Contents

Original Papers

Evaluating the Effectiveness of Interactive Virtual Patients for Medical Education in Zambia: Randomized Controlled Trial (e43699)
Rebecca Horst, Lea-Mara Witsch, Rayford Hazunga, Natasha Namuziya, Gardner Syakantu, Yusuf Ahmed, Omar Cherkaoui, Petros Andreadis, Florian Neuha n, Sandra Bartelt .................................................. 3

The Educational Impact of Web-Based, Faculty-Led Continuing Medical Education Programs in Type 2 Diabetes: A Survey Study to Analyze Changes in Knowledge, Competence, and Performance of Health Care Professionals (e40520)
Stewart Harris, Shannon Idzik, Adriano Boasso, Sola Neunie, Alexander Noble, Helen Such, Joanna Van .................................................. 49

Defining a Role for Webinars in Surgical Training Beyond the COVID-19 Pandemic in the United Kingdom: Trainee Consensus Qualitative Study (e40106)
Emma Barlow, Wajiha Zahra, Jane Hornsby, Alex Wilkins, Benjamin Davies, Joshua Burke .................................................. 66

Effects of Practicing With and Obtaining Crowdsourced Feedback From the Video-Based Communication Assessment App on Resident Physicians’ Adverse Event Communication Skills: Pre-post Trial (e40758)
Andrew White, Ann King, Angelo D’Addario, Karen Brigh t, Suzanne Dintzis, Emily Fay, Thomas Gallagher, Kathleen Mazor .................................................. 76

A Fork in the Road for Emergency Medicine and Critical Care Blogs and Podcasts: Cross-sectional Study (e39946)
Michelle Lin, Mina Phipps, Yusuf Yilmaz, Christopher Nash, Michael Gisondi, Teresa Chan .................................................. 85

Software-Based Simulation on a 3D Environment for Vaccination Teaching and Learning: Design Science Research (e35712)
Daniel Domingueti, Darlinton Barbosa Feres Carvalho, Diego Colombo Dias, Valéria Oliveira .................................................. 94

Perceptions of US Medical Students on Artificial Intelligence in Medicine: Mixed Methods Survey Study (e38325)
David Liu, Jake Sawyer, Alexander Luna, Jihad Aoun, Janet Wang, Lord Boachie, Safwan Halabi, Bina Joe .................................................. 119

Google Images Search Results as a Resource in the Anatomy Laboratory: Rating of Educational Value (e37730)
Alexandra Wink, Amanda Telfer, Michael Pascoe .................................................. 130

Developing and Integrating Asynchronous Web-Based Cases for Discussing and Learning Clinical Reasoning: Repeated Cross-sectional Study (e38427)
Sonny Tat, Haroon Shaukat, Pavan Zaveri, Maybelle Kou, Lenore Jarvis .................................................. 143
The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation (e47049)
Zuheir Khlaif, Allam Mousa, Muayad Hattab, Jamil Itmazi, Amjad Hassan, Mageswaran Sanmugam, Abedalkarim Ayyoub.  149

Reviews

Health Literacy in Health Professionals Two Years into the COVID-19 Pandemic: Results From a Scoping Review (e39023)
Eva-Maria Grepmeier, Maja Pawellek, Janina Curbach, Julia Sommogyi, Karl Drewitz, Claudia Hasenpusch, Eva Bitzer, Christian Apfelbacher, Uwe Mattern.  15

Effectiveness of Shared Decision-making Training Programs for Health Care Professionals Using Reflexivity Strategies: Secondary Analysis of a Systematic Review (e42033)
Ndeye Diouf, Angèle Musabyimana, Virginie Blanchette, Johanie Lépine, Sabrina Guay-Bélanger, Marie-Claude Tremblay, Maman Dogba, France Légare.  25

Viewpoints

The Role of Academic Health Systems in Leading the “Third Wave” of Digital Health Innovation (e32679)
Adeel Faruki, Richard Zane, Jennifer Wiler.  44

Emoji Education: How Students Can Help Increase Health Awareness by Making Emojis (e39059)
Sammer Marzouk, Shuhan He, Jarone Lee.  112

Incorporating Paid Caregivers Into Medical Education to Enhance Medical Student Exposure to This Essential Workforce (e38329)
Samir Kamat, George Danias, Aneesh Agarwal, Sumanth Chennareddy, Joseph Han, Samuel Lee.  115
Evaluating the Effectiveness of Interactive Virtual Patients for Medical Education in Zambia: Randomized Controlled Trial

Rebecca Horst1*, MSc; Lea-Mara Witsch1*, MD; Rayford Hazunga1,2, MD; Natasha Namuziya1,2, MD; Gardner Syakantu2, MD; Yusuf Ahmed2, MD; Omar Cherkaoui3, MD; Petros Andreadis4, PhD; Florian Neuhann1,2, MD; Sandra Barteit1, MA, MSc, DrScHum

1Faculty of Medicine and University Hospital, Heidelberg Institute of Global Health (HIGH), Heidelberg University, Heidelberg, Germany
2Levy Mwanawasa Medical University, Lusaka, Zambia
3AMBOSS Global Health Initiative, Berlin, Germany
4SolidarMed Zambia, Lusaka, Zambia
*these authors contributed equally

Corresponding Author:
Sandra Barteit, MA, MSc, DrScHum
Faculty of Medicine and University Hospital, Heidelberg Institute of Global Health (HIGH)
Heidelberg University
Im Neuenheimer Feld 130.3
Heidelberg, 69120
Germany
Phone: 49 062215634030
Email: barteit@uni-heidelberg.de

Abstract

Background: Zambia is facing a severe shortage of health care workers, particularly in rural areas. Innovative educational programs and infrastructure have been established to bridge this gap; however, they encounter substantial challenges because of constraints in physical and human resources. In response to these shortcomings, strategies such as web-based and blended learning approaches have been implemented, using virtual patients (VPs) as a means to promote interactive learning at the Levy Mwanawasa Medical University (LMMU) in Zambia.

Objective: This study aimed to evaluate the students’ knowledge acquisition and acceptance of 2 VP medical topics as a learning tool on a Zambian higher education e-learning platform.

Methods: Using a mixed methods design, we assessed knowledge acquisition using pre- and posttests. In a randomized controlled trial setting, students were assigned (1:1) to 2 medical topics (topic 1: appendicitis and topic 2: severe acute malnutrition) and then to 4 different learning tools within their respective exposure groups: VPs, textbook content, preselected e-learning materials, and self-guided internet materials. Acceptance was evaluated using a 15-item questionnaire with a 5-point Likert scale.

Results: A total of 63 third- and fourth-year Bachelor of Science clinical science students participated in the study. In the severe acute malnutrition–focused group, participants demonstrated a significant increase in knowledge within the textbook group ($P=.01$) and the VP group ($P=.01$). No substantial knowledge gain was observed in the e-learning group or the self-guided internet group. For the appendicitis-focused group, no statistically significant difference in knowledge acquisition was detected among the 4 intervention groups ($P=.62$). The acceptance of learning materials exhibited no substantial difference between the VP medical topics and other learning materials.

Conclusions: In the context of LMMU, our study found that VPs were well accepted and noninferior to traditional teaching methods. VPs have the potential to serve as an engaging learning resource and can be integrated into blended learning approaches at LMMU. However, further research is required to investigate the long-term knowledge gain and the acceptance and effectiveness of VPs in medical education.

Trial Registration: Pan African Clinical Trials Registry (PACTR) PACTR202211594568574; https://pactr.samrc.ac.za/TrialDisplay.aspx?TrialID=20413

(JMIR Med Educ 2023;9:e43699) doi:10.2196/43699

https://mededu.jmir.org/2023/1/e43699

JMIR Med Educ 2023 | vol. 9 | e43699 | p.3
(page number not for citation purposes)
KEYWORDS

global health; Zambia; health care workers; medical skills; e-logbook, digital global health

Introduction

Background

The critical need to enhance Zambia’s health care workforce, particularly involving doctors, nurses, and other health care workers (HCWs), is driven by a substantial shortage of qualified HCWs. In this study, we define HCWs as individuals whose primary professional goal is to maintain or enhance the health of others, including those who provide patient care and diagnostic and treatment services across various clinical settings. The scarcity of HCWs is particularly acute in Zambia’s rural areas, highlighting the importance of addressing this issue to ensure equitable access to health care and improved health outcomes across the country. Addressing this HCWs deficit is essential for enhancing public health outcomes and providing equitable care across the country. Exacerbating this situation are infrastructural obstacles such as the protracted process of developing educational facilities. The convergence of these factors impede the capacity expansion and the enhancement of health profession training quality. Consequently, addressing these constraints is vital for improving public health outcomes and fostering a robust HCW workforce. These constraints contribute to adverse public health outcomes, including compromised disease treatment efficacy, increased child mortality, and poor maternal health [1]. In 2002, Zambia introduced medical licentiate practitioners (MLPs) through an initial 3-year diploma program, followed by a 2-year medical licentiate, to address the shortage of qualified HCWs, particularly in rural areas, and to upgrade the health care delivery scope of existing clinical officers. MLPs, who have a distinct role compared with traditional medical graduates, now complete a 4-year training program, for example, at the Levy Mwanawasa Medical University (LMMU), earning a Bachelor of Science (BSc) in clinical science [2-4]. This specialized training enables MLPs to perform a limited number of emergency surgeries, including cesarean sections, and to prescribe medications [2].

Their education focuses on 4 primary disciplines that align with the country’s health priorities, particularly in rural areas: surgery, pediatrics, obstetrics and gynecology, and internal medicine [3,4]. The MLP training program comprises a balanced structure, including 2 years of theoretical instruction followed by 2 years of hands-on skills training acquired through rotational assignments at a diverse range of hospitals and health clinics across Zambia.

In low-resource learning environments, such as Zambia, e-learning and self-directed internet materials play a crucial role in overcoming the challenges posed by the scarcity of qualified HCWs, limited infrastructure, and constrained budgets. e-Learning allows for the expansion and enhancement of medical education by providing students with access to up-to-date information and resources regardless of their location. This approach is particularly beneficial for students in rural areas, where there may be a lack of qualified medical educators, limited access to learning resources, and inadequate infrastructure to support traditional face-to-face instruction [3,5].

Furthermore, self-directed internet materials encourage learners to take charge of their own education, allowing them to study at their own pace and focus on the topics most relevant to their professional development. This flexibility is crucial in low-resource settings, where e-learning can provide access to up-to-date information, reducing the reliance on on-site classes in areas facing a shortage of medical educators [3-5]. The integration of e-learning and self-directed internet materials into medical education programs in low-resource settings can help bridge the gap between theoretical knowledge and practical application, thus improving the overall quality of health care [6-8]. For example, virtual patient (VP) scenarios can provide students with a more interactive and engaging learning experience, enhancing their clinical reasoning skills and compensating for the scarcity of senior HCWs for face-to-face training [9]. By leveraging e-learning and self-directed internet materials, medical education programs in low-resource settings can better prepare students for the challenges they will face in their professional careers, ultimately leading to improved health care outcomes.

Interactive Medical Learning Through VPs

VPs are defined by the American Association of Medical Colleges as “a specific type of computer-based program that simulates real-life clinical scenarios” [10]. VPs offer the advantage of addressing multiple cognitive levels while enhancing the learning experience through supplementary channels such as visual and auditory information. Students may develop clinical reasoning skills by using VPs, which bridge the gap between theoretical clinical knowledge and practical clinical application [3-5]. Assuming the role of a clinician, VPs enable students to practice diagnostic, treatment, and follow-up procedures. This approach may bolster student motivation as they become more aware of the importance of practice [10-12], rather than solely focusing on academic performance or examination scores [13,14]. A systematic review conducted by Kononowicz et al [11] revealed that VPs are capable of improving skills and knowledge as effectively as, or even surpassing, other prevalent educational methodologies. They observed improvements in clinical reasoning, procedural skills, and a combination of procedural and team skills in both low-income and high-income settings [11]. Bediang et al [15] assessed the impact of VP training on the clinical skills of Cameroonian health care professionals and found that such training could contribute to the advancement of users’ clinical operational skills [15].

In the Zambian context, implementing VPs as a learning resource can help mitigate challenges associated with the shortage of senior medical lecturers and infrastructure limitations. By integrating VPs into the existing e-learning platform, we can provide a more interactive and engaging learning experience, enhance clinical reasoning skills, and compensate for the scarcity of senior HCWs for face-to-face training. Furthermore, VPs can be accessed remotely, allowing students in rural areas or those with limited access to traditional...
educational resources to benefit from this innovative learning approach [9]. To improve medical education at LMMU within the BSc clinical sciences program, we devised and assessed 2 VP scenarios. VPs could be a highly accepted and effective tool for integration into the existing e-learning platform for MLPs at LMMU, thus expanding the range of digital learning resources available. Our study specifically focused on examining 2 primary research questions that aimed to evaluate the effectiveness and impact of these scenarios on student learning and clinical reasoning skills. By addressing these questions, we sought to provide valuable insights into the potential benefits and limitations of implementing VP scenarios in medical education, particularly within the context of low-resource settings such as Zambia. Outcomes may differ from those in high-income countries because of differences in educational systems, infrastructure, and technological access. Evaluating the potential impact of VPs in a context such as LMMU can inform targeted interventions aimed at enhancing medical education, which may ultimately contribute to improved health care outcomes within such settings.

In the context of LMMU, our study aimed to address two primary research questions: (1) How effectively do VPs contribute to knowledge acquisition in comparison with the traditional learning resources prevalent in Zambia, such as learning from textbooks, using free internet searches, and accessing preselected static resources on a medical e-learning platform? and (2) How does student acceptance of VPs compare with their acceptance of traditional textbooks, internet searches, and a medical e-learning platform as learning tools?

**Methods**

**Overview**

We conducted a noninferior, randomized controlled trial with a mixed methods research design (convergent) to evaluate the effectiveness of VPs in terms of acceptance and knowledge acquisition (Figure 1). The analysis team was blinded to the study. The students, who were informed through a flyer distributed beforehand, were aware that the study aimed to investigate VPs as a learning method. The CONSORT (Consolidated Standards of Reporting Trials) checklist was used for reporting this study [16] (for the CONSORT checklist, refer to Multimedia Appendix 1). The study participants were recruited on November 29, 2021, and the study took place on December 10, 2021, at the main campus of LMMU in Lusaka, Zambia. All third- and fourth-year BSc clinical science students aged ≥18 years were eligible to participate in this study.

![Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram detailing the randomized controlled trial, following the stages of enrollment, allocation, and analysis.](image-url)

**Randomization, Blinding, and Implementation**

The study participants were recruited through digital messaging services, email, and the local university administration. A total of 63 third- and fourth-year BSc clinical science students aged ≥18 years were invited to participate in this study. Randomization was implemented in a 2-step process. Initially, participants were assigned to 1 of the 2 study groups (appendicitis or severe acute malnutrition [SAM]) based on their study ID. Subsequently, within each study group,
participants were stratified according to their academic year. We used stratified randomization to guarantee a balanced allocation of participants, considering the stratum of academic year. This procedure ensured an equitable distribution across study groups while addressing potential variations associated with participants’ academic advancement. To maintain the integrity of the study, the data analysis team remained independent of the data collection process. Only third- and fourth-year BSc clinical science students were asked to participate in the study, as they had previously been exposed to the 2 medical topics of appendicitis and SAM during their first 2 years of university training. The learning resources that the students were exposed to are presented in Textbox 1.

Textbox 1. Students’ learning resources.

- Interactive virtual patient (VP) medical topics
  - Severe acute malnutrition (SAM): The VP medical topic was developed using materials from the World Health Organization’s country guidelines on managing SAM in infants and children [17], web-based resources [18,19], and relevant sections from Nelson’s Textbook of Pediatrics [20] (refer to Multimedia Appendix 2 for the detailed VP medical topic).
  - Appendicitis: The VP medical topic was developed using materials from the AMBOSS e-learning platform, specifically the website on appendicitis [21] (refer to Multimedia Appendix 3 for the detailed VP medical topic).

- Textbook contents aligned with the Bachelor of Science clinical science curriculum

- e-Learning materials were preselected from the medical e-learning platform AMBOSS [23], which was made available on a complementary basis to the Levy Mwanawasa Medical University faculty and students.

- Self-guided internet materials were made accessible to study participants, allowing them to independently investigate 1 of the 2 topics (appendicitis or SAM) using their own search terms. This approach facilitates autonomous exploration and information gathering on the subject through internet resources.

Both VP medical topics were uploaded to LMMU’s Moodle e-learning platform [3] but remained inaccessible to participants until the day of the trial.

All participants, regardless of their assigned study group, completed a pretest before accessing their designated learning resource for a 30-minute period. After the intervention, a posttest identical to the pretest was administered to all the participants. Furthermore, each participant completed a questionnaire evaluating their acceptance of the respective learning resource (Multimedia Appendix 4). During the entire 4-hour study period, participants were explicitly instructed to refrain from communicating with one another.

Data Collection

To evaluate knowledge acquisition from the 4 learning resources, we administered multiple-choice question (MCQ) tests before (pretest) and after (posttest) the intervention. The appendicitis-related MCQ test comprised 20 questions (maximum score: 1000 points), whereas the SAM-related test contained 15 questions (maximum score: 720 points; refer to Multimedia Appendix 5 for pre- and posttests). Each question was presented with 4 answer options, with 1 correct answer. All the groups received identical questions. An internal pilot study was conducted before the randomized controlled trial to ensure that participants could successfully pass the tests using any of the 4 learning resources. The pilot study involved a small sample of participants (n=6), including students (n=4) and faculty members (n=2), who were not part of the main study. The primary objective of this pilot study was to assess the clarity, comprehensibility, and effectiveness of the study materials, including the questionnaires and VPs. This preliminary testing helped to identify any potential issues, ambiguities, or biases in the questions and to evaluate the overall efficacy of the questionnaire. By addressing these concerns, we aimed to enhance the validity and reliability of the study instruments and, ultimately, the quality of the data collected in the main study.

The pilot study addressed uncertainties regarding the effectiveness and time allocation of the various study methods. This facilitated the refinement of the learning materials and ensured appropriate time durations for each method, allowing for adequate knowledge acquisition. The pilot study also confirmed that each learning resource covered course objectives, preventing participants from focusing solely on pretest questions, and enabled the assessment of pre- and posttest questions to accurately measure knowledge acquisition across learning methods. By integrating the pilot study findings into the main study design, the research team mitigated concerns about the effectiveness of various study methods and the sufficiency of the time allocated for each approach.

The VPs were developed to be consistent with the Zambian context, incorporating relevant resources and guidelines, such as the World Health Organization’s country-specific guidance on managing SAM in infants and children [17], and the curricular standards of LMMU. This method ensured that the VPs were customized to suit the requirements and expectations of BSc clinical science students as well as to address the specific demands of the local health care system.

Acceptance Questionnaire

We assessed the acceptance of the 4 learning resources using a questionnaire adapted from the study by Davis [24] on the technology acceptance model. The technology acceptance model...
includes six dimensions: (1) perceived usefulness, (2) perceived ease of use, (3) attitude toward use, and (4) behavioral intention to use, (5) job relevance, and (6) perceived enjoyment. The original fifth component (actual system use) was excluded from our study’s questionnaire, as it did not pertain to our research objectives. We incorporated 2 additional dimensions—(5) and (6)—based on the study by Salloum et al [25]. The questionnaire comprised 15 items, with responses recorded on a 5-point Likert scale.

**Data Analysis**

Before conducting the analyses, we examined all data for normal distribution using the Shapiro-Wilk test. P values <.05 were considered statistically significant. We analyzed both study groups separately. Descriptive statistics including frequency, percentage, mean, median, and SD were used to evaluate the distribution of age, sex, and prior medical knowledge within the groups. Here, prior medical knowledge refers to the participants’ preexisting knowledge specifically related to appendicitis and SAM.

The pre- and posttest outcomes were evaluated through the following analyses:

1. We used an ANOVA test followed by a paired 1-tailed t test with Bonferroni correction as a post hoc test to assess any variations in prior knowledge across the 4 study groups for each of the 2 study groups’ topics (appendicitis and SAM).

2. We applied the same approach described in the previous point to evaluate the differences in postintervention knowledge levels across the groups. This analysis aimed to identify whether there were any significant differences in the knowledge levels between the groups after exposure to different learning resources, which could indicate the relative effectiveness of each resource.

3. To assess within-group knowledge acquisition, we compared the pre- and posttest results for each group. We used a Wilcoxon rank test for the 3 groups using textbook contents, e-learning materials, and self-guided internet materials. For the group using VP, we used a t test as the data were normally distributed. This analysis helped determine the extent of knowledge gain within each group after using their assigned learning resource.

The 5-point Likert scale was converted to numerical values (1=strongly agree, 2=agree, 3=neutral, 4=disagree, and 5=strongly disagree). We assessed acceptance across the 6 dimensions using descriptive statistics. To evaluate whether there was a statistically significant difference in the acceptability of the 4 learning resources among all intervention groups, we applied the Kruskal-Wallis test. As a post hoc analysis, we conducted a Wilcoxon rank test with Bonferroni correction.

**Ethics Approval, Informed Consent, and Participation**

This study was approved by the Heidelberg University Hospital Ethical Committee on August 30, 2021 (S-685/2021) and the LMMU Research Ethics Committee on November 29, 2021 (LMMU-REC 00005/21). We informed all potential and selected participants about the study’s objectives and procedures as well as their right to withdraw at any time without consequences. Before participation, each individual provided written informed consent, ensuring their voluntary involvement in the study.

**Results**

**Demographics**

A total of 63 students consented to participate and were included in the study. The mean age of the participants was 39.56 (SD 6.05) years, with age ranging from 22 to 46 years. The sample consisted of 39 male students and 23 female students, with 1 participant identifying as “diverse.” Regarding the participants’ academic progression, 32 were in their third year of study, whereas 31 were in their fourth year (refer to Table 1 for a detailed breakdown).

**Table 1.** Overview of the study groups’ composition and demographic characteristics (N=63).

<table>
<thead>
<tr>
<th>Intervention group</th>
<th>Participants, n</th>
<th>Age (years), mean (SD)</th>
<th>Sex, n (%)</th>
<th>Study year, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Study group 1—medical topic: severe acute malnutrition (n=32)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual patient</td>
<td>8</td>
<td>29.38 (1.77)</td>
<td>4 (50)</td>
<td>4 (50)</td>
</tr>
<tr>
<td>Textbook</td>
<td>8</td>
<td>27 (5.04)</td>
<td>3 (38)</td>
<td>5 (62)</td>
</tr>
<tr>
<td>Preselected e-learning materials</td>
<td>8</td>
<td>30.38 (6.82)</td>
<td>4 (50)</td>
<td>3 (38)</td>
</tr>
<tr>
<td>Self-guided internet materials</td>
<td>8</td>
<td>32.75 (9.66)</td>
<td>0 (0)</td>
<td>8 (100)</td>
</tr>
<tr>
<td>Study group 2—medical topic: appendicitis (n=31)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Virtual patient</td>
<td>8</td>
<td>29.39 (4.57)</td>
<td>5 (62)</td>
<td>3 (57)</td>
</tr>
<tr>
<td>Textbook</td>
<td>7</td>
<td>29.17 (3.87)</td>
<td>2 (29)</td>
<td>5 (71)</td>
</tr>
<tr>
<td>Preselected e-learning materials</td>
<td>8</td>
<td>33.75 (4.74)</td>
<td>1 (13)</td>
<td>7 (87)</td>
</tr>
<tr>
<td>Self-guided internet materials</td>
<td>8</td>
<td>32.28 (7.51)</td>
<td>4 (50)</td>
<td>4 (50)</td>
</tr>
</tbody>
</table>
MCQ Pre- and Posttests

Pre- and Posttests for SAM

The pretest revealed a significant difference in knowledge between the VP group (mean score 480, SD 76.97) and the textbook group (mean score 456, SD 67.88) participants (P=0.01) as well as a difference between the VP group and self-guided internet group (P=0.05; refer to Table 2 for detailed results). Participants in the VP group achieved the highest pretest scores, with a mean of 68% (43/63; mean score 480, SD 76.97) of the questions correctly answered, followed closely by the e-learning group, who correctly answered 63% (40/63) of the questions (mean score 456, SD 67.88). Students in the self-guided internet group scored 50% (mean score 360, SD 105.79), whereas students in the textbook group scored 48% (mean score 342, SD 82.89). Although differences in scores between the groups persisted in the posttest compared with the pretest, the overall knowledge gap between the groups narrowed:

The VP group scored 79% (mean score 570, SD 82.89), the e-learning group scored 68% (mean score 468, SD 132.7), the self-guided internet group scored 59% (mean score 426, SD 165.16), and the textbook group scored 59% (mean score 426, SD 82.89; refer to Table 2 for details).

No significant increase in knowledge was observed in the e-learning group (mean score 468, SD 132.7) and the self-guided internet group (mean score 426, SD 165.16). However, significant knowledge growth was identified in both the textbook group (P=0.01) and the VP group (P=0.01; refer to Multimedia Appendix 6 for details).

Table 2. Overview of the 4 exposure intervention groups that were exposed to the medical topic of severe acute malnutrition as well as their relative test scores (pre- and posttest) and average knowledge increase.

<table>
<thead>
<tr>
<th>Intervention group (severe acute malnutrition)</th>
<th>Pretest score, mean (SD)</th>
<th>Posttest score, mean (SD)</th>
<th>Knowledge gain score, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual patient</td>
<td>480 (76.97)</td>
<td>570 (82.89)</td>
<td>90 (90.48)</td>
</tr>
<tr>
<td>Textbook contents</td>
<td>342 (82.89)</td>
<td>426 (82.89)</td>
<td>84 (80.11)</td>
</tr>
<tr>
<td>Preselected e-learning materials</td>
<td>456 (67.88)</td>
<td>468 (132.7)</td>
<td>12 (114.02)</td>
</tr>
<tr>
<td>Self-guided internet materials</td>
<td>360 (105.79)</td>
<td>426 (165.16)</td>
<td>66 (122.88)</td>
</tr>
</tbody>
</table>

Pre- and Posttests for Appendicitis

The pretest for the medical topic of appendicitis did not reveal a significant difference in knowledge acquisition between the 4 intervention groups (P=0.62; refer to Table 3 for detailed results). Participants in the e-learning group achieved the highest pretest score, with a mean score 75% (mean score 750, SD 46.29), closely followed by participants in the self-guided internet group with 74% (mean score 73.5, SD 112.6), the textbook group with 74% (mean score 735, SD 146.3), and the VP group with 72% (mean score 718, SD 106.7). In the posttest, a difference was observed between the intervention groups, although this difference was not substantial (Table 3). On the basis of the individual learning curves, moderate knowledge acquisition was observed in all 4 groups, although these changes were not substantial (refer to Multimedia Appendix 7 for details).

Table 3. Overview of the 4 exposure intervention groups that were exposed to the medical topic of appendicitis as well as their relative test scores (pre- and posttest) and average knowledge increase.

<table>
<thead>
<tr>
<th>Intervention group (appendicitis)</th>
<th>Pretest score, mean (SD)</th>
<th>Posttest score, mean (SD)</th>
<th>Knowledge acquisition, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual patient</td>
<td>718 (106.7)</td>
<td>800 (128.17)</td>
<td>81.25 (106.7)</td>
</tr>
<tr>
<td>Textbook</td>
<td>735 (146.3)</td>
<td>821.43 (128.64)</td>
<td>85.71 (85.22)</td>
</tr>
<tr>
<td>Preselected e-learning materials</td>
<td>750 (46.29)</td>
<td>887.5 (74.4)</td>
<td>137.5 (87.63)</td>
</tr>
<tr>
<td>Self-guided internet materials</td>
<td>735 (112.6)</td>
<td>850 (75.59)</td>
<td>112.5 (99.1)</td>
</tr>
</tbody>
</table>

Acceptance Questionnaire

We observed that the acceptance of learning resources varied depending on the medical topic. For the topic of SAM, the response to the statement “If given the opportunity, I would favor this learning resource over others” showed a significant difference between the e-learning group (mean 2.5, SD 0.33; P=0.01) and the self-guided internet group (mean 1.25, SD 0.46). In contrast, the VP group had a favorable score (mean 1.62, SD 0.72). For the topic of appendicitis, the mean response to the same statement in the e-learning group was 1.38 (SD 0.52), which is significantly lower than the mean response in the VP group (mean 3.62, SD 1.41; P=0.02).

Regarding the statement “I think this learning resource is a good instrument to acquire knowledge,” a difference between the e-learning group (mean 1.5, SD 0.53) and the self-guided internet group (mean 3.12, SD 1.13) with P=0.02 was observed for the topic of appendicitis. This finding indicated that the e-learning group received more positive feedback than the self-guided internet group (Table 4).

One misinterpreted question in the study was removed because of its outlier status (question item: “Interacting with the learning mode required considerable effort.”).
Table 4. The acceptance questionnaire results for both study groups by intervention groups showing mean and SD.

<table>
<thead>
<tr>
<th>Study arm and questions items (acceptance questionnaire)</th>
<th>Intervention groups, mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Virtual patient group</td>
</tr>
<tr>
<td>Study arm 1: severe acute malnutrition</td>
<td></td>
</tr>
<tr>
<td>&quot;I think this learning resource is a good instrument to acquire knowledge.&quot;</td>
<td>1.75 (0.89)</td>
</tr>
<tr>
<td>&quot;If given the opportunity, I would favor this learning resource over others.&quot;</td>
<td>1.62 (0.74)</td>
</tr>
<tr>
<td>Study arm 2: appendicitis</td>
<td></td>
</tr>
<tr>
<td>&quot;I think this learning resource is a good instrument to acquire knowledge.&quot;</td>
<td>2.12 (0.64)</td>
</tr>
<tr>
<td>&quot;If given the opportunity I would favor this learning resource over others.&quot;</td>
<td>3.62 (1.41)</td>
</tr>
</tbody>
</table>

Discussion

Principal Findings

In our study, we developed and assessed 2 interactive VP medical topics focusing on SAM and appendicitis, aiming to evaluate their effectiveness and acceptance compared with other prevalent learning resources at LMMU. The efficacy of transferring knowledge to students, and the precise impact of certain VP features, had previously been ambiguous. Our study aimed to address these aspects.

The primary aim of this study was to evaluate the acceptance and knowledge acquisition of BSc clinical sciences students at LMMU when using VPs as a learning resource in comparison with textbooks, preselected e-learning materials, and self-guided internet materials. A key finding of this study was that all 4 learning resources demonstrated their effectiveness in promoting knowledge gain within the study setting. Furthermore, VPs were well received by the students and proved to be noninferior compared with the other 3 learning methods.

Comparison With Prior Work

Knowledge acquisition significantly increased in the VP and textbook groups but not in the e-learning or self-guided internet groups. The differences in knowledge acquisition between these groups can be attributed to various factors. Each learning resource provides different levels of structure and guidance, with some students preferring visual or interactive content (eg, VPs) and others opting for text-based resources (eg, textbooks). These individual preferences may influence the effectiveness of each learning method, thus impacting knowledge acquisition across the groups. Motivation and engagement may also play a role, as VPs and textbooks potentially offer a more structured and engaging learning experience, which could lead to increased motivation and improved information retention. In contrast, general e-learning and self-guided internet materials may require a higher degree of self-discipline and motivation to effectively navigate and absorb the content. Another aspect to consider is the familiarity with learning resources. Students might be more familiar with traditional learning resources, such as textbooks, compared with newer methods, such as VPs or e-learning platforms. This familiarity could influence the ease with which students can use and learn from these resources, thus affecting their knowledge acquisition. Finally, access to specific, targeted learning materials is essential. The VP and textbook groups had access to well-organized, systematic learning materials, making it easier for students to focus on relevant content and efficiently grasp key concepts. In contrast, participants in the e-learning and self-guided internet groups had to navigate and search for pertinent information independently. This process could be time consuming and challenging, as students might encounter a vast amount of information of varying complexity that is not always directly related to course objectives. Consequently, these students may have faced difficulties in identifying and assimilating the critical knowledge required for the subject matter, leading to a smaller increase in knowledge acquisition compared with their counterparts in the VP and textbook groups. However, this observation was not reflected in the acceptance questionnaire. In the second trial group (appendicitis), the pre- and posttest results revealed no significant differences in knowledge acquisition among participants in the intervention groups, although all 4 groups demonstrated an increase in knowledge. The acceptance questionnaires indicated similar responses for all 4 learning resources across 6 technology acceptance dimensions (perceived usefulness, perceived ease of use, attitude toward using, behavioral intention to use, job relevance, and perceived enjoyment) but showed mixed results when comparing the 2 medical subjects of SAM and appendicitis.

Participants exposed to the SAM VP intervention displayed a higher preference for this learning resource, which was not observed in the group exposed to the appendicitis VP. This difference could be attributed to the SAM VP’s integration of more images and a visually appealing design, making it more engaging for students. The varying success of VPs for the 2 subjects may be attributed to differences in content and design. The appendicitis VP might have lacked the engaging elements found in the SAM VP, leading to a lower preference among participants. In addition, the participants’ higher familiarity with or the lower complexity of 1 topic could have contributed to the observed differences in the success of the respective VPs. However, it is important to consider that the design of the 2 VP
medical topics may have acted as a confounding factor, potentially influencing the outcomes. As such, we cannot definitively attribute the observed differences solely to the subject matter or the VP medical topic design. Further research is needed to identify the specific factors that contribute to the success of VPs in medical education.

In the SAM group, the preselected e-learning materials (AMBOSS platform) received the lowest mean rating among all intervention groups. Conversely, the appendicitis group demonstrated a positive response to the AMBOSS platform but displayed indifference toward self-guided web-based learning materials. The disparity in the acceptance of the AMBOSS platform as a learning resource between the 2 study groups might be attributed to the platform’s content, which primarily targets the global north.

The content of the AMBOSS platform may not be adequately tailored to the specific learning needs and objectives of the SAM group, possibly because of differences in guidelines, treatment protocols, or context-specific challenges in managing SAM between the United States and Zambia. In addition, the AMBOSS platform may use terminology, examples, or scenarios predominantly familiar to US-based learners, which could pose comprehension difficulties for Zambian students when addressing the specific topic of SAM. SAM is a pressing issue in Zambia, with treatment priorities and modalities that may differ from those in the United States. In contrast, appendicitis holds similar importance in both countries.

Pre- and Posttests

Upon comparing the pre- and posttest results of all study participants, the most substantial improvement was observed among fourth-year students. This finding aligns with the study by Kiesewetter et al [26], who reported that students with less prior knowledge experienced a greater cognitive load compared with those with more prior knowledge. A potential advantage of using VPs is their accessibility, as textbooks can be expensive, sometimes scarce, and may contain outdated content when published. Incorporating VPs into the curriculum can help overcome these limitations and provide students with up-to-date and readily available learning resources. Overall, our pre- and posttest findings indicate that VPs are as effective in promoting learning as other widely used learning resources. Previous research has indicated that the use of VPs leads to significant increases in knowledge, enhanced understanding, and improved problem-solving skills when compared with lecture-based small seminar groups [3,27,28]. These studies also evaluated long-term knowledge retention, revealing no discernible differences between the 2 groups over a 4- to 6-week period.

Acceptance Questionnaire

The observed disparity in the acceptance of VPs between the 2 study groups, SAM and appendicitis, could potentially be attributed to the differences in the design of the 2 VP medical topics. Peddle et al [29] conducted a study involving student interviews to better understand the acceptance of VPs and discovered that incorporating images improved student comprehension and facilitated knowledge retention. The study emphasized the benefits of using short videos to promote knowledge acquisition.

In general, responses to the acceptance questionnaire were predominantly positive. Participants frequently selected responses that ranged from positive to neutral, whereas negative responses were rare. This pattern was also observed in other studies [30]. The study by Krumpal [31] described this phenomenon as individuals considering risks and losses when determining a response, as they seek social acceptability.

To address this potential bias, we communicated with the participants before the study, emphasizing that their responses to the questionnaires would be handled anonymously and would not affect their academic performance.

Future studies should carefully control for potential confounding factors, such as differences in design, when examining the effectiveness of VPs across different medical subjects. This would allow for a more accurate assessment of the impact of subject matter and case design on learning outcomes.

Strengths and Limitations

Our study, which encompassed a majority of students and investigated 4 distinct learning methods, provides valuable insights into how these approaches might improve student learning. However, several limitations of this study should be acknowledged.

First, the generalizability of our findings is limited because of the study population, which exclusively consisted of third- and fourth-year BSc clinical science students. Consequently, our results may not be directly applicable to other contexts, such as different disciplines.

Second, the content of the AMBOSS platform may not have been adequately tailored to the specific learning needs and objectives of SAM in Zambia. Disparities in the content and design of VP between the 2 case scenarios may also result in differences in the study outcomes. Future studies should consider developing a more robust method for comparing VP cases, such as using a larger sample of case scenarios or ensuring that the cases are matched in terms of difficulty and complexity.

Third, owing to a 1-hour delay at the study’s onset caused by technical issues, some participants might have experienced time pressure during the later stages (posttest and acceptance questionnaire), potentially introducing bias. To mitigate this concern, we requested all participants to wait until the last participant completed the study. However, the delay could have resulted in increased fatigue among the participants, affecting their concentration, motivation, and overall performance during the learning sessions and the pre- and posttests. This factor could have potentially impacted knowledge acquisition, leading to lower scores across all intervention groups. Moreover, the delay may have induced stress or frustration that could have influenced their approach to the learning sessions and the pre- and posttests, resulting in less accurate or reliable data and affecting the overall interpretation of the study results.

Fourth, the study did not examine long-term knowledge retention or evaluate traits such as clinical reasoning using MCQs. Therefore, additional research is warranted to investigate...
long-term knowledge acquisition, as this study concentrated solely on immediate knowledge gain.

Fifth, in our study, we acknowledge the possibility that participants may have been hesitant to provide negative feedback because of concerns regarding anonymity and potential implications for their academic performance. To address this concern, we implemented several measures to ensure anonymity of the data collected. These measures included (1) emphasizing the confidentiality of the study in the information provided to the participants, both verbally and in written form; (2) assigning unique participant identification numbers, which were not linked to personal information, to protect the identity of the participants during data collection and analysis; (3) ensuring that the questionnaires were completed individually and without peer or instructor influence; and (4) storing the collected data securely and restricting access to only the researchers directly involved in the study.

By implementing these measures, we aimed to minimize any potential bias arising from the participants’ reluctance to provide negative feedback. Nevertheless, it is important to recognize that a certain degree of social desirability bias may still be present, which is common in self-reporting studies. Future research could explore alternative methods of data collection or use more indirect questioning techniques to further reduce the impact of such biases on the study results.

Future Directions
At present, LMMU is in the process of revising its e-learning strategy, with the aim of fully integrating e-learning into the curriculum in the future. In light of the findings from our study, it is important to consider potential methodological concerns, such as the content of the AMBOSS platform not being adequately tailored to the specific learning needs and objectives of SAM for Zambia, as well as disparities in the content and design of VP between the 2 case scenarios that may result in differences in study outcomes. Despite these concerns, the updated e-learning strategy and our study results may support the potential inclusion of VPs within the curriculum as a means to enhance medical education at LMMU. Future research should address these methodological concerns to ensure that the implementation of VP scenarios is tailored to the specific needs and contexts of medical education in Zambia, ultimately leading to more robust and reliable outcomes.

Conclusions
The primary aim of this study was to assess the acceptability and effectiveness of VPs for knowledge acquisition in the BSc clinical science program at LMMU in Zambia, comparing their performance with 3 other prevalent learning resources: textbook content, preselected e-learning materials, and self-guided internet materials. In the context of a low-resource setting, our findings demonstrate that although VPs are well accepted, their effectiveness in terms of knowledge acquisition may vary depending on the specific case scenario and content design.

These results underscore the importance of adapting VP designs to Zambian needs and addressing the limitations observed in the appendicitis case before making broad statements regarding their comparability with other learning methods. When appropriately tailored to local contexts, VPs can function as an engaging and interactive learning strategy to enhance web-based or blended learning programs in settings with a high student-to-faculty ratio and limited teaching resources.

Nonetheless, further research on the acceptability and effectiveness of VPs is warranted, as the incorporation of additional VP medical topics into the blended learning program at LMMU for BSc clinical science students is planned. Expanding the evidence base will ensure that VPs continue to contribute positively to medical education in low-resource settings, support the ongoing development and refinement of these learning resources, and address potential disparities in content and design that may impact their effectiveness.

Acknowledgments
This work was supported by the Else Kröner-Fresenius-Stiftung (2019_HA25). For the publication fee, the authors acknowledge financial support from the Deutsche Forschungsgemeinschaft within the funding program “Open Access Publikationskosten” as well as from the Heidelberg University. The authors would like to thank Engagement Global GmbH for their support within the framework of the ASA program.

Data Availability
The data sets generated and analyzed during this study are not publicly available because of data protection regulations but are available from the corresponding author upon reasonable request.

Conflicts of Interest
OC was working for AMBOSS during the study but he was not involved in the data analysis nor did his affiliation influence any study outcomes.

Multimedia Appendix 1
CONSORT eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 97 KB - mededu_v9i1e43699_app1.pdf ]
Multimedia Appendix 2
Virtual patient medical topic: severe acute malnutrition.
[DOCX File, 34975 KB - mededu_v9i1e43699_app2.docx ]

Multimedia Appendix 3
Virtual patient medical topic: appendicitis.
[DOCX File, 14053 KB - mededu_v9i1e43699_app3.docx ]

Multimedia Appendix 4
Acceptance questionnaire.
[PDF File (Adobe PDF File), 82 KB - mededu_v9i1e43699_app4.pdf ]

Multimedia Appendix 5
Knowledge pre- and posttests.
[DOCX File, 25 KB - mededu_v9i1e43699_app5.docx ]

Multimedia Appendix 6
Individual learning trajectories of study participants in the context of severe acute malnutrition.
[PNG File, 119 KB - mededu_v9i1e43699_app6.png ]

Multimedia Appendix 7
Individual learning trajectories of study participants in the context of appendicitis.
[PNG File, 96 KB - mededu_v9i1e43699_app7.png ]

References
1. Physicians (per 1,000 people). The World Bank. URL: https://data.worldbank.org/indicator/SH.MED.PHYS.ZS [accessed 2023-03-17]


**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSc</td>
<td>Bachelor of Science</td>
</tr>
<tr>
<td>CONSORT</td>
<td>Consolidated Standards of Reporting Trials</td>
</tr>
<tr>
<td>HCW</td>
<td>health care worker</td>
</tr>
<tr>
<td>LMMU</td>
<td>Levy Mwanawasa Medical University</td>
</tr>
<tr>
<td>MCQ</td>
<td>multiple-choice question</td>
</tr>
<tr>
<td>MLP</td>
<td>medical licentiate practitioner</td>
</tr>
<tr>
<td>SAM</td>
<td>severe acute malnutrition</td>
</tr>
<tr>
<td>VP</td>
<td>virtual patient</td>
</tr>
</tbody>
</table>
Review

Health Literacy in Health Professionals Two Years into the COVID-19 Pandemic: Results From a Scoping Review

Eva-Maria Grepmeier¹, MA; Maja Pawellek², MSc; Janina Curbach¹, PhD; Julia von Sommoggy⁴, PhD; Karl Philipp Drewitz¹, MPH; Claudia Hasenpusch¹, MSc; Eva Maria Bitzer⁵, PhD; Christian Apfelbacher¹, PhD; Uwe Mattern², PhD

¹Institute of Social Medicine and Health Systems Research, Otto von Guericke University Magdeburg, Magdeburg, Germany
²University Children’s Hospital Regensburg, University of Regensburg, Regensburg, Germany
³Department of Business Studies, Ostbayerische Technische Hochschule Regensburg, Regensburg, Germany
⁴Medical Sociology, Institute for Epidemiology and Preventive Medicine, University of Regensburg, Regensburg, Germany
⁵Department of Public Health and Health Education, Freiburg University of Education, Freiburg, Germany

Corresponding Author:
Eva-Maria Grepmeier, MA
Institute of Social Medicine and Health Systems Research
Otto von Guericke University Magdeburg
Leipziger Str 44
Magdeburg, 39120
Germany
Phone: 49 39124300
Email: eva-maria.grepmeier@med.ovgu.de

Abstract

Background: Health literacy (HL) is an important public health goal but also crucial in individuals providing medical care. During the pandemic, COVID-19–related HL of health professionals (HPs) has gained momentum; it helps to minimize the risk of self-infection, on the one hand, and to protect patients and relatives from infection, on the other. However, comprehensive information about the levels of individual pandemic-related HL in HPs is scarce.

Objective: In this paper, we aimed at describing the extent of existing research on HL (concept) conducted in HPs (population) in the COVID-19 pandemic (context). The review intends to map the literature on HL in HPs, thereby highlighting research gaps.

Methods: This scoping review was conducted using the methodology of Khalil et al (2016). This involved an electronic search of PubMed (MEDLINE) and PsycInfo and a hand search. The included studies were iteratively examined to find items representing the four HL dimensions of access, understand, critically appraise, and apply COVID-19–related health information.

Results: The search yielded a total of 3875 references. Only 7 (1.4%) of the 489 included studies explicitly stated to have addressed HL; 2 (0.4%) studies attempted to develop an instrument measuring COVID-19–related HL in HPs; 6 (1.2%) studies included an HL measure in an observational survey design. Of the remainder, the vast majority used a cross-sectional design. The dimensions access and understand were frequently examined, but few studies looked at the dimensions critical appraisal or apply. Very few studies reported an intervention aiming to improve a COVID-19–related HL outcome.

Conclusions: High levels of COVID-19–related HL among HPs are necessary to ensure not only safe practice with necessary protection of HPs, their patients, and relatives, but also successful care delivery and subsequently improved health outcomes in the long term. To advance our understanding of how high COVID-19–related HL manifests itself in HPs, how it relates to health outcomes, and how it can be improved, more research is necessary.

Trial Registration: Open Science Framework dbfa5; https://osf.io/dbfa5/

(JMIR Med Educ 2022;8(4):e39023) doi:10.2196/39023

KEYWORDS

SARS-CoV-2; COVID-19; health competence; COVID-19–related health literacy; health care worker

https://mededu.jmir.org/2022/4/e39023

JMIR Med Educ 2022 | vol. 8 | iss. 4 | e39023 | p.15

(page number not for citation purposes)
Introduction

Background

Since late 2019, the world has been challenged by a new coronavirus, SARS-CoV-2. Besides its health, economic, social, and psychological impact [1], the pandemic has posed unprecedented challenges, particularly for health professionals (HPs) [2]. Many HPs are in a particularly exposed position during a pandemic [2]. Being in direct contact with COVID-19–infected patients in intensive care units or COVID-19 wards, or as general or specialist practitioners continuing to provide a safe service to patients, requires adaptation to new daily routines and workloads. HPs need to provide care for infected patients, continue to provide the necessary care for noninfected patients, and make sure not to infect patients nor themselves, their family, or significant others. In addition, HPs have a major societal responsibility to stop or mitigate the spread of the pandemic. They are not only required to provide health care services to patients [3] in their actions, but they must also consider their own, their patients’, and their family’s health [4]. HPs represent a population of individuals who are at an increased risk of infection due to the setting in which they work, and simultaneously they could pose a high risk to others due to high frequency of contacts. Therefore, their pandemic-related behavior is crucial to protect themselves and others from infection.

While guidance on how to organize these routines is provided by governmental policies and professional organizations, HPs may still face difficulty in meeting the many new demands placed on them during a pandemic [4]. HPs may also feel at the core of a dilemma. While encountering the ethical responsibility and moral obligation to spend time in places and situations where infection is more likely, they may also feel the concurrent need to protect themselves from infection [5,6]. The majority is willing to go to work [7,8], but the decision appears to be influenced by the preparedness of the organization [9] and other factors [8]. Sufficient availability of evidence-based protective measures, including personal protective equipment (PPE), is at the core of a health care facility’s preparedness [10].

In all these and many other scenarios, the concept of health literacy (HL) can be considered as a key aspect for HPs’ ability to adequately deal with a pandemic’s ubiquitous demands and challenges. HL “represents the cognitive and social skills which determine the motivation and ability of individuals to gain access to, understand and use information in ways which promote and maintain good health” [11]. It can also be understood to entail “the motivation, knowledge and competencies to access, understand, appraise and apply health information in order to make judgements and take decisions in everyday life concerning health care, disease prevention and health promotion to maintain or improve quality of life throughout the course of life” [12]. COVID-19–related HL can be understood as the level or extent of knowledge, motivation, and abilities of individuals to find, understand, and appraise pandemic-related health information and apply the results when making COVID-19–related health decisions. This includes, for example, knowledge about the application of measures to prevent COVID-19 infections, including vaccination-related aspects, detecting infections at an early stage (eg, through regular testing), and seeking medical assistance in case of a positive test or symptoms. Especially in a pandemic situation, which the world has been facing since 2019, one is dealing with a very rapidly changing evidence landscape. This makes it more important to find out what skills people working in the health sector have in terms of information access and understanding, information appraisal, and application.

Often, it is the HL of the general population that matters, because it has been established that HL is not only associated with a range of health outcomes [13,14] but also with social determinants of health. Large parts of the population report difficulties in accessing, understanding, appraising, and applying general health information [15-17], which is an important point, especially considering that compliance to infection prevention measures by each individual is critical in mitigating pandemics. However, comprehensive information about the levels of individual pandemic–related HL in HPs is scarce.

Objectives

It was the aim of this scoping review to describe the extent of research on HL (concept), conducted in HPs (population) in the COVID-19 pandemic (context). The review intends to map the literature on HL in HPs, thereby highlighting research gaps.

Methods

Overview

This scoping review was performed according to the methodological framework as put forward by Khalil et al [18]. As a first step, goal and research question of the scoping review were predefined. Second, in identifying the relevant studies, adjustments to the framework were made; for instance, unlike what has been recommended by Khalil et al [18], no search was performed for gray literature sources for practical and economical research efficiency reasons. PubMed (MEDLINE) and PsycINFO searches were performed by 1 author (UM) on January 20, 2022. Third, the studies were carefully screened in a five-stage procedure by the team of researchers: (1) abstracts were screened in dyads, irrelevant studies were excluded, and duplicates were removed (UM, CH, MP, KPD, JC, EG, and JvS); (2) full texts were screened; (3) further studies were excluded (EG, UM, and CH; Figure 1) and (4) categorized; and finally, (5) the results were collated.

The protocol for this review was registered at OSF Registries on February 19, 2022 [19]. Reporting in this scoping review follows the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist (Multimedia Appendix 1) [20]. The checklist contains 20 essential and 2 optional items, following what has been recommended by Khalil et al [18], but they must also consider their own, their patients’, and their family’s health [4]. HPs represent a population of individuals who are at an increased risk of infection due to the setting in which they work, and simultaneously they could pose a high risk to others due to high frequency of contacts. Therefore, their pandemic-related behavior is crucial to protect themselves and others from infection.
Eligibility
To identify studies for this review, we used the Population, Concept and Context framework by the Joanna Briggs Institute for scoping reviews [21]. The inclusion criteria should meet the Population, Concept and Context framework as follows.

Population
This review will include studies that focus on HPs. Only studies conducted with licensed or registered and practicing HPs were included. Studies exclusively conducted in students, trainees, and non–health care professionals or in the general population were excluded.

Concept
The concept of interest for this scoping review is health literacy, including the dimensions of access, understand, critically appraise, and apply COVID-19–related health information. We excluded studies focusing on knowledge about professional techniques and methods as well as studies focusing on mental health. Mental health and mental health literacy were outside the scope of this review.

Context
The context of this review is the COVID-19 pandemic. All studies from the onset of the COVID-19 (December 2019) pandemic reporting on any dimensional level or facet of COVID-19–related HL with or without explicitly referring to HL and conducted in the context of the COVID-19 pandemic were included. Studies published in peer-reviewed journals and written in English or German language were included. This was a pragmatic decision reflecting the author team’s language proficiency. Moreover, restricting to English language publications appears to have little influence on the introduction of bias in reviews [22,23]. No restrictions were applied regarding study design. Narrative reviews, books or chapters, commentaries, or prefaces were excluded.

Data Sources and Search Strategy
One of the authors (UM) conducted a search in PubMed (MEDLINE) and PsycINFO on January 20, 2022. All citations were downloaded to Citavi (Swiss Academic Software), and duplicates were removed. The search terms are reported in Multimedia Appendix 2.

Study Selection
UM, CH, MP, KPD, JC, EG, and JvS independently screened the titles and abstracts for inclusion in dyads. Full texts of the short-listed articles were obtained and independently reviewed in duplicate by 3 authors (EG, UM, and CH), and studies not meeting the inclusion criteria were excluded. Title and abstract screening as well as full-text screening took place in Rayyan (Rayyan Systems Inc) [24].

Data Collection Process
Three authors (EG, CH, and UM) independently extracted data from the included studies using a pretested data extraction form in Microsoft Excel. Consensus was achieved through discussions and arbitration within the review team.
Data Items

The included studies were iteratively examined to find items representing the 4 HL dimensions of *access*, *understand*, critically appraise, and apply COVID-19–related health information. All authors participated in this process, and findings were discussed until consensus was reached.

Ultimately, a total of 10 categories were developed based on these items, which are presented in the results section. In addition, data on study design were extracted.

Synthesis of Results

Findings were synthesized descriptively and narratively to provide a systematic classification of HL dimensions studied. Furthermore, we provide tables including frequency counts wherever possible. We did not conduct a critical appraisal of the included studies, because it was not within the scope of this review.

Results

Selection

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram shown in Figure 1 describes the study selection process. The electronic searches in PubMed (MEDLINE) and PsycINFO and the hand search yielded 3875 references. Following removal of duplicates and the title and abstract screening, 946 full-text articles were assessed for eligibility. A further 457 full-text articles were excluded with reasons, leading to 489 studies finally included in this scoping review. The list of included studies can be found in Multimedia Appendix 3.

Population

The population included in the studies encompassed HPs such as nurses, pharmacists, community health workers, and medical doctors of most specializations. A clear distinction between the specific HPs studied and the setting in which data collection took place was often not made. Of the 489 included studies, 277 (56.6%) studied HPs, 86 (17.6%) were conducted in dental settings, 45 (9.2%) among nurses, 30 (6.1%) among pharmacy settings, 28 (5.7%) among medical doctors (eg, physicians, surgeons, pediatricians, and general practitioners), the rest (n=22, 4.5%) among a variety of settings such as community health work; emergency medical services; intensive care unit; ear, nose, and throat care; eye care; radiology; and physiotherapy. Only 1 (0.2%) study referred to HPs in a residential home setting.

Types of Studies

Most of the included studies had used a cross-sectional design. Very few studies reported an intervention aiming to improve a COVID-19–related HL outcome. Of these, only 1 was conducted as a randomized controlled trial; all others used a pre-post design (Table 1). Interventions aimed at improving infection prevention and control knowledge or competencies such as proper PPE application or hygiene measures.

Table 1. Study design used in the included studies (N=489).

<table>
<thead>
<tr>
<th>Study design</th>
<th>Frequency, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-sectional</td>
<td>415 (85)</td>
</tr>
<tr>
<td>Longitudinal</td>
<td>7 (1.4)</td>
</tr>
<tr>
<td>Interventional</td>
<td>27 (5.5)</td>
</tr>
<tr>
<td>Systematic review</td>
<td>9 (1.8)</td>
</tr>
<tr>
<td>Narrative review</td>
<td>1 (0.2)</td>
</tr>
<tr>
<td>Qualitative</td>
<td>18 (3.7)</td>
</tr>
<tr>
<td>Mixed method</td>
<td>4 (0.8)</td>
</tr>
<tr>
<td>Other</td>
<td>8 (1.6)</td>
</tr>
</tbody>
</table>

Concept and Dimensions of HL

Concept of the scoping review was HL. The authors developed the following categories of HL, based on the four HL dimensions; these categories are as follows: (1) HL (objective versus subjective), (2) sources of COVID-19 information, (3) knowledge (objective versus subjective), (4) ability to understand COVID-19–related information, (5) critically evaluate COVID-19–related information, (6) perceived skills or confidence and perceived preparedness in applying COVID-19–related information, (7) development of educational resource to improve any HL dimension, (8) reported receipt of infection control–related health education training, (9) COVID-19–related HL instrument development, and (10) interventions to improve any of the four HL dimensions.

Almost no included study explicitly referred to or introduced the concept of HL (n=482, 99%). However, all examined studies at least implicitly mentioned one dimension of COVID-19–related HL (Table 2). Of the 489 included studies, the HL dimension of *access* information was investigated by 191 (39.1%) studies; the HL dimension of *understand* information was represented by 434 (88.8%) studies included by a measure of COVID-19–related knowledge; the HL dimension of critically appraise information was only examined in 1 (0.2%) study. Moreover, 59 (12.07%) studies measured the HL dimension ability to apply COVID-19–related information. The HL dimension apply was reviewed using the two categories perceived skills or confidence (n=28, 5.7%) and perceived preparedness (n=31, 6.3%).
COVID-19 knowledge was most frequently examined as objectively measured knowledge. Fewer studies assessed knowledge subjectively, and only 31 (6.3%) studies measured knowledge complementarily by objective and subjective items (Table 3).

Of the 489 studies, 14 (2.9%) reported on the development of an educational training resource to increase (inferred) HL facets. In 148 (30.3%) studies, a measure of reported receipt of infection prevention and control-related training was assessed. Only 7 (1.4%) studies explicitly stated to have addressed HL (Table 4). As the studies presented below show, HL has many possible fields of application.

Table 2. Health literacy (HL) dimensions implicitly examined in the included studies (N=489).

<table>
<thead>
<tr>
<th>HL dimensions</th>
<th>Frequency, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access</td>
<td></td>
</tr>
<tr>
<td>Sources of information</td>
<td>191 (39.1)</td>
</tr>
<tr>
<td>Understand</td>
<td></td>
</tr>
<tr>
<td>Knowledge (any)</td>
<td>434 (88.7)</td>
</tr>
<tr>
<td>Other than knowledge</td>
<td>13 (2.7)</td>
</tr>
<tr>
<td>Critically appraise</td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td>1 (&lt;1)</td>
</tr>
<tr>
<td>Apply</td>
<td></td>
</tr>
<tr>
<td>Perceived skills or confidence</td>
<td>28 (5.7)</td>
</tr>
<tr>
<td>Perceived prepared-ness</td>
<td>31 (6.3)</td>
</tr>
</tbody>
</table>

Multiple entries possible; hence, numbers do not add up to 489 or 100%.

Table 3. Type of knowledge assessment in the included studies (n=434).

<table>
<thead>
<tr>
<th>Knowledge assessment</th>
<th>Value, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjective (perceived) knowledge</td>
<td>81 (18.7)</td>
</tr>
<tr>
<td>Objective knowledge</td>
<td>280 (64.5)</td>
</tr>
<tr>
<td>Subjective (perceived) and objective knowledge</td>
<td>31 (7.1)</td>
</tr>
<tr>
<td>Knowledge unclear</td>
<td>42 (9.7)</td>
</tr>
</tbody>
</table>

No multiple entries.

Table 4. Studies explicitly referring to health literacy (HL).

<table>
<thead>
<tr>
<th>Study</th>
<th>Study design or objective</th>
<th>Type of HL</th>
<th>Instrument used</th>
<th>Subjectively or objectively assessed</th>
<th>Validated instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alam et al [25], 2021</td>
<td>Cross-sectional survey</td>
<td>Vaccine literacy</td>
<td>Author developed HLS-SF12, eHEALS</td>
<td>Unclear</td>
<td>No</td>
</tr>
<tr>
<td>Do et al [26], 2020</td>
<td>Cross-sectional survey</td>
<td>General HL; digital HL</td>
<td>Author developed</td>
<td>Subjective</td>
<td>Yes; partly</td>
</tr>
<tr>
<td>Fatteh et al [27], 2022</td>
<td>Cross-sectional survey</td>
<td>COVID-19 HL</td>
<td>Author developed</td>
<td>Subjective and objective</td>
<td>No</td>
</tr>
<tr>
<td>Hara et al [28], 2021</td>
<td>Cross-sectional survey</td>
<td>Vaccine literacy</td>
<td>Author developed</td>
<td>Subjective</td>
<td>Partly</td>
</tr>
<tr>
<td>Heiniger et al [29], 2021a</td>
<td>Instrument development</td>
<td>Hygiene competence</td>
<td>HygiKo; author developed</td>
<td>Objective</td>
<td>Yes</td>
</tr>
<tr>
<td>Hiltrop et al [30], 2021</td>
<td>Instrument development</td>
<td>COVID-19 HL</td>
<td>HL-COV-HP; author developed</td>
<td>Subjective</td>
<td>Partly b</td>
</tr>
<tr>
<td>Nahidi et al [31], 2021</td>
<td>Cross-sectional survey</td>
<td>COVID-19 HL</td>
<td>Author developed</td>
<td>Subjective</td>
<td>Content validated</td>
</tr>
</tbody>
</table>

Referring to objectively assessed competence.
Exploratory and confirmatory analyses conducted in same sample.
HL referred to in Discussion.
Measurement of COVID-19–Related Health Literacy

A total of 2 (0.4%) studies attempted to develop an instrument, measuring COVID-19–related HL in HPs. Heininger et al [29] developed an objective test, the situational judgement test, HygiKo, to assess hygiene competence. It comprises 20 picture vignettes. Each vignette shows at least one HP and a patient in clinical situations in which hygiene is a pertinent subject. Item-response analyses demonstrated that HygiKo is appropriate for assessing hygiene competence and that it allows distinguishing between persons demonstrating different levels of ability.

Another study [30] developed HLS-COV-HP to measure subjective COVID-19–related HL in HPs. It was adapted from the HLS-EU-Q16 and contains in its present form 12 items to assess the perceived motivation and ability of HPs to find, understand, evaluate, and use COVID-19 information. However, exploratory and confirmatory analyses were performed using the same sample.

A total of 7 (1.4%) studies included an HL measure in an observational survey design. Alam et al [25] examined the motivation to receive a COVID-19 vaccination using a cross-sectional survey design. They also assessed vaccine literacy (VL) by 6 questions from a self-report questionnaire. VL levels were found to differ as a function of gender, age, occupation, or type of organization. The relationship with vaccination motivation was not examined. Do et al [26] evaluated the psychometric properties of an instrument measuring digital HL (eHEALS) and examined associations of subjective general and digital HL with adherence to infection prevention and control procedures among other constructs by conducting a cross-sectional survey in HPs. They found a positive relationship between both general and digital HL, on the one hand, and adherence to infection prevention and control procedures, on the other.

Fatteh et al [27] administered a self-developed questionnaire measuring subjective and objective aspects of COVID-19–related HL to the workforce of a large medical center. They found a positive relationship between medical education level and COVID-19–related HL.

Hara et al [28] conducted a cross-sectional survey in HP and the general population assessing VL and vaccine hesitancy. HPs were found to have higher levels of VL compared with the general population, but the levels of vaccine hesitancy were similar between the groups.

Nahidi et al [31] conducted a cross-sectional survey in critical care nurses. They assessed the ease or difficulty of knowledge acquisition across 11 key information areas of COVID-19, such as the use of PPE, infection prevention and control, or signs and symptoms. Most participants reported a “good” to “very good” level of knowledge about COVID-19 and obtained up-to-date COVID-19 information from a variety of credible sources.

Discussion

Overview

This scoping review provides a summary of research on COVID-19–related HL conducted during the COVID-19 pandemic in HPs. HL is considered a key competence to protect oneself, one’s patients, but also one’s relatives from potential COVID-19 infection; it also entails competencies regarding vaccination-related aspects, detecting infections at an early stage (eg, through regular testing) and seeking medical assistance in case of a positive test or symptoms. Definitions of health literacy by Nutbeam [11] and Sorensen [12] were used to guide the conduct of this scoping review. For a scoping review focusing on health literacy in the context of the pandemic, it seems appropriate to present in more detail those studies that address HL explicitly as a concept. Thus, in the context of this review, we also present a fundus of what an exemplary engagement with HL might explicitly look like. Our results suggest that HL in HPs during a pandemic has rarely been studied in light of a theoretically founded framework of HL. However, a large body of studies measured variables subsumed to be HL dimensions without explicitly referring to HL as a theoretical construct.

Principal Results

A comprehensive literature search identified 489 studies having examined COVID-19–related HL in HPs. The vast majority, while examining at least one HL dimension, had not intended to study HL as a distinct construct. Of the included studies, only 7 (1.4%) studies explicitly addressed HL. More specifically, 3 (0.6%) studies directly addressed COVID-19–related HL [27,30,31], 2 (0.4%) studies examined general HL in the context of the COVID-19 pandemic [26,32], and another 2 (0.4%) examined vaccination-related HL [25,28]. Digital HL was examined in 1 (0.2%) [26] and hygiene competence in another (n=1, 0.2%) study [29].

Although the overall body of identified studies was heterogeneous, most reviewed studies used a cross-sectional observational survey design to assess among other constructs subjective or objective knowledge related to COVID-19. Because respondents often overestimate their levels of knowledge, competence, or abilities when assessed by subjective self-report [29], it is noteworthy that most of these studies assessed objective knowledge. Some authors [33,34] recommend a complementary assessment of subjective and objective HL because subjective HL should be considered as a separate concept from objective HL [35]. Few studies assessed the HL dimension understand by other means beside knowledge. Our analysis found that the HL dimension access or find information was only represented by the measurement of sources of information. We do not think this can be considered a sufficient approach to measuring this HL dimension. Sources of information was the second most frequently reported dimension in the included studies.

In contrast to the many studies reporting on the HL dimensions access and understand information, only 1 study reported on the critical appraisal of COVID-19–related information in HPs. Though working in a health care environment, HPs may also be exposed to conflicting information and misinformation.
Critical HL is crucial in individuals’ ability to distinguish fact from fake [36].

Approximately 20% of the included studies examined a measure of the HL dimension apply COVID-19–related information, by for instance, perceived skills or confidence and perceived preparedness regarding use of infection prevention and control-related measures. It is stressed that the HL dimension apply does not pertain to the intention or motivation to enact protective behavior nor the actual behavior itself. Within the HL framework, apply (like all other dimensions) is a reflection of the mere abilities or competences rather than the realizations or manifestations of these abilities.

While we came across interventions aiming to improve infection prevention and control-related knowledge and abilities, we found no studies trying to improve vaccination-related knowledge or competencies.

Limitations

In this scoping review, our systematic search was limited to 2 major databases, and no gray-literature search was conducted. Although it is generally recommended to include gray literature in a scoping review, we decided against including all possible sources for practical and economical research efficiency reasons: The COVID-19 pandemic led to an unprecedented high-speed publication of a large body of scientific literature, both peer-reviewed as well as gray, and to handle this large volume of publications would have required more resources.

Owing to the large number of included studies, we refrained from reviewing the included literature more thoroughly, for instance, regarding quality assessment. This reflects the nature of a scoping review, which is intended to provide a summary of the state of research without addressing the quality of individual studies, according to the standard guidelines for observational or intervention studies.

A more detailed categorization and charting of the HL dimensions may have been beneficial. For instance, we would have liked to provide a more elaborate analysis on the type of COVID-19–related knowledge as knowledge could refer to transmission, course, symptoms, or the prevention of COVID-19.

Future Research

The observed paucity of research in HPs applying empirically developed HL formulations to pandemic contexts calls for future research. From the current review, many questions remain unanswered. While all areas of HPs’ working environments appeared to have provided studies for our review, it is surprising that only 1 study was conducted among staff in nursing or residential homes. In many countries, these facilities were the ones most hard hit by COVID-19 and should thus be considered more strongly in future research [37].

Our scoping review also revealed that there is a need to use more comprehensive approaches to the measurement of HL dimensions. Altogether, most studies provided very little evidence about the psychometric properties of the used instruments (results not shown). We identified 2 instruments for COVID-19–related HL assessment in HPs, but further validation and refinement appears necessary. There is also a need for instruments objectively measuring a broader range of COVID-19–related HL dimensions in HPs.

Investigations aiming to assess change in HL over time, for instance by repeated surveys attempting to monitor HL levels over the course of the pandemic in HPs, would also be desirable. As there is a need to conduct more robust experimental studies to examine the effectiveness of HL interventions among HPs, such instruments could be used to examine long-term effects of these interventions.

It would probably be profitable for future research to provide more comprehensive reviews, including gray literature and larger bodies of literature by searching more than 2 databases.

Conclusions

Based on the existing literature on HL in general and related to other health issues, we assume that high levels of COVID-19–related HL among HPs are necessary to ensure not only safe practice with necessary protection of HPs, their patients, and relatives but also successful care delivery. Subsequently, health outcomes may be improved in the long term.

To advance our understanding of how high COVID-19–related HL manifests itself in HPs, how it relates to health outcomes, and how it can be improved, more research is necessary.

Acknowledgments

This scoping review was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—409800133.

Authors’ Contributions

CA and UM conceptualized and designed the study; UM searched the literature; EMG, CH, UM, JvS, KPD, and MP screened the literature; UM, EMG, and CH extracted data; EMG drafted the initial manuscript and reviewed and revised the manuscript with contributions from EMB, UM, MP, JvS, CH, JC, KPD, and CA. All authors approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.
References


34. Mayer A. Subjective and objective measures of health literacy: Do they provide complementary or redundant information? In: Information Literacy in the Workplace. Cham, Switzerland: Springer International Publishing; 2018.


Abbreviations

**HL:** health literacy
**HP:** health professional
PPE: personal protective equipment
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
VL: vaccine literacy
Review

Effectiveness of Shared Decision-making Training Programs for Health Care Professionals Using Reflexivity Strategies: Secondary Analysis of a Systematic Review

Ndèye Thiab Diouf¹,²,³, MSc, MPH; Angèle Musabyimana¹,²,³, MPH, MD; Virginie Blanchette¹,²,⁴, MSc, DPM, PhD; Johanie Lépine¹,², PhD; Sabrina Guay-Bélanger¹,², PhD; Marie-Claude Tremblay²,⁵,⁶, PhD; Maman Joyce Dogba²,⁵,⁶, PhD; France Légaré¹,²,⁶, BSc, MSc, MD, PhD, CCFP

¹Canada Research Chair in Shared Decision Making and Knowledge Translation (Tier 1), Quebec, QC, Canada
²VITAM - Centre de recherche en santé durable, Centre intégré universitaire de santé et services sociaux de la Capitale-Nationale, Quebec, QC, Canada
³Department of Community Health, Faculty of Nursing and Faculty of Medicine, Université Laval, Quebec, QC, Canada
⁴Department of Human Kinetic and Podiatric Medicine, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada
⁵Office of Education and Continuing Professional Education, Université Laval, Quebec, QC, Canada
⁶Department of Family Medicine and Emergency Medicine, Faculty of Medicine, Université Laval, Quebec, QC, Canada

Corresponding Author:
France Légaré, BSc, MSc, MD, PhD, CCFP
VITAM - Centre de recherche en santé durable
Centre intégré universitaire de santé et services sociaux de la Capitale-Nationale
2480 chemin de la canardièr Quebeck, QC, G1G 2G1
Canada
Phone: 1 418 663 5313
Email: france.legare@fmed.ulaval.ca

Abstract

Background: Shared decision-making (SDM) leads to better health care processes through collaboration between health care professionals and patients. Training is recognized as a promising intervention to foster SDM by health care professionals. However, the most effective training type is still unclear. Reflexivity is an exercise that leads health care professionals to question their own values to better consider patient values and support patients while least influencing their decisions. Training that uses reflexivity strategies could motivate them to engage in SDM and be more open to diversity.

Objective: In this secondary analysis of a 2018 Cochrane review of interventions for improving SDM by health care professionals, we aimed to identify SDM training programs that included reflexivity strategies and were assessed as effective. In addition, we aimed to explore whether further factors can be associated with or enhance their effectiveness.

Methods: From the Cochrane review, we first extracted training programs targeting health care professionals. Second, we developed a grid to help identify training programs that used reflexivity strategies. Third, those identified were further categorized according to the type of strategy used. At each step, we identified the proportion of programs that were classified as effective by the Cochrane review (2018) so that we could compare their effectiveness. In addition, we wanted to see whether effectiveness was similar between programs using peer-to-peer group learning and those with an interprofessional orientation. Finally, the Cochrane review selected programs that were evaluated using patient-reported or observer-reported outcome measurements. We examined which of these measurements was most often used in effective training programs.

Results: Of the 31 training programs extracted, 24 (77%) were interactive, among which 10 (42%) were considered effective. Of these 31 programs, 7 (23%) were unidirectional, among which 1 (14%) was considered effective. Of the 24 interactive programs, 7 (29%) included reflexivity strategies. Of the 7 training programs with reflexivity strategies, 5 (71%) used a peer-to-peer group learning strategy, among which 3 (60%) were effective; the other 2 (29%) used a self-appraisal individual learning strategy, neither of which was effective. Of the 31 training programs extracted, 5 (16%) programs had an interprofessional orientation, among which 3 (60%) were effective; the remaining 26 (84%) of the 31 programs were without interprofessional orientation, among which 8 (31%) were effective. Finally, 12 (39%) of 31 programs used observer-based measurements, among which more than half (7/12, 58%) were effective.
Conclusions: Our study is the first to evaluate the effectiveness of SDM training programs that include reflexivity strategies. Its conclusions open avenues for enriching future SDM training programs with reflexivity strategies. The grid developed to identify training programs that used reflexivity strategies, when further tested and validated, can guide future assessments of reflexivity components in SDM training.

(JMIR Med Educ 2022;8(4):e42033) doi:10.2196/42033

KEYWORDS
shared decision-making; reflexivity; training; health care professionals; implementation

Introduction

Background

There is increasing recognition of the ethical imperative to support patients to be engaged in their care, especially in health-related decisions. Shared decision-making (SDM) is a collaborative process whereby health care professionals support patients in making decisions that are informed by the best evidence and by what matters to them [1]. SDM improves the health care experiences of patients and health care professionals and leads to better health care processes, patient outcomes, and lower health costs [2-5]. SDM is the best practice for informed consent and is fundamental to patient- and family-centered care [6]. However, SDM has not yet been widely implemented in clinical practice because of several perceived barriers [7]. In some contexts, SDM implementation is encouraged by health policies, but certain challenges related to patients, patient–health care professional relationships, and organizational factors remain in the way of its concrete adoption [8].

Reflexivity

Training programs for health care professionals [9,10] are believed to be crucial to the implementation of SDM. However, SDM training programs for health care professionals are highly heterogeneous [10], and we still do not know what makes them effective [9]. One promising approach is reflexivity, which has been shown to increase health care professionals’ willingness to be more engaged in the health care offer and collaborate with other professionals [11]. SDM training programs that use reflexivity strategies have the potential to be more effective than those that do not [12].

Reflexivity is a form of learning based on reflection on one’s experiences applied in a professional or interprofessional context [13]. Concepts related to reflexivity are found in various disciplinary fields under different names. Writing about reflective practice (a practice based on reflexivity), Schon [14] states that the concept entails critical thinking concerning the actions and stances one takes [14,15]. In health care practices, being reflexive can mean different things: acknowledging and questioning the power dynamics implicit in a health care encounter; identifying the assumptions that underlie a health care situation; or examining the influences, such as values and beliefs, that shape health care practices [13].

On the basis of this multiplicity of definitions, Sandars [16] developed a guide that classifies them into 3 main approaches. According to Sandars’s work [16], reflection is a form of learning that is based on three common aims: (1) reflection for learning, (2) reflection to develop a therapeutic relationship, and (3) reflection to develop professional practice [14,15]. Indeed, reflexivity includes questioning the premises of an action, such as the values, norms, and beliefs that a professional may hold, as well as how such actors justify their actions [17]. In sum, reflexivity is a good strategy for motivating both health care professionals and patients to engage in patient care in a collaborative way.

Reflexivity and SDM

Reflexivity is appropriate in the context of SDM because the latter emphasizes a partnership between patients and clinicians in making decisions and establishing care plans. SDM aims to repose the knowledge of patients and clinicians on an equal footing, adjusting the asymmetrical power relationship between patients and health care professionals [18,19]. A prerequisite for implementing SDM in care settings is that health care professionals not only have the knowledge and skills but also the willingness to engage patients in the decision-making process [20].

Training based on reflexivity may lead health care professionals to question the power issues inherent in a more traditional conception of the health system and of their role and may motivate them to adopt SDM [11]. Thus, we hypothesized that SDM training programs that integrate reflexivity strategies would be more effective in increasing the adoption of SDM than those that do not.

Interprofessional Training

Research also shows that SDM training programs developed with an interprofessional orientation are to be encouraged [21]. Interprofessionality is defined by D’amour et al [22] as the development of a cohesive practice between professionals from different disciplines for the care of a single patient. Specifically, interprofessional collaboration involves collegial, authentic, constructive, open and honest communication as well as mutual trust and respect between professionals who are committed to achieving a common set of goals. However, although an interprofessional orientation is highly encouraged in SDM because of the need for better teamwork and professionals’ openness to other forms of practices, few training sessions with an interprofessional orientation are available [10]. In addition, there is little evidence that interprofessionally oriented training is more effective than other training approaches. Given the similarity between the methods used in training programs with an interprofessional orientation and those used in training programs with certain reflexivity strategies (eg, peer-to-peer group learning), it is important to examine how effective both these training types are.
**Outcome Reporting**

In the Cochrane review, one of the criteria for selecting programs was that the type of measurement followed must be a patient-reported outcome measurement (PROM) or an observer-reported outcome measurement (OBOM). A PROM is an instrument used to collect information directly from patients. PROMs do not require amendments or interpretation by a clinician or another observer [23]. An OBOM is any instrument used by a third-party observer to report observable concepts such as signs or behaviors to assess, for example, the decision-making process during an encounter between a patient and family and their health care professional when facing a health treatment or screening decisions [23]. It seemed useful to see what types of measurements were most common, especially among the effective training programs that used reflexivity strategies. This could offer another potential reason for a training program’s effectiveness. As we were interested in interventions targeting health care professionals, it was important to discern whether the type of measurement used to evaluate the intervention was equipped to examine the different levels of effectiveness as defined by Kirkpatrick, the creator of one of the most common evaluation models for assessing training programs targeting health care professionals [24]. The Kirkpatrick model evaluates training programs based on 4 categories: the satisfaction of the participants with the training, improvement of their knowledge, improvement of their care practices, and improvement of patient health. The first category is very important in the evaluation process because, according to Kirkpatrick, learner appreciation is a key factor in motivating participants to learn from training and apply what they learned.

Therefore, based on our analysis of the Cochrane review of interventions for increasing the use of SDM by health care professionals, we first sought to determine whether the SDM training programs for health care professionals that used reflexivity strategies were more frequently classified as effective than the training programs that did not. Second, if such programs were more effective, we aimed to explore the strategy (peer-to-peer group learning or self-appraisal individual learning) that seemed to be more effective. Third, we aimed to examine whether training programs with an interprofessional orientation tended to be evaluated as effective. Finally, we examined whether the type of measurement used (OBOM or PROM) made it possible to classify the training programs in terms of effectiveness.

**Methods**

**Study Design**

We performed a secondary analysis of a published 2018 Cochrane review that aimed to evaluate the effectiveness of interventions for increasing the use of SDM by health care professionals [7]. As there are no reporting guidelines for the secondary analyses of systematic reviews, we used the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 for all applicable items [25]. This study took place from February 2021 to April 2022.

**Data Sources and Search Strategy**

The search strategy for the Cochrane review serving as the basis of the current secondary analysis was launched on June 15, 2017. Details of the Cochrane search strategy and data sources can be found in the published review [7]. The Cochrane review [7] included interventions classified according to the three target categories of the Effective Practice and Organization of Care (EPOC) taxonomy of interventions [26]: (1) interventions targeting patients (eg, patient-mediated interventions), (2) interventions targeting health care professionals (eg, distribution of printed educational material, educational meetings, audit and feedback, reminders, and educational outreach visits), and (3) interventions targeting both patients and health care professionals (eg, a patient-mediated intervention combined with an intervention targeting health care professionals).

**Eligibility Criteria**

In the published Cochrane review, the participants in the training programs could be any type of health care professional (eg, physicians, nurses, pharmacists, or social workers), including professionals in training (eg, resident physicians). Studies that recruited eligible health care professionals along with other types of participants (eg, patients and managers) were also included, as were training programs evaluating a multicomponent intervention (eg, SDM training for health care professionals with the use of patient decision aids). All types of training formats were incorporated into the review (eg, in class, group workshop, web-based training, and synchronous or asynchronous training). A training program was defined as a capacity-building activity conducted live for a group or a single individual, such as a web-based course or a traditional course (ie, a course integrated into an academic program), that used a recognized instructional method such as lectures, workshops, case studies, demonstrations, role plays, and small group discussions [10]. There were no restrictions on the comparison groups, which were all included.

The same primary outcomes of interest reported in the Cochrane review, namely SDM outcomes, were maintained in this secondary analysis, as were the outcome measurements used [7]. The Cochrane review grouped secondary outcomes into 2 categories: patient outcomes (eg, affective-cognitive outcomes, behavioral outcomes, and health outcomes) and process outcomes (eg, consultation length, costs, and equity). The 4 eligible study designs were randomized controlled trials, nonrandomized controlled trials, controlled before-and-after studies, and interrupted time series.

In this secondary analysis, we selected all interventions targeting health care professionals from the Cochrane review (43/87, 49%), excluding interventions that only targeted patients (44/87, 51%). Of the 43 that targeted health care professionals, we excluded interventions that were not a training program (12/87, 14%), for example, demonstrations on how to use a decision aid. Therefore, the total number of articles included in this study was 31.

**Study Selection Process**

One of the reviewers (NTD) examined all the content available concerning the interventions used in each of the 87 articles.
included in the Cochrane review and selected interventions that met the inclusion criteria [7]. Another reviewer (JL) reviewed both the selected and excluded programs to validate the rigor of the selection process. Finally, all the articles involving an eligible training program targeting health care professionals were selected (N=31) (refer to the eligibility criteria section).

Data Extraction Process
Data extraction was performed by 2 pairs of independent reviewers (NTD and AM or NTD and VB). Differences between the 2 reviewers were resolved by consensus based on discussion and by referring to the definitions provided in the extraction grid (Multimedia Appendix 1). For the remaining conflicts, a third reviewer (SG-B or MCT) intervened to facilitate a consensus. The articles believed to include reflexivity were submitted to MCT for validation. Data extracted included (1) article and study characteristics—year, country and language of publication, and measurement of study design and type; (2) information on the training program—country, training language, context of care, type of health care professional trained, and training format (eg, unidirectional or interactive); and (3) the use or otherwise of reflexivity strategies for interactive training programs and, among those including reflexivity strategies, the type of strategy selected (peer-to-peer reflective group learning or self-appraisal learning).

Classification of Articles
Classification of Training Formats
First, 2 reviewers (NTD and AM or NTD and VB) classified all the included articles (n=31) according to training format (unidirectional or interactive). The interactive training programs were then classified into 2 groups: programs using reflexivity strategies and programs that do not (refer to the details given in the Reflexivity Strategies Assessment section). On the basis of how reflexivity is defined in the literature, in this analysis, only interactive training programs were considered to have the potential to involve reflexivity. The training programs classified here as unidirectional were those in which the trainer delivered the whole message without asking for learner input other than questions, whereas interactive training programs are delivered in a 2-way manner, requiring the active contribution of learners (ie, the trainer delivers the information to the learners and encourages them to contribute to an exchange process). Interactive training by its very nature has the potential for reflexivity through, for example, role play or case discussions. For the purposes of this study, training programs using reflexivity strategies involve at least a minimal contribution from learners in the reflection process. Second, training programs classified as including reflexivity strategies were also categorized into 2 further groups (peer-to-peer group learning or self-appraisal individual learning) by 2 reviewers (NTD and AM or NTD and VB, validated by MCT).

Reflexivity Strategies Assessment
To the best of our knowledge, there is no validated set of criteria that defines the minimal components required to qualify a training program as reflexivity based. Thus, we developed a grid informed by a preliminary rapid literature review that synthesized the most common approaches and concepts related to reflexivity strategies used in health care professional training [27]. The grid contains minimal criteria that a training program must meet to be considered as including reflexivity strategies. The 2 following questions (A and B) from our grid were used to assess whether an interactive training program incorporated reflexivity strategies.

(A) Does the Training Program Include Any Reflexivity Approaches?
A reflexivity training approach could be, but is not limited to, the following: group-based reflections with peers (with or without a trainer), self-competence improvement with case-based reflections, electronic platforms with reflective portfolios, reflective journals, Balint groups, on-site reflective writing exercises, and the like. When these approaches were not clearly specified, we looked for the common reflexivity concepts.

(B) Does the Training Program Include Any Common Reflexivity Concepts?
The following were considered reflexivity concepts: critical thinking, metacognition, self-reflection, reflective dialogue, reflection-in-action, reflection-in-practice, reflection-on-action, reflection-on-practice, reflective practice, reflective learning, reflective approaches, reflective dialogues, critical self-reflection, reflective thinking, reflection on error, and the like.

Once we identified articles that included reflexivity strategies based on questions A (presence of reflexivity approaches) and B (presence of reflexivity concepts), we further subcategorized them according to 2 types of strategies: peer-to-peer group learning or self-appraisal individual learning.

Peer-to-Peer Group Learning
The main objective of peer-to-peer group learning (small or large group) is to stimulate interaction between participants. In peer-to-peer groups, the participants learn from each other’s reflections while being supported by experienced trainers or facilitators. This strategy can be organized in different ways, for example, a few days of practice followed by a day of reflection among peers or presentation of a topic followed by a group reflection among professionals during which they discuss their practice experience. Various approaches such as reflective writing exercises or groups with colleagues can be incorporated. Reflections may be based on real cases (ie, cases seen in practice) or fictitious ones. Everyone is called upon to give their point of view, and lessons are learned as the reflection progresses [28]. During and at the end of the exercise, the trainer or facilitator reframes the interactions, guides the discussion, and corrects errors or discrepancies resulting from the reflection [28].

Self-appraisal Individual Learning
A self-appraisal individual learning strategy is any individual learning process in which the learner is subjected to reflection exercises or cases to be solved. The learner can also be questioned on their practice. The exercise might involve traps to allow learners to detect their own errors. This exercise is performed individually, for example, in a reflective journal
where professionals might write down reflections on their practice such as all the events (positive or negative) experienced, what these events meant for them, and what they learned from these experiences [29]. They may then reflect on how such experiences could help them in similar future circumstances. This exercise can also be performed during a group training session but where participants reflect individually on a case (fictitious or real) [30]. A self-appraisal individual learning strategy can also be applied during web-based courses. In a self-appraisal individual learning process, exercises such as self-reflection; reflective learning; reflection on one’s own values, beliefs, and thoughts; and metacognition are often used.

**Assessing Training Effectiveness**

We determined whether the included training programs were classified as effective in the Cochrane review [7]. The same outcomes of interest reported in the Cochrane review, that is, SDM outcomes, were maintained in this secondary analysis.

**Analysis**

First, unidirectional and interactive training programs were compared to see what percentage of each was classified as effective by the Cochrane review. Second, among the interactive programs, those that used reflexivity strategies were compared with those that did not to see which were more frequently effective. Third, reflexivity strategies (peer-to-peer group learning and self-appraisal individual learning) were compared. After these 3 steps, we carried out an additional analysis to see whether there were elements that could explain why a training program was effective or otherwise, apart from those cited earlier. For example, we compared the proportion of training programs based on interprofessional orientation that were effective with the proportion of programs using reflexivity that were effective. In addition, to see how SDM training programs can be better evaluated, we classified training programs according to the type of measurement they incorporated and whether they were classified as effective.

**Results**

**General Characteristics of Results**

**Main Characteristics of the Included Studies**

All the 87 studies included in the Cochrane review were evaluated [7]. Of these, 43 (49%) interventions targeted health care professionals, and 44 (51%) targeted only patients. Among the 43 interventions targeting health care professionals, 31 (72%) were found to be training programs (Figure 1) and were analyzed in this study.

**Figure 1.** Selection process of the included studies. HCP: health care professional.
Of the 31 articles included, 10 (32%) were published between 2002 and 2010 (the first publication of the Cochrane review) [31-40], 14 (45%) were published between 2011 and 2014 (the first update of the Cochrane review) [41-54], and the other 7 (23%) were published between 2015 and 2017 (the most recent update) [55-62]. A total of 11 (35%) included articles were published in the United States [38, 43, 46, 50, 51, 53-55, 58, 60, 61, 63], followed by 8 (26%) in Germany [3, 32, 35, 37, 39, 44, 47, 57, 64-66] (Table 1). All the studies included were published in English, and 29 (94%) of them were RCTs. For the primary outcome assessment, 15 (48%) of the 31 studies were evaluated using PROMs alone, 12 (39%) were evaluated using OBOMs, and 3 (10%) were evaluated using both. Information about the 1 (3%) study [35] that used health care professional–reported outcome measurements (HCPROMs) was directly collected from the article. Cochrane did not include outcomes measured by HCPROMs in its analysis [7]. However, seeing that this paper [35] was included in the analysis of some secondary outcomes in the Cochrane review, we examined the results directly from the article to analyze information related to our variable of interest (SDM).

Table 1. Characteristics of included studies (N=31).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Year of publication</strong></td>
<td></td>
</tr>
<tr>
<td>2002-2010</td>
<td>10 (32)</td>
</tr>
<tr>
<td>2011-2014</td>
<td>14 (45)</td>
</tr>
<tr>
<td>2015-2017</td>
<td>7 (23)</td>
</tr>
<tr>
<td><strong>Country</strong></td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>11 (35)</td>
</tr>
<tr>
<td>Germany</td>
<td>8 (26)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>4 (13)</td>
</tr>
<tr>
<td>Canada</td>
<td>3 (10)</td>
</tr>
<tr>
<td>Norway</td>
<td>2 (6)</td>
</tr>
<tr>
<td>Other (Netherlands, Belgium, and Switzerland)</td>
<td>3 (10)</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>31 (100)</td>
</tr>
<tr>
<td><strong>Study design</strong></td>
<td></td>
</tr>
<tr>
<td>Randomized controlled trials</td>
<td>29 (94)</td>
</tr>
<tr>
<td>Nonrandomized controlled trials</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Before-and-after studies</td>
<td>1 (3)</td>
</tr>
<tr>
<td><strong>Outcome measure assessors</strong></td>
<td></td>
</tr>
<tr>
<td>PROMa</td>
<td>15 (48)</td>
</tr>
<tr>
<td>OBOMb</td>
<td>12 (39)</td>
</tr>
<tr>
<td>PROM and OBOM</td>
<td>3 (10)</td>
</tr>
<tr>
<td>PROM and HCPROMc</td>
<td>1 (3)</td>
</tr>
</tbody>
</table>

aPROM: patient-reported outcome measurement.
bOBOM: observer-reported outcome measurement.
cHCPROM: health care professional–reported outcome measurement.

Main Characteristics of the Training Programs

Of the 31 training programs, 11 (35%) were developed in the United States, and 8 (26%) in Germany (Table 2). A total of 22 (71%) programs were in English, followed by 4 (13%) in German and 2 (6%) in Dutch. Of the remaining programs, 1 (3%) was in French, 1 (3%) was in both English and Spanish, and language was not reported for one of the programs. Of the 31 training programs, 20 (65%) were developed in a primary health care context, and 11 (35%) were developed in specialized care. Regarding the type of health care professionals trained, 18 (58%) targeted physicians, and 26 (84%) were developed for fully trained health care professionals (Table 2). Only 5 (16%) training programs out of the 31 were developed with an interprofessional orientation, that is, with the promotion of interprofessionality as one of its training objectives.

A total of 24 (77%) of the 31 programs had an interactive format [3,32-39,41-45,51-60,64], and the remaining 7 (23%) were unidirectional training programs [40,46-50,61]. Among the 24 interactive training programs, 7 (29%) were developed using reflexivity strategies [32-34,41,42,55,56,67,68], among which...
5 (71%) were classified as peer-to-peer group learning [32,33,41,42,56,64,69] and 2 (29%) as self-appraisal individual learning [34,55,62]. Details on how the training programs were classified according to the reflexivity strategy used are reported in Table 3.

Table 2. Training program characteristics (N=31).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Country</strong></td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>11 (35)</td>
</tr>
<tr>
<td>Germany</td>
<td>8 (26)</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>4 (13)</td>
</tr>
<tr>
<td>Canada</td>
<td>3 (10)</td>
</tr>
<tr>
<td>Norway</td>
<td>2 (7)</td>
</tr>
<tr>
<td>Australia, Netherlands, Switzerland, and Germany (collaboration)</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Others (Netherlands and Belgium)</td>
<td>2 (7)</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>22 (71)</td>
</tr>
<tr>
<td>German</td>
<td>4 (13)</td>
</tr>
<tr>
<td>Dutch</td>
<td>2 (7)</td>
</tr>
<tr>
<td>French</td>
<td>1 (3)</td>
</tr>
<tr>
<td>English and Spanish</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Not reported</td>
<td>1 (3)</td>
</tr>
<tr>
<td><strong>Context of care</strong></td>
<td></td>
</tr>
<tr>
<td>Primary care</td>
<td>20 (65)</td>
</tr>
<tr>
<td>Specialized care</td>
<td>11 (36)</td>
</tr>
<tr>
<td><strong>Types of professional trained</strong></td>
<td></td>
</tr>
<tr>
<td>Physiciansa</td>
<td>18 (58)</td>
</tr>
<tr>
<td>Nurses and geneticists</td>
<td>8 (26)</td>
</tr>
<tr>
<td>Nurses</td>
<td>2 (7)</td>
</tr>
<tr>
<td>Physicians and nurses</td>
<td>2 (7)</td>
</tr>
<tr>
<td>Physicians and midwives</td>
<td>1 (3)</td>
</tr>
<tr>
<td>Fully trained</td>
<td>26 (84)</td>
</tr>
<tr>
<td>Not fully trained</td>
<td>5 (16)</td>
</tr>
<tr>
<td>Interprofessional orientation</td>
<td>5 (16)</td>
</tr>
</tbody>
</table>

aOne training program was designed for oncologists and gynecologists, and the other included medical residents.
Table 3. Quotations illustrating how reflexivity strategies are reported.

<table>
<thead>
<tr>
<th>Articles</th>
<th>Main quotations</th>
<th>Types of reflexivity strategies</th>
</tr>
</thead>
</table>
| Krones et al [32], 2008 | • “Practical communication strategies”  
• “After role-play feedback was given by their peers”\textsuperscript{a}  
• “Educational outreach...members were invited to moderate the sessions”  
• “Using the script-like decision aid was practiced through role playing”  
• “Participants received feedback from peers in their groups” | Peer-to-peer reflective group learning |
| Murray et al [33], 2010 | • “Self and peer appraisal during role play”  
• “Participants will evaluate decision-support skills to self-appraise their own and workshop peers’ quality of decision support during the case studies and role-play activities”  
• “To train nurses and medical residents in self-appraisal” | Peer-to-peer reflective group learning |
| Sanders et al [56], 2017 | • “The training was based on the learning principles described by Kolb. In the training sessions, group discussion, theory, role-playing, and reflections on personal behaviour were alternated. This tool was generated in the first session when the GPs reflected on their training experiences” | Peer-to-peer reflective group learning |
| Fossli et al [41], 2011 | • “The course consisted of a 50/50 mix of theory and 45 min group sessions (3-7 participants and two teachers per group) including role-plays, with plenary debriefs after each group”  
• “Our course was based on the same content as the 5-day course Communication Skills Intensive offered by Kaiser Permanente”  
• “At the conclusion of the course, all participants received a one-sheet overview of the Four Habits to carry in their pockets as reminder in everyday work” | Peer-to-peer reflective group learning |
| Kennedy et al [42], 2013 | • “Skills to encourage a structured approach to self-care support in consultations. Interactive role play (small groups) techniques to help deal with difficult issues during consultations. Interactive role play (small groups). Brief presentation with discussion. DVD exemplar of use plus manual involving (whole group). Explanatory models to encourage discussion about the causes and consequences of long-term conditions... Presentation with discussion. DVD exemplar of use plus manual (involving whole group)... As a practice–develop skills to solve problems that come up in the work of the practice. Problem-solving techniques involving whole practice systems within practice to improve self-care support for patients.”  
• “Problem-solving techniques involving whole practice ways to engage patients with self-care support.” | Peer-to-peer reflective group learning |
| Elwyn et al [34], 2004 | • “Practitioners attended two workshops. During the first workshop, the background literature on SDM was outlined and participants were asked to debate its relevance to clinical practice. The skills of SDM were described and demonstrated using simulated consultations. This provided opportunities for all the participants to comment on the method, using an observational competence checklist. Simulated patients were also encouraged to comment. Participants were asked to consult with the simulated patients using preprepared scenarios involving the study conditions. At the second workshop, participants were asked to consider the competences in more depth. By the end of the workshop, all participants had conducted and received feedback from at least one consultation with a simulated patient” | Self-appraisal individual learning |
| Epstein et al [55], 2017 | • “A 2-session in-office physician training (1.75 hours) using a brief video, feedback from standardized patients portraying roles of patients with advanced cancer; audio recorded study patient visits, and (2) (...), plus up to 3 follow-up phone calls (Table 1)” | Self-appraisal individual learning |

\textsuperscript{a}Text in italics illustrates possible reflexivity approaches and concepts (related to questions A and B of our criteria grid).

Effectiveness of the Training Programs in Connection With Different Variables

Training Formats and Effectiveness

Based on the Cochrane review classification of the effectiveness of the included interventions, 10 (42%) of the 24 interactive training programs were deemed effective, as opposed to 1 (14%) of the 7 unidirectional programs (Table 4).
<table>
<thead>
<tr>
<th>Training format and articles</th>
<th>$\text{SMD}^d$, $\text{EMD}^b$, or $\text{RD}^c$ (95% CI)</th>
<th>Narrative results</th>
<th>Effective$^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interactive format (including reflexivity strategies)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Krones et al [32], 2008; Hirsch et al [64], 2010</td>
<td>0.40 (0.28 to 0.52)</td>
<td>N/A$^e$</td>
<td>Yes</td>
</tr>
<tr>
<td>Murray et al [33], 2010</td>
<td>3.75 (2.46 to 5.03)</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td>Sanders et al [56], 2017</td>
<td>0.85 (0.54 to 1.16); 0.93 (0.62 to 1.25)</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td>Elwyn et al [34], 2004; Edwards et al [67], 2004; Longo et al [68], 2006</td>
<td>$-0.30 \pm -1.19$ to 0.59; 0.05 $\pm -0.17$ to 0.27</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Fossli et al [41], 2011</td>
<td>0.38 $\pm -0.17$ to 0.94</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Kennedy et al [42], 2013; Kennedy et al [69], 2010</td>
<td>$-0.05 \pm -0.12$ to 0.01</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Epstein et al [55], 2017; Butow et al [62], 2015</td>
<td>0.00 $\pm -0.24$ to 0.24</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td><strong>Interactive format (not including reflexivity strategies)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bieber et al [35], 2006; Bieber et al [65], 2007</td>
<td>N/A</td>
<td>“An ANOVA for repeated measurements comparing the SDM group with the information group revealed that patients’ appraisal of the interaction quality was higher in the SDM group”</td>
<td>Yes</td>
</tr>
<tr>
<td>Stacey et al [36], 2008</td>
<td>2.07 (1.26 to 2.87)</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td>Loh et al [37], 2007</td>
<td>N/A</td>
<td>“In the intervention group, significantly higher patient participation from pre- to postintervention was found.”</td>
<td>Yes</td>
</tr>
<tr>
<td>Haskard et al [38], 2008</td>
<td>N/A</td>
<td>“Training significantly improved physicians’ health behaviour counseling of their patients.”</td>
<td>Yes</td>
</tr>
<tr>
<td>Deinzer et al [39], 2009</td>
<td>N/A</td>
<td>“The degree of SDM was significantly higher in the SDM group at baseline and after one-year visits. The results of the SDM sum score on actually practiced SDM exhibited in both groups significantly increased, but the control group did not reach the score of the study group after one year.”</td>
<td>Yes</td>
</tr>
<tr>
<td>Feng et al [43], 2013</td>
<td>N/A</td>
<td>“Significant difference in favour of the intervention group, high risk of bias.”</td>
<td>Yes</td>
</tr>
<tr>
<td>Tinsel et al [44], 2013</td>
<td>0.32 (0.17 to 0.46)</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td>Hamann et al [3], 2007</td>
<td>0.16 $\pm -0.28$ to 0.61</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Bernhard et al [45], 2011; Butow et al [62], 2015</td>
<td>N/A</td>
<td>“There was no effect for this variable for SGA doctors (estimated population mean difference =0.52, SE =1.39, $ES^d=0.04; P=0.71$)” “After the training workshop, doctors in the experimental group within the ANZ cohort displayed more behaviours designed to establish the SDM framework than doctors in the control group (estimated population mean difference=3.42, SE =1.50, $ES=0.30, P=0.03$). However, the ES was small” “There was considerable variation in patient outcomes between the SGA and ANZ cohorts and no substantial training effect”</td>
<td>No</td>
</tr>
</tbody>
</table>

$^d$ Significant difference in favour of the intervention group, high risk of bias.

$^e$ N/A indicates no data available.

$^f$ SGA = Doctors, ANZ = Australian and New Zealand.

$^g$ Estimated population mean difference.

$^h$ Estimated standard error.

$^i$ Estimated effect size.

$^j$ Estimated standard error of the effect size.
**Reflexivity Strategies and Effectiveness**

Regarding the effectiveness of the programs, 3 (43%) of the 7 programs including reflexivity strategies were deemed effective [32,33,56]. The number of effective programs among programs with reflexivity strategies was similar to that among interactive programs without reflexivity strategies. Concerning the latter, 7 (41%) out of 17 were deemed effective (Table 4) [35-39,43,44].

Table 4 shows that 3 (60%) of the 5 training programs using a peer-to-peer group learning strategy were effective [32,33,56], whereas none of those using a self-appraisal individual learning strategy were effective (Table 4).

**Interprofessional Approach and Effectiveness**

Of the 5 training programs developed with an interprofessional orientation, 3 (60%) were classified by the Cochrane review as effective [32,33,46], and 2 (67%) of these included reflexivity strategies [32,33] (Table 5). In other words, the 2 programs with an interprofessional orientation that included reflexivity strategies were both deemed effective.

<table>
<thead>
<tr>
<th>Training format and articles</th>
<th>SMD&lt;sup&gt;a&lt;/sup&gt;, EMD&lt;sup&gt;b&lt;/sup&gt;, or RD&lt;sup&gt;c&lt;/sup&gt; (95% CI)</th>
<th>Narrative results</th>
<th>Effective&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tai-Seale et al [58], 2016; Dillon et al [63], 2017</td>
<td>0.35 (−0.53 to 1.24); 0.51 (0.19 to 0.84); −0.29 (−1.17 to 0.60); 0.00 (−0.32 to 0.32)</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Ampe et al [59], 2017</td>
<td>−0.10 (−0.96 to 0.76)</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Cox et al [60], 2017</td>
<td>0.11 (−0.21 to 0.42)</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td><strong>Unidirectional</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hess et al [46], 2012</td>
<td>2.82 (2.43 to 3.21)</td>
<td>N/A</td>
<td>Yes</td>
</tr>
<tr>
<td>O’Cathain et al [40], 2002</td>
<td>−0.02 (−0.05 to 0.01)</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Koerner et al [47], 2014</td>
<td>−0.08 (−0.26 to 0.11)</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Mathers et al [48], 2012</td>
<td>−0.09 (−0.23 to 0.05)</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Rise et al [49], 2012; Rise et al [72], 2016</td>
<td>0.13 (−0.32 to 0.58)</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Sheridan et al [50], 2014</td>
<td>−0.17 (−0.35 to 0.00)</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>Coylewright et al [61], 2016</td>
<td>0.51 (−0.05 to 1.07)</td>
<td>N/A</td>
<td>No</td>
</tr>
</tbody>
</table>

<sup>a</sup>SMD: standardized mean difference.
<sup>b</sup>EMD: effect size mean difference.
<sup>c</sup>RD: risk difference.
<sup>d</sup>Scored as “Yes” if the 95% CI reported in the Cochrane review did not include 0 for the SMD, RD, and MD values or when the 95% CI did not include 1 for the OR values. In some studies, with ≥2 scales, we referred to the conclusion of the authors.
<sup>e</sup>N/A: not applicable.
<sup>f</sup>SGA: Switzerland, Germany, and Australia.
<sup>g</sup>ES: effect size.
<sup>h</sup>ANZ: Australia and New Zealand.

<sup>i</sup>This study found no effect of shared decision-making training on the primary outcomes, which were similar between both the groups. However, training did contribute to improved observer-rated shared decision-making skills in physicians and reduced anxiety and depression in patients, particularly in women with breast cancer.

<sup>j</sup>The primary outcome measure was CollaboRATE, a patient-reported experience with care. While the odds ratios (ORs) from the ASK (Ask Share Know) clinic (OR 1.417) and the OpenComm plus ASK clinic (OR 1.134) were greater than 1, their 75% CIs included 1, which suggests no difference from the usual care clinic. Our findings suggest that something could be done to improve the patient experience. We view the results as promising evidence of the intervention’s efficacy and as meaningful signals of its likely effects on patient experience.
Table 5. Interprofessional orientation and type of measurement.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Effective</th>
<th>IP\textsuperscript{a} approach</th>
<th>Outcome assessors</th>
<th>Training format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Krones et al [32], 2008; Hirsch et al [64], 2010</td>
<td>Yes</td>
<td>Yes</td>
<td>• PROM\textsuperscript{b}</td>
<td>Interactive format (including reflexivity strategies)</td>
</tr>
<tr>
<td>Murray et al [33], 2010</td>
<td>Yes</td>
<td>Yes</td>
<td>• OBOM\textsuperscript{c}</td>
<td>Interactive format (including reflexivity strategies)</td>
</tr>
<tr>
<td>Sanders et al [56], 2017</td>
<td>Yes</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (including reflexivity strategies)</td>
</tr>
<tr>
<td>Elwyn et al [34], 2004; Edwards et al [67], 2004; Longo et al [68], 2006</td>
<td>No</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (including reflexivity strategies)</td>
</tr>
<tr>
<td>Fossli et al [41], 2011</td>
<td>No</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (including reflexivity strategies)</td>
</tr>
<tr>
<td>Kennedy et al [42], 2013; Kennedy et al [69], 2010</td>
<td>No</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (including reflexivity strategies)</td>
</tr>
<tr>
<td>Epstein et al [55], 2017; Butow et al [62], 2015</td>
<td>No</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (including reflexivity strategies)</td>
</tr>
<tr>
<td>Bieber et al [35], 2006; Bieber et al [65], 2007</td>
<td>Yes</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Stacey et al [36], 2008</td>
<td>Yes</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Loh et al [37], 2007</td>
<td>Yes</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Haskard et al [38], 2008</td>
<td>Yes</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Deinzer et al [39], 2009</td>
<td>Yes</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Feng et al [43], 2013</td>
<td>Yes</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Tinsel et al [44], 2013</td>
<td>Yes</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Hamann et al [3], 2007</td>
<td>No</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Bernhard et al [45], 2011; Butow et al [62], 2015</td>
<td>No</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Cooper et al [54], 2011</td>
<td>No</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Légaré et al [52], 2012; Allaire et al [70], 2012; Légaré et [71], 2013</td>
<td>No</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Cooper et al [53], 2013</td>
<td>No</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Wilkes et al [51], 2013</td>
<td>No</td>
<td>No</td>
<td>• PROM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Hârter et al [57], 2015; Bieber et al [66], 2018</td>
<td>No</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Tai-Seale et al [58], 2016; Dillon et al [63], 2017</td>
<td>No</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Ampe et al [59], 2017</td>
<td>No</td>
<td>No</td>
<td>• OBOM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Cox et al [60], 2017</td>
<td>No</td>
<td>Yes</td>
<td>• OBOM</td>
<td>Interactive format (not including reflexivity strategies)</td>
</tr>
<tr>
<td>Hess et al [46], 2012</td>
<td>Yes</td>
<td>Yes</td>
<td>• OBOM</td>
<td>Unidirectional</td>
</tr>
</tbody>
</table>
**Outcome Assessors and Effectiveness**

Among the selected articles, based on the Cochrane review, 8 (67%) out of the 12 programs using OBOMs were classified as effective [33,36,38,39,43,46,56], while 3 (21%) of the 14 programs using PROMs were classified as effective [32,44]. Meanwhile, 2 (67%) of the 3 training programs that included reflexivity strategies and were classified as effective were assessed using OBOMs. The only effective program in the unidirectional category was assessed using OBOMs (Table 5).

The main findings related to the different elements analyzed in the study are summarized in Table 6.

### Table 6. Summary of main findings (N=31).

<table>
<thead>
<tr>
<th>Level of analysis and training categories</th>
<th>Training programs</th>
<th>Effective program</th>
<th>IP(^a) approach</th>
<th>Effective program with IP(^a) approach</th>
<th>Type of assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Format</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactive</td>
<td>31</td>
<td>24 (7)</td>
<td>24</td>
<td>10 (42)</td>
<td>OBOM(^b): 10/24 (42); PROM(^c): 10/24 (42); PROM/OBOM: 3/24 (12); PROM/HCPROM: 1/24 (4)</td>
</tr>
<tr>
<td>Unidirectional</td>
<td>31</td>
<td>7 (23)</td>
<td>7</td>
<td>1 (14)</td>
<td>OBOM: 2/7 (29); PROM: 5/7 (71)</td>
</tr>
<tr>
<td><strong>Interactive</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflexivity strategies—yes</td>
<td>24</td>
<td>7 (30)</td>
<td>7</td>
<td>3 (43)</td>
<td>OBOM: 3/7 (43); PROM: 3/7 (43); PROM/OBOM: 1/7 (14)</td>
</tr>
<tr>
<td>Reflexivity strategies—no</td>
<td>24</td>
<td>17 (70)</td>
<td>24</td>
<td>7 (30)</td>
<td>OBOM: 7/17 (41); PROM: 7/17 (41); PROM/OBOM: 2/17 (12); PROM/HCPROM: 1/17 (6)</td>
</tr>
<tr>
<td><strong>Peer-to-peer group learning</strong></td>
<td>7</td>
<td>5 (71)</td>
<td>5</td>
<td>3 (60)</td>
<td>OBOM: 3/5 (60); PROM: 2/5 (40)</td>
</tr>
<tr>
<td>Self-appraisal learning</td>
<td>7</td>
<td>2 (29)</td>
<td>5</td>
<td>0 (0)</td>
<td>N/A(^d)</td>
</tr>
</tbody>
</table>

\(^{a}\)IP: interprofessional.
\(^{b}\)OBOM: observer-reported outcome measure.
\(^{c}\)PROM: patient-reported outcome measure.
\(^{d}\)N/A: not applicable.
Risk of Bias Assessment

The risk of bias assessment was reported in the initial Cochrane review, which used criteria for EPOC reviews [73] and the Cochrane Handbook for Systematic Reviews of Interventions [74] for interrupted time series designs.

Certainty of Evidence

For our variable of interest (the primary outcome), the certainty of evidence was assessed in the Cochrane review according to the GRADE (Grading of Recommendations Assessment, Development and Evaluation) guidelines and methods described in Chapter 12 of the Cochrane Handbook for Systematic Reviews of Interventions [75]. The method proposed in EPOC worksheets was used to determine which secondary outcomes should be assessed [76].

Discussion

Principal Findings

This secondary analysis of a 2018 Cochrane review on interventions for increasing the use of SDM by health care professionals aimed to identify training programs that included reflexivity and to ascertain how effective they were. Our study is the first to evaluate the effectiveness of SDM training programs that include reflexivity strategies. Among the 31 SDM training programs for health professionals, 23% (n=7) included reflexivity and 77% (n=24) did not. More of those that included reflexivity were deemed effective as a percentage of the whole (3/7, 43%) than those that did not (8/24, 33%). Among the interactive training programs, there was little difference in effectiveness between those that used reflexivity strategies (3/7, 43%) and those that did not (7/17, 41%). However, when comparing interactive training programs with unidirectional ones, there were a great many more programs deemed effective among the former (10/24, 42%) than among the latter (1/7, 14%). Among the training programs that included reflexivity (n=7), most programs using a peer-to-peer group learning strategy were found to be effective (3/5, 60%), whereas those using a self-appraisal individual learning strategy were not (0/2, 0%). The training programs with an interprofessional orientation were more frequently classified as effective (3/5, 60%) than those without (2/5, 40%). Finally, the percentage of effective training programs in studies using OBOMs to assess training was higher than the percentage in studies using PROMs. These findings led to the observations that follow.

A Larger Percentage of Reflexivity-Based Training Programs Were Deemed Effective

Our results confirm the findings of Leyland et al [77] and Chaffey et al [12] that in general, programs that include reflexivity have more positive effects than those that do not [12,77]. Training using reflexivity strategies has been shown to increase medical students’ ability to integrate alternative sources of knowledge and critically reflect on their own practices [12]. According to Chaffey et al [12], it is difficult to assess reflexivity alone, but when applied in an intervention, it yields more positive results. Although only a few SDM training programs include reflexivity strategies to date, these results suggest that reflexivity could be a core component of effective training and, as such, may be understood as an effective implementation strategy for change. However, further studies with a larger sample are needed to confirm this hypothesis. For Kolb [78], learning is a complex process driven partly by individuals’ ability to find a sense for themselves in operationalizing change in their day-to-day practices. As our results suggest, training that uses reflexivity strategies could provide health care professionals with that personal sense of the purpose of SDM and the motivation needed to implement SDM in their practices. However, future research could more precisely evaluate the impact of reflexivity strategies on SDM uptake by health care professionals if more programs were explicit in their use or promotion of reflexivity. The training programs included in this review were not designed to use or promote reflexivity. An analysis that compared SDM training programs that were explicitly based on reflexivity with those that were not would be more useful and appropriate. Nevertheless, our results suggest that SDM training for health care professionals based on reflexivity strategies is effective in motivating trainees to adopt SDM in a manner that facilitates positive patient experiences in health care systems.

A Peer-to-Peer Group Learning Strategy Would Appear to Be More Effective Than a Self-appraisal Learning Strategy

Our findings also showed that more training programs using peer-to-peer group learning were classified as effective than those using self-appraisal individual learning. These findings suggest that a self-appraisal strategy can be more effective than training focused on an individual if it is part of a group learning process. Interaction among learners seems to be a powerful strategy for encouraging reflection, even self-reflection. Research could further compare the types of reflexivity learning strategies that are most effective in SDM training for health care professionals.

Training Using Reflexivity Strategies and an Interprofessional Orientation Can Lead to Better Results

Our findings indicated that training programs with an interprofessional orientation have proven to be more effective than those without this orientation. All programs that used interprofessional orientation and reflexivity strategies were also deemed effective. In addition, these programs used a peer-to-peer group learning strategy in which participants were encouraged to learn from each other’s reflections by sharing their points of view and experiences [28]. One of the main goals of interprofessional SDM is to encourage the recognition of other professionals’ values and competencies [21,79], in other words, to recognize other members of the care team as peers [22]. Leyland et al [77] also considered that reflexivity is needed for health care professionals to recognize each other’s professional skills. This suggests that the peer-to-peer group learning strategy is especially appropriate for implementing an interprofessional approach. In addition, Tremblay et al [27] defined the goal of “formative reflexivity” (ie, reflexive practices for learning) as developing new visions of professional experience and working in collaboration. Thus, blending these 2 goals (interprofessional orientation and reflexivity) is a
promising avenue for fostering the uptake of interprofessional SDM in health practice. Furthermore, encouraging reflection and questioning of one’s practice can help to correct the paternalistic way in which SDM still seems to be undertaken [19,77].

**Measuring SDM Training for Health Care Professionals Only With OBOM or PROM, a Limiting Approach**

The Cochrane review used only OBOMs and PROMs (or both) to evaluate intervention effectiveness, eschewing programs that used HCPROMs. In the review, most programs classified as effective were assessed using OBOMs (8/11, 73%) rather than PROMs (3/11, 27%). OBOMs are considered a more rigorous assessment option than self-reported measures because they are more independent [23]. Based on this justification, one can say that the apparent objectivity of OBOM assessment yields better results than the potential subjectivity arising when patients report an outcome. Furthermore, our findings showed that 2 (67%) of the 3 training programs using reflexivity and classified as effective were assessed with OBOMs, as opposed to only 1 (33%) with PROMs. However, neither PROMs nor OBOMs respect the first Kirkpatrick model criterion, which is the satisfaction of the participants with the training. According to Kirkpatrick, when participants like a training program, they may be willing to adopt what they have learned from it. Based on the importance of this first level of evaluation, we suggest that the perspective of the training beneficiaries (health care professionals) should be considered in evaluations, as is the case with HCPROMs. This does not exclude the use of PROM or OBOM measures to assess Kirkpatrick’s other 3 levels of improvement: knowledge improvement, practice improvement, and health-related outcome improvement (the levels considered in the Cochrane review). Furthermore, a reflexivity-based approach focuses not only on outcomes but also on the process or experience of health care, which may be as important to patients as outcomes [80]. If trainees reflect on their own practice and question their own values, they will better consider patients’ values and experiences in the health care-seeking process. Therefore, another relevant measure of effectiveness would be patient-reported experience measures (PREMs), which help improve patient care [81]. In a 2019 systematic review, Müller et al [9] analyzed the methods used by 41 studies that assessed SDM training programs for health care professionals and concluded that the diversity of assessment methods limits the ability to compare training program effectiveness and is a barrier to conclusive evidence. Therefore, it now seems important to develop a harmonized SDM training assessment measure that includes all the 4 perspectives to enable better comparison.

**Multiplicity of Elements Did Not Facilitate the Analysis**

In this review, some interventions involved multiple components, for example, they included a workshop, a web-based tutorial, and a decision aid tool or a workshop with audit and feedback. In these cases, it was difficult to evaluate the effects of the components separately. In addition, some of the training programs included patients. If the training includes patients, it can be difficult to know whether this inclusion has an additional impact on effectiveness. Another example is the diversity of the comparators. The articles included in the review used different types of comparators (usual care or another differing intervention). Finally, even if the primary outcome (SDM) was our focal interest, it may be defined differently from one study to another. For example, while some focus their analysis on the uptake of SDM as a whole, others may analyze only one component (eg, decision conflict or decision regret) or separate them. In some articles, analysis used ≥2 scales, which made our judgment difficult, especially if effectiveness-related results were different. To classify these types of articles in terms of effectiveness, we referred to the conclusions of the authors. Based on all these observations, we suggest using a core set of assessment methods and validated outcomes for all learning levels in SDM training programs to identify the most effective strategies and better compare them. Future research could explore methods to specifically assess reflexivity strategies included in SDM training programs targeting health care professionals for their inclusion in this core model. Our reflexivity grid, when validated and published, along with the results from this analysis, can be considered as the first step in assessing SDM training that uses reflexivity strategies, but further work is needed to guide training approaches in this field.

**Limitations**

This study has a few limitations. First, the systematic review included evidence only up to June 2017, and another update has not been performed since then. The Cochrane review has been updated twice, with each update including all the studies in the earlier versions; yet, its conclusions regarding effectiveness have varied little since the first review in 2010. At its first publication, it included only 5 RCTs, but in 2018, with 84 RCTs, it still concluded that “a great variety of activities exist to increase shared decision-making by health care professionals, but we cannot be confident about which of these activities work best because the certainty (or the confidence) of the evidence has been assessed as very low.”

Second, seeing that the training programs included in the systematic review were not explicit about promoting reflexivity, the selection of those that included what could be accurately described as reflexivity strategies was not an easy task. Nevertheless, using our grid based on a preliminary rapid review, we identified 7 programs. This was a small number for further analysis, although not surprising, as reflexivity is a new concept in the context of health professional education about SDM. It is possible that we missed a few programs that included reflexivity, as we relied solely upon published data and did not contact the authors for additional information. To minimize this risk, we analyzed all additional materials cited in the articles that were linked with the training programs and considered the most recent publications to discuss our results. Finally, the grid used in the analysis was developed for the purpose of this study, and although it is an important contribution to the field of implementation, its validation will require further use and assessment.

**Conclusions**

Our results suggest that SDM training programs for health care professionals using reflexivity strategies could increase SDM
implementation. Our study is the first to evaluate the effectiveness of SDM training that includes reflexivity and raises several questions: (1) Are peer-to-peer learning strategies more effective than self-appraisal strategies? (2) How can reflexivity and interprofessional orientation strategies best complement each other? (3) Are OBOMs and PROMs the only appropriate means of evaluating SDM training programs? The grid developed for identifying reflexivity strategies in training programs, including reflexivity-related approaches and concepts, will be a useful guide for developing reflexivity training and is to be validated in future studies.

Acknowledgments
The authors wish to thank Louisa Blair, scientific editor, for the language assessment and scientific editing, along with Anne McBryde of Abaca Traduction for revision services. They also thank Lionel Adisso, PhD candidate in epidemiology, for orienting the data analysis.

Authors’ Contributions
NTD and FL were involved in the conceptualization. NTD and JL were involved in article selection. NTD, VB, and AM were involved in the classification of training programs and data extraction. NTD and JL were involved in data analysis. NTD and FL were involved in the investigation. NTD, M-CT, FL, MJD, and SG-B were involved in the development of the collection grid. FL was involved in resource collection. FL, MJD, and M-CT were involved in supervision. NTD wrote the first draft, and JL edited it. NTD, JL, FL, MJD, and M-CT were involved in writing, reviewing, and editing.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Extraction grid for training program based on reflexivity strategies.
[XLSX File (Microsoft Excel File), 32 KB - mededu_v8i4e42033_app1.xlsx ]

References


Abbreviations

EPOC: Effective Practice and Organization of Care
HCPROM: health care professional–reported outcome measurement
OBOM: observer-reported outcome measurement
PREM: patient-reported experience measure
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROM: patient-reported outcome measurement
RCT: randomized controlled trial
SDM: shared decision-making

©Ndeye Thiab Diouf, Angèle Musabyimana, Virginie Blanchette, Johanie Lépine, Sabrina Guay-Bélanger, Marie-Claude Tremblay, Maman Joyce Dogba, France Légaré. Originally published in JMIR Medical Education (https://mededu.jmir.org), 07.12.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.

https://mededu.jmir.org/2022/4/e42033
Viewpoint

The Role of Academic Health Systems in Leading the “Third Wave” of Digital Health Innovation

Adeel A Faruki1, MD, MBA; Richard D Zane2, MD; Jennifer L Wiler2, MD, MBA

1Department of Anesthesiology, University of Colorado Hospital School of Medicine, Aurora, CO, United States
2Department of Emergency Medicine, University of Colorado Hospital School of Medicine, Aurora, CO, United States

Corresponding Author:
Adeel A Faruki, MD, MBA
Department of Anesthesiology
University of Colorado Hospital School of Medicine
13001 East 17th Place
7th Floor
Aurora, CO, 80045
United States
Phone: 1 7208486723
Email: adeel.a.faruki@gmail.com

Abstract

Investors, entrepreneurs, health care pundits, and venture capital firms all agree that the health care sector is awaiting a digital revolution. Steven Case, in 2016, predicted a “third wave” of innovation that would leverage big data, artificial intelligence, and machine learning to transform medicine and finally achieve reduced costs, improved efficiency, and better patient outcomes. Academic medical centers (AMCs) have the infrastructure and resources needed by digital health intrapreneurs and entrepreneurs to innovate, iterate, and optimize technology solutions for the major pain points of modern medicine. With large unique patient data sets, strong research programs, and subject matter experts, AMCs have the ability to assess, optimize, and integrate new digital health tools with feedback at the point of care and research-based clinical validation. As AMCs begin to explore digital health solutions, they must decide between forming internal teams to develop these innovations or collaborating with external companies. Although each has its drawbacks and benefits, AMCs can both benefit from and drive forward the digital health innovations that will result from this journey. This viewpoint will provide an explanation as to why AMCs are ideal incubators for digital health solutions and describe what these organizations will need to be successful in leading this “third wave” of innovation.

(JMIR Med Educ 2022;8(4):e32679) doi:10.2196/32679

KEYWORDS
innovation; academic hospitals; academic health systems; health technology; entrepreneur; disruption; digital health; research programs; cost; investment; intrapreneur

Introduction

In 2016, Steve Case, health care futurist, entrepreneur, investor, and former CEO of America Online, declared the health care industry ripe for disruption and predicted a “third wave” of innovation that would alter how we practice medicine [1]. This innovation would leverage big data, artificial intelligence, and machine learning to transform health care and finally achieve the reduced costs and improved outcomes demanded by the public. Case’s prediction appears imminent, but even extremely successful companies have already tried and failed to disrupt our complex health care system as seen with the joint venture Haven [2]. Ambitious ideas combined with a revenue of US $534 billion dollars and a high-profile leadership team could not tackle the perverse incentives and market dynamics of the current health care system [3]. Therefore, it is imperative that this change must develop within health care organizations.

Medical research insights and knowledge in health care overall have been rapidly expanding in the past 70 years due to advances in technology, and this momentum will only accelerate as early adoption of health care innovation continues [4]. As Balas and Chapman [5] stated, health care innovation is based on the criteria of “being novel, nonobvious and useful.” With this definition, innovation can be seen as a gradual process improvement by updating outdated processes and tools, or a complete disruption of current systems. Critically, innovation...
in health care will never achieve significant positive impact on patient outcomes without accepting the risks of early adoption and without the integration of digital health solutions into common practice. This viewpoint will further provide an explanation as to why academic medical centers (AMCs) are ideal incubators for digital health solutions; it will also describe what these organizations will need to be successful in leading this “third wave” of innovation.

Why AMCs Should Lead the “Third Wave”?

AMCs across the United States serve as community hospitals, safety net institutions, and as state-of-the-art research-intensive quaternary referral centers. Even though AMCs only account for 6% of the US health system in 2012, they account for “47% of organ transplants, 60% of level one trauma centers, and 66% of burn units” [6,7]. These referral centers are known to provide care for the most complex patients with rare or difficult-to-manage disease processes and multiple comorbidities [8]. The combination of patients with disease processes at various stages and large catchment areas results in enormous amounts of unique data that are constantly growing and stored in relatively easily accessible electronic medical records. These large stores of unique patient data can be leveraged to ensure any digital health solutions developed are applicable to diverse patient populations. Historically, these data had primarily been abstracted by billing coders, nurses, cancer registrars, quality improvement teams, and health information management professionals at an institutional level [9]. The purposes of the abstraction have included billing, patient registry functions, quality improvement initiatives, and clinical research. We note that most health care institutions have the infrastructure to abstract data successfully from the electronic medical record (EMR) for digital health innovation; however, at AMCs, this can be performed with increased safeguards to safeguard patient’s protected health information through institutional review board (IRB)–approved research projects.

Research has always been one of the integral components of the mission of any academic institution and AMCs; medical schools and universities continue to be extensively funded to perform this mission with US $29.5 billion dollars of funding from the National Institutes of Health extramural awards in fiscal year 2019 alone [10-12]. Landmark medical developments from AMCs range from the HER2-Herceptin breast cancer treatment clinical trial at University of California Los Angeles in 2001 to the discovery of altered mRNA initiating a protective immune response at the University of Pennsylvania in 2005 [13-15]. These studies resulted in significant innovations in disease management or prevention and were the result of structured research projects at AMCs. The health care system, as a whole, has relied on AMCs to use basic science and clinical research, both funded and unfunded, to innovate in diagnostic modalities, therapeutics, care delivery, patient safety, and now in technology integration. AMCs have structured processes with technology transfer offices and IRBs to perform high-quality translational research. One significant benefit to any technology company who collaborates with an AMC will be the ability to clinically validate their digital health innovations through research studies that can be published as scientific manuscripts. This research supports the academic mission of the organization and acts as a differentiator for a health technology start-up company. Finally, AMCs have access to clinical faculty who deliver patient care on the frontline and can advise on those digital health implementations that will add the greatest value to clinical quality and patient safety. The collaboration between AMCs and technology companies has already started, but it has been a gradual process. Institutions such as the University of Pittsburgh Medical Center and University of California, San Francisco are already integrating artificial intelligence and machine learning into their EMR to help clinicians identify chronic diseases and improve imaging interpretation accuracy through co-development with external technology companies [16,17]. Moreover, UCHealth University of Colorado Hospital is performing clinical validation for a wireless, wearable patient motion sensor in the intensive care unit that communicates with the EMR to inform care teams if patients are at risk of pressure ulcer injuries [18]. Although, other types of health care organizations may have access to large stores of patient data and have legal infrastructure with contracting capabilities to collaborate with technology companies, AMCs have more extensive infrastructure to perform IRB-approved research studies. AMCs have experts in the field who are incentivized by their institutions to continually develop new knowledge, intellectual property, and novel technologies, which is why they are uniquely poised to lead the next wave of health care advancement through technology integration and validation.

What Will AMCs Need to Lead the “Third Wave” of Digital Health Innovation?

To successfully lead in digital health innovation, AMCs will need to dedicate resources, both financial and human capital, to support these endeavors. The first determination is whether the AMC wants to develop digital health tools internally with their own employees acting as entrepreneurs within organizations, otherwise known as intrapreneurship, or through collaboration with external companies. Intrapreneurship, as explained by Bill Aulet, the Managing Director of the Marin Trust Center for the Massachusetts Institute of Technology, is as follows:

[C]reating value with new products, new ways of running businesses, and with a number of assets that [the company or organization] control. [19]

The concept of intrapreneurship has always existed in AMCs as quality improvement and process improvement projects. The specialists who have led these projects in the past would be an ideal foundational group for a digital health development team in an AMC. However, as health care technology often integrates with existing EMR capabilities and may require bidirectional data flow from the record, it is critical to include a dedicated group of information technology (IT) specialists as collaborators. Additionally, this team will need training in implementation science, which focuses on converting research findings and evidence-based practices and implementing them into routine practice to impact patient care [20]. This training will enable
the team to understand how to create digital health solutions that solve the problems faced by frontline health care providers or patients. Finally, these intrapreneurship teams will require members with strong clinical experience or ad hoc subject matter experts to ensure the innovations being developed will either improve clinical care quality, efficiency in processes, or patient safety. Often the best innovators can create tools that fail to help the target audience after they implement phases; therefore, repeated experimentation in the form of pilot studies is critical to success [21]. Many AMCs have developed innovation centers to facilitate this work, but these centers will only be successful with adequate funding to enable development of the digital health solutions. If successful, these solutions can also outgrow into separate companies, be sold to other health care organizations, and serve as a form of revenue for the founding AMC.

If AMCs collaborate with entrepreneur outside of the organization, such as health technology start-up companies, they can consider either codeveloping a digital health solution or adopting an existing solution [22]. As mentioned by Marwaha et al [23], some of the pitfalls when adopting an existing solution is the lack of understanding of the AMC’s problems and of the existing IT infrastructure in which the digital health tool must integrate. If codeveloping with a start-up company, the AMC may consider investing in the start-up, which will be high financial risk but can also result in a future source of revenue [23]. When codeveloping digital health solutions with a start-up company, AMCs have the added ability to provide clinical expertise as well as gain access to data and resources to help the start-up succeed while simultaneously customizing the solution to their own organization. The potential risk of adopting or partnering with digital health technology start-up companies is the Silicon Valley fail fast mentality [24]. Failing fast is ideal for early-stage start-up companies and venture capital firms; however, this often leaves early adopters without functioning technology and wasting their own resources on an innovation that either no longer exists or is no longer supported. If switching to an early-stage digital health solution, AMCs may disrupt existing workflows or technologies they use to provide patient care with the hope of improved efficiency or outcomes. Furthermore, there is significant up-front cost in the form of financial resources and time to integrate a new digital health tool into the current IT systems. Any collaboration, co-development, or investment by an AMC in an external company for digital health innovation must be vetted carefully. This ensures the digital health solution will resolve issues faced by the AMC and the external company will be dedicated to continual support and ongoing development of the innovation. With health care organizations in the United States already spending approximately 2.5%-2.8% of their annual revenue on IT costs, it is critical to ensure any added cost will be worthwhile for the long term [25].

Whether AMCs choose to develop in-house digital health innovations or integrate technology from external companies, they will need the full support of the organization’s senior leadership team. Many AMCs have incorporated innovation into their values and strategic foci, such as Cedars-Sinai and Brigham and Women’s Hospital, which enables conversations with board members and financial officers to provide funding to support these endeavors [26,27]. Also, it is critical to vet any new partnerships, ideas, and spun-off companies through the AMC’s legal department and ensure any research studies are approved through the organization’s IRB for clinical studies.

Despite AMCs comprising only 6% of the United States’ health care system, they have large unique patient data sets, advanced health information management systems, data abstraction infrastructure, a constant desire to improve patient care quality, research missions, and the scientific method mindset [6]. These characteristics make AMCs a unique and ideal environment for intrapreneurship or collaboration with external companies to develop digital health solutions that can be validated to ensure they improve efficiency, patient safety, and clinical care quality for patients. As the “third wave” of digital health innovation begins to swell, AMCs should lead this transformation for all health care.

Conflicts of Interest
None declared.

References


Abbreviations

AMC: academic medical center
EMR: electronic medical record
IRB: institutional review board
IT: information technology
©Adeel A Faruki, Richard D Zane, Jennifer L Wiler. Originally published in JMIR Medical Education (https://mededu.jmir.org), 09.11.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
The Educational Impact of Web-Based, Faculty-Led Continuing Medical Education Programs in Type 2 Diabetes: A Survey Study to Analyze Changes in Knowledge, Competence, and Performance of Health Care Professionals

Stewart B Harris1, MPH, MD, CM; Shannon Idzik2, BS, MS, DNP; Adriano Boasso3, BSc, MSc, PhD; Sola Quasheba Neunie3, BSc, MSc; Alexander Daniel Noble3, BSc, MSc; Helen Elaine Such3, BA, PgCE; Joanna Van4, BS, DO

1 Schulich School of Medicine and Dentistry, Western University, London, ON, Canada
2 University of Maryland School of Nursing, Baltimore, MD, United States
3 touchIME, Stockport, United Kingdom
4 University Clinical Investigators (d.b.a. Diabetes Research Center), Tustin, CA, United States

* all authors contributed equally

Corresponding Author:
Joanna Van, BS, DO
University Clinical Investigators (d.b.a. Diabetes Research Center)
2492 Walnut Avenue
Suite 130
Tustin, CA, 92780
United States
Phone: 1 410 706 8129
Email: jvan@uciinc.net

Abstract

Background: The treatment landscape for type 2 diabetes (T2D) is continually evolving; therefore, ongoing education of health care professionals (HCPs) is essential. There is growing interest in measuring the impact of educational activities, such as through use of the Moore framework; however, data on the benefits of continuing medical education (CME) in the management of T2D remain limited.

Objective: This study aimed to evaluate HCP satisfaction; measure improvements in knowledge, competence, and performance following short, case-based, multidisciplinary web-based CME activities; and identify the remaining educational gaps.

Methods: Two faculty-led, CME-accredited, web-based educational activities on T2D and obesity, touchIN CONVERSATION and touch MultiDisciplinary Team, were developed and made available on a free-to-access medical education website. Each activity comprised 3 videos lasting 10 to 15 minutes, which addressed learning objectives developed based on a review of published literature and faculty feedback. Participant satisfaction (Moore level 2) was evaluated using a postactivity questionnaire. For both activities, changes in knowledge and competence (Moore levels 3 and 4) were assessed using questionnaires completed by representative HCPs before or after participation in the activities. A second set of HCPs completed a questionnaire before and after engaging in activities that assessed changes in self-reported performance (Moore level 5).

Results: Each activity was viewed by approximately 6000 participants within 6 months. The participants expressed high levels of satisfaction (>80%) with both activities. Statistically significant improvements from baseline in knowledge and competence were reported following participation in touchIN CONVERSATION (mean score, SD before vs after activity: 4.36, 1.40 vs 5.42, 1.37; \( P < .001 \)), with the proportion of learners answering at least six of 7 questions correctly, increasing from 22% (11/50) to 60% (30/50). A nonsignificant improvement in knowledge and competence was observed following participation in touch MultiDisciplinary Team (mean score, SD 4.36, 1.24 vs 4.58, 1.07; \( P = .35 \)); however, baseline knowledge and competence were relatively high, where 80% of the respondents (40/50) answered at least four of 6 questions correctly. A significant improvement in HCP self-reported performance was observed in a combined analysis of both activities (mean score, SD 2.65, 1.32 vs 3.15, 1.26; \( P = .03 \)), with the proportion of learners selecting the answer representing the best clinical option for all 4 questions increasing from 32% (11/34) to 59% (20/34) after the activity. Several unmet educational needs were self-reported or identified from the
primary care providers play a key role in ensuring that patients in the prevalence of T2D in countries such as the United States, practitioners, and physician assistants. Thus, given the increase in primary care providers, such as family physicians, internists, nurse practitioners, and physician assistants. Thus, given the increase in the prevalence of T2D in countries such as the United States, primary care providers play a key role in ensuring that patients who do not require specialist care remain at low risk of complications and comorbidities and can be effectively managed in a primary care setting [11]. The focus of primary care providers on the prevention of T2D progression or worsening also makes them well placed to lead a person-centered approach to diabetes care with the aim of achieving both good glycemic control and reducing the risk of complications [11].

**Introduction**

**Epidemiology and Burden of Type 2 Diabetes**

Diabetes is a major public health concern worldwide. In 2021, it was estimated to affect 537 million adults (9.8% of the world’s population) and was responsible for 6.7 million deaths [1]. Type 2 diabetes (T2D) is the most common type of diabetes, accounting for more than 90% of all cases worldwide, and is often associated with lifestyle factors, such as an unhealthy diet and obesity [1]. It is well established that reducing levels of glycated hemoglobin (HbA1c) in patients with diabetes can delay the onset and progression of microvascular and macrovascular complications [2,3]. An HbA1c <7% is recognized by both the American Diabetes Association and the European Association for the Study of Diabetes as an appropriate glycemic target [4,5], although the proportion of patients with T2D who achieve HbA1c <7% varies from approximately 20% to 50% in different regions of the world [6,7]. Thus, there remains a need to achieve optimal glycemic control in patients with T2D [6]. However, diabetes management has become increasingly complex for health care professionals (HCPs) owing to multiple medication classes and treatment combinations, the need to avoid hypoglycemia or hyperglycemia, multiple medical device options, and the need to facilitate patients’ lifestyles [8]. This multiplicity of treatment options, combined with the need to manage the risk of complications in patients with T2D, underscores the need for HCP education to ensure optimal patient management according to the most recent guidelines and evidence-based practice [9,10]. T2D management has also evolved from a glucocentric approach aimed at achieving glycemic control to a holistic approach aimed at preventing complications and improving quality of life, with a 2018 consensus report from the American Diabetes Association and European Association for the Study of Diabetes, highlighting the importance of person-centered care [5,11]. Specifically, it is now recommended that the selection of add-on therapy after metformin should be based on factors other than just HbA1c, and decision-making should also take into account the presence of comorbidities such as atherosclerotic cardiovascular disease, heart failure, and chronic kidney disease, as well as the patient’s clinical characteristics, risks for side effects, and socioeconomic factors [11]. This person-centered approach to T2D management is particularly relevant to primary care providers, such as family physicians, internists, nurse practitioners, and physician assistants. Thus, given the increase in the prevalence of T2D in countries such as the United States, primary care providers play a key role in ensuring that patients who do not require specialist care remain at low risk of complications and comorbidities and can be effectively managed in a primary care setting [11]. The focus of primary care providers on the prevention of T2D progression or worsening also makes them well placed to lead a person-centered approach to diabetes care with the aim of achieving both good glycemic control and reducing the risk of complications [11].

**Education in T2D Management**

As the treatment landscape and guidelines for T2D are continually evolving, innovative educational activities are required to ensure that HCPs remain up to date with clinical developments in the management of the disease. In addition, a multidisciplinary and person-centered approach is recommended for the management of patients with diabetes [12]. In support of this, a position statement published by the Insights for Diabetes Excellence, Access, and Learning Group in 2020 emphasized the need for responsive and effective HCP education to meet the increasing needs for diversity, specialization, cultural competence, advancing practice, and person-centeredness in diabetes care delivery [13]. They also highlighted the importance of proactive rather than reactive diabetes care to avoid therapeutic inertia in timely treatment intensification [13]. Many activities focusing on HCP education in the management of T2D have been developed. Although some peer-reviewed publications have described outcomes from educational activities in diabetes, these are highly heterogeneous, may include patients with type 1 diabetes, and may include a range of activities and a variety of end points [9,10,14-20].

**Medical Education for HCPs**

Traditionally, ongoing medical education for HCPs worldwide involves live symposia, face-to-face workshops, and training events. However, for many HCPs, these can be cost and time prohibitive [21,22]. As an alternative, web-based distance learning offers many advantages, including ease of access, ability to take a course from any location, lower cost of delivery, and availability at any time [21,22]. With the onset of the COVID-19 pandemic, the need to digitize medical education programs became even more urgent to ensure that HCPs had continued access to education in the absence of face-to-face events [23]. Several studies and commentaries published during this time illustrated that web-based events can be effective and can reduce barriers to access [24-26]. Effective web-based educational activities depend on many factors, including well-designed course content and well-prepared and fully supported instructors [27]. Education should also be in the...
context of patient care, answer HCP questions, and be directly applicable to their work [28]. Precanvassing the target audience, for example, asking potential learners to provide specific questions or cases that they would like the activity to cover, can be valuable, as it theoretically allows learners to become more vested in the activity by contributing questions on their own key educational areas of interest. Involving learners in the identification of educational gaps is a well-known tool for designing educational activities that can effectively impact theoretical and practical knowledge [29]. Precanvassing the audience allows us to involve learners in activity development and encourages participation and engagement. The value of adopting surveys of the target audience to identify needs and shape educational programs has been previously demonstrated [30]. Delivering content using an engaging format may be particularly important in the digital age, given the competing demands for individuals’ attention [31].

The need for interdisciplinary medical education has arisen because medicine has become increasingly specialized in recent decades; this can be effectively met by various specialties presenting content to a multidisciplinary audience of HCPs [32]. With studies demonstrating the high use of freely available medical education by a multidisciplinary audience, it is important for medical education providers to address this need by providing interdisciplinary programs [32]. The World Health Organization has also highlighted that medical education providers and programs should deliver education that helps HCPs acquire wide-ranging competencies, including multidisciplinary patient care [33].

Assessing the Impact of Education

For many years, HCPs only had to provide documentation of attendance at educational activities to qualify for certification by their professional associations or the reregistration of their medical licenses [34]. However, there is growing recognition of the need to assess the impact of continuing medical education (CME) on HCP performance and health outcomes [34]. The Accreditation Council for Continuing Medical Education (ACCME) now requires CME providers to demonstrate changes in learner competence, performance, or patient outcomes because of the program [35]. Similarly, the American Nurses Credentialing Center requires that an accredited provider measures the impact of its educational activities in relation to improving the knowledge, skills, and practices of registered nurses [36].

In 2009, Moore et al [34] developed an expanded 7-level framework for planning and assessing the outcomes of a CME program. Level 1 measures the number of HCPs who participate in an activity, and level 2 measures the extent to which they are satisfied with it, using a questionnaire completed after the activity. Levels 3 and 4 measure knowledge and competence, respectively. Knowledge can be assessed either objectively through pre- and postactivity tests or subjectively through self-reports of knowledge gain. Competence can be assessed either objectively by observation in an educational setting or subjectively by self-reporting competence or intention to change. Level 5 measures performance and can be objectively assessed through performance in a patient care setting or subjectively assessed through self-reporting of performance. Levels 6 and 7 measure the degree to which education can improve the health status of patients or a community of patients through an analysis of health status measures in patient charts or databases or of epidemiological data [34]. Moore levels have been widely used to measure the outcomes of educational programs and have been included in many consensus documents and best practice recommendations [37].

In this study, we developed and implemented 2 faculty-led, CME-accredited, web-based educational activities on T2D and obesity and analyzed the educational outcomes up to Moore level 5. The objectives of this analysis were to (1) evaluate the learners’ satisfaction with the educational activities, as well as the changes in knowledge, competence, and performance that were achieved following their implementation and (2) identify the remaining educational gaps in the clinical management of T2D and obesity.

Methods

Educational Activities

Educational gaps were identified at the start of activity development, in March 2021, by touch Independent Medical Education (touchIME), a provider of independent medical education for the global HCP community. The identification process included a thorough review of the relevant published literature on T2D and feedback from expert faculty specializing in diabetes care and research.

The expert faculty and patient faculty member were identified and recruited by the medical directors at touchIME. The expert faculty was identified through searches of the literature, relevant congress websites, and web-based educational videos, for diabetes experts with an established background in diabetes research and clinical practice. The Patient faculty member was identified through searches for videos or blogs detailing the firsthand experience of a patient with T2D and obesity. Recruitment was conducted by email invitation, which included details of the proposed activity. Conflict of interest statements from all faculty participants were gathered during the recruitment stage. All the expert faculty members involved in the educational activities are authors of this manuscript or mentioned in the acknowledgments.

Learning objectives were designed based on the educational gaps, and 2 faculty-led, web-based, CME-accredited activities were developed, each comprising 3 recorded 10- to 15-minute videos that addressed the learning objectives (Multimedia Appendix 1). The identified educational gaps and corresponding learning objectives for touchIN CONVERSATION and touch MultiDisciplinary Team (touchMDT) are listed in Table S1 in Multimedia Appendix 2. Both activities were recorded remotely using a web-based video conferencing platform and were made available to the HCPs for a maximum of 24 months after launch.

The first activity, touchIN CONVERSATION, featured an endocrinologist and a diabetes specialist and focused on the management of specific patient cases in the clinic. The learning objectives were to (1) evaluate the unmet need for achieving glycemic control and the associated reasons, (2) decide how to
apply individualized glycemic targets according to patient characteristics, and (3) choose appropriate treatments with properties relevant to the individual patient to help achieve glycemic control. For the activity to be immediately relevant to participants’ daily practice, the target audience was precanvassed for questions related to specific patient cases. Precanvassing was carried out by touchIME starting 4 weeks before the videos were recorded, whereby HCPs were invited to submit questions based on patient cases on the following key topics: (1) challenges faced in achieving glycemic control, (2) applying individualized glycemic targets according to patient characteristics, and (3) treatment choices for achieving glycemic targets safely. Canvassing through social media took place using Facebook, LinkedIn, and Twitter, with announcements targeting relevant HCPs, identified using keywords in their profiles linked to diabetes and endocrinology. Canvassing through organic social media involved nontargeted announcements on the same 3 channels, as well as on touchENDOCRINOLOGY and touchCARDIO websites. In addition, an announcement was sent directly to 11,586 HCPs who had subscribed to emails from the touchENDOCRINOLOGY and touchCARDIO sites. No financial incentives were provided to submit the questions. For each video, 3 questions were developed by the faculty for discussion. The questions were based on the precanvassing and learning objectives of the activity.

The second activity, a touchMDT, featured an endocrinologist, a primary care physician (PCP), a nurse specializing in diabetes, and a patient with T2D. This study focused on the relationship between T2D and obesity, and the learning objectives were to (1) describe the relationship between T2D and obesity, (2) predict the beneficial effects of weight loss with glucagon-like peptide-1 (GLP-1) receptor agonist (GLP-1 RA)–based therapy and/or sodium-glucose cotransporter 2 inhibitor (SGLT2i) therapy on outcomes in patients with T2D and obesity, and (3) perform appropriate selection of antihyperglycemic therapy with weight loss benefits for patients with T2D and obesity. Each of the three 10-minute discussions involved a different combination of clinicians and the patient and was based on 3 or 4 discussion points focused on the practical management of patients with T2D and obesity from both the clinicians’ and the patient’s perspectives.

Both activities are available as free access on the touchENDOCRINOLOGY website (Multimedia Appendix 1) [38], a web-based HCP education community, from October 2021 to October 2023. The maximum 24-month viewing period was set in accordance with ACCME requirements. To reach a global target audience of HCPs specializing in diabetes, endocrinology, or primary care, a combination of communication channels was used, including emails—using touchMAIL, touchIME’s proprietary software—to touchENDOCRINOLOGY subscribers within the first 12 weeks and then 6 months after activity launch; medical society partnerships (website publicity) throughout the lifetime of activity; and HCP-targeted social media announcements on Facebook, LinkedIn, and Twitter throughout the lifetime of activity. The announcements on these social media platforms were paid. No financial incentives were provided to participate in this activity.

CME accreditation for both activities was provided by the University of South Florida Health, which is accredited as a provider of continuing professional development by the ACCME and the American Nurses Credentialing Center.

Assessment of Educational Outcomes

Outcomes for both activities were assessed according to the Moore expanded outcomes framework (levels 1-5) [34]. For levels 1 to 4 (participation, satisfaction, knowledge and competence), the outcomes were assessed for the 2 activities independently. For level 5 (performance), outcomes were assessed for the 2 activities combined to account for the overlap in content, learning objectives (ie, treatment selection and intensification for patients with T2D and overweight or obesity), and target audience.

Level 1 was assessed over the first 6 months after launch as 2 variables: the number of participants who engaged in the activity and the average time spent by participants viewing the videos. Google Analytics was used to capture geolocation, participant numbers, and the overall average time HCPs spent on the activity. Data on specialty and the country from which participants connected were collected from HCPs who viewed the activity using their touchENDOCRINOLOGY account and from learners who completed the level 3 and level 4 outcome questionnaires.

Levels 2 to 5 were assessed using the outcome questionnaires. To avoid bias, all data from the level 2 to level 5 questionnaires were collected by an independent third-party vendor (nuaxia Limited) that was not involved in the development of the activities. A target audience was specified for fielding the questionnaires so that the sample was taken from relevant respondents (HCPs who completed the preactivity questionnaire) and learners (HCPs who participated in the activity and completed the postactivity questionnaire). Financial incentives were provided by nuaxia Limited for the HCPs to complete the questionnaires. For both activities, the target audiences were predefined by specialty (diabetologists, endocrinologists, and primary care specialists) and country (France, Germany, Italy, Spain, and the United States). A database of 203,744 HCPs was sampled based on a predefined target audience. To avoid any pre-exposure bias and obtain a statistically representative sample size, data were collected using an independent sample model both before and after the launch of each activity. All questionnaires were fielded to the database and then closed once a prespecified number of HCPs responded. Levels 3 and 4 are assessed using a single questionnaire. Preactivity scores were obtained by fielding this questionnaire 1 to 2 weeks before launch (to ensure that the sample was from HCPs who had not interacted with the activity) and were closed after 50 respondents had completed it. Postactivity scores were obtained by fielding this questionnaire to another set of HCPs immediately after the launch and closed after 50 learners responded. The level 2 questionnaire assessing satisfaction with the activities was included with the level 3 and level 4 questionnaire that was fielded after the activity. For level 5, the questionnaire was fielded 1 to 2 weeks before the launch—to a different set of HCPs to those who answered the level 3 and level 4 questionnaires—and was closed after 50 respondents had
completed it. At 26 weeks after the launch, the level 5 questionnaire was administered to the same 50 learners who had responded before the activity; of these, 34 (68%) responded to the postactivity fielding, and data are presented as paired samples from these 34 learners only. For levels 2 to 5, the learners who responded to the postactivity questionnaires viewed the activity as part of the questionnaire process.

The level 2 satisfaction questionnaire included the following 5 statements that were to be scored using a 1- to 5-point Likert scale (where 5 is the highest satisfaction): this activity was of high quality, this activity met the stated learning objectives, the presenters were knowledgeable and effective, the activity contained content relevant to my clinical practice, and the information presented is likely to help change my management strategies in this therapeutic area. Levels 3 and 4 (knowledge and competence) and level 5 (performance) were assessed using questionnaires developed by the medical directors at touchIME and approved for scientific and medical accuracy by the faculty (Textbox 1). Satisfaction data were collected from participants immediately after engaging in the activity and before answering the level 3 and level 4 questionnaires. The level 3 and level 4 questionnaires comprised 7 questions for touchIN CONVERSATION (Table S2 in Multimedia Appendix 2) and 6 questions for touchMDT (Table S3 in Multimedia Appendix 2). All the questions were multiple-choice, with 3 to 4 possible answers, of which only one was correct. Data were analyzed for the overall participant groups and in subgroups defined by country, specialty, and years of experience. The level 5 questionnaire is a subjective assessment based on self-reported change in performance. It included 4 multiple-choice questions with 4 possible answers. All answers were plausible, but one was the best possible clinical option (Table S4 in Multimedia Appendix 2).

Textbox 1. Topics included in the level 3, level 4, and level 5 outcome questionnaires.

### Levels 3 and 4

To assess levels 3 and 4, separate questionnaires were developed for touchIN CONVERSATION and touch MultiDisciplinary Team activities

- touchIN CONVERSATION
  - Factors contributing to clinical inertia
  - Achieving glycemic control in nonadherent patients
  - Selecting individualized glycemic targets and add-on therapies
  - Emerging therapies for patients with type 2 diabetes (T2D) and overweight or obesity
- touchMDT (touch MultiDisciplinary Team)
  - Mechanisms linking obesity to T2D
  - Benefits of weight loss for T2D prevention
  - Clinical benefits of glucagon-like peptide-1 receptor agonists (GLP-1 RAs) and sodium-glucose cotransporter-2 inhibitors (SGLT2is)
  - Treatment intensification after a GLP-1 RA or metformin

### Level 5

To assess level 5, a single questionnaire was sent to the respondents of both activities

- Appropriate second-line treatment selection for patients with T2D and overweight or obesity
- Eligibility criteria for treatment with a GLP-1 RA
- Outcomes expected for patients treated with an SGLT2i
- Treatment intensification in patients with T2D, obesity, and atherosclerotic cardiovascular disease, who have not achieved their glycemic target

### Intention to Change Practice

To assess the impact of the educational activities on HCPs’ willingness to change their clinical practice, learners who took part in the level 2 to level 5 questionnaires after participating in the activities were asked the following multiple-choice question: “As a result of your participation in this session, will you make a change in your practice?” There were 5 possible mutually exclusive responses: yes, uncertain—more education needed, uncertain—practical limitations, no—more education needed, and no—practical limitations.

### Identification of Remaining Educational Gaps

To collect information on the learners’ perspective on the need for further education in the management of T2D, those who completed the level 2 to level 5 questionnaires after the activity were asked the question, “What do you think is the most important unmet educational need in this therapy area?” They were required to rank 4 predefined potential educational gaps (12 in total over the 3 questionnaires) by importance. Potential educational gaps were drafted by the medical directors at touchIME, with input from the faculty on the respective activities, and were included at the end of the questionnaires after the multiple-choice questions. The results were analyzed using a single transferable vote system. In the first round of
voting, all first-choice votes were counted to determine the most important educational gap for the participants; in the second round, all second-choice votes were counted to determine the second most important educational gap. Any first-choice vote, not from the winning option in the first round, was also counted in the second round. The voting rounds continued until all options were placed in order. In addition, questions in the level 3 and level 4 questionnaires that were answered incorrectly by ≥40% of learners after completion of the activity were identified as outstanding educational gaps.

Statistical Analysis

Data were analyzed using SPSS Statistics (version 28.0.1; IBM Corp). On the basis of target population of learners and the sample size, a statistical power calculation was used to determine the number of respondents (N=50) and learners (N=50) required to detect a statistically significant difference between surveys conducted before and after the activity, with a margin of error of approximately 10% for both touchIN CONVERSATION and touchMDT. For the satisfaction (level 2) questionnaire, the mean scores were calculated for the individual questions, and an overall satisfaction score was calculated as the average across all satisfaction fields, with a maximum possible satisfaction score of 5 points out of 5. For the knowledge and competence (levels 3 and 4) analysis, the mean and median numbers of correct answers were calculated for both the pre- and postactivity data sets, and the results were compared using an independent sample 2-tailed t test. To analyze the results by country, specialty, and experience, 2-way ANOVA was used. Individual questions were first analyzed using a paired sample t test and then using 1-way ANOVA. Data collection for performance (level 5) was performed using a matched sampling method. Pre- and postactivity data were compared using a paired sample t test. Country, specialty, and experience analyses were conducted using the same methods as for levels 3 and 4 using 2-way ANOVAs.

Ethics Approval

The faculty for touchIN CONVERSATION and touchMDT consented to the necessary use, distribution, and reproduction of their contribution to the activities and assigned the entire copyright and all other intellectual property rights existing in their contributions to touchIME. According to the European Union General Data Protection Regulation [39], HCPs who responded to the outcome questionnaires were informed before their input that, as with all research, their identity and personal data were strictly confidential and would not be revealed without their explicit further consent. This study did not report experiments on human participants; therefore, institutional review board approval and informed consent were not applicable.

Results

Assessment of Educational Activities

Level 1—Participation

Data collected between 6 and 7 months after launch showed that 6759 and 5998 participants had engaged with the touchIN CONVERSATION and touchMDT activities, respectively. The average length of participation was 8.50 minutes for the touchIN CONVERSATION and 13.09 minutes for the touchMDT (Table 1). For both activities, most participants specialized in endocrinology or diabetes, with only a small proportion working in primary care. Most participants were physicians (8869/12,757, 69.5%, for both activities combined), with the remainder being either nurse practitioners (2551/12,757, 20%) or physician assistants (1335/12,757, 10.5%). Participants from 25 countries engaged in each activity. The largest proportion of HCPs who engaged in the touchIN CONVERSATION activity was based in the United States, followed by the Philippines and Italy. The largest proportion of HCPs who participated in the touchMDT activity were based in the United Kingdom, followed by the United States, Portugal, and Italy. All other countries were represented by fewer than 10% of participants for each activity (Table 1).
Table 1. Engagement results and demographics of participants in the touchIN CONVERSATION and touch MultiDisciplinary Team (touchMDT) activities.

<table>
<thead>
<tr>
<th></th>
<th>touchIN CONVERSATION</th>
<th>touchMDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant engagement, n</td>
<td>6759</td>
<td>5998</td>
</tr>
<tr>
<td>Countries reached, n</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Length of participation (minutes), mean (SD)</td>
<td>8.50</td>
<td>13.09</td>
</tr>
<tr>
<td>Specialty, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endocrinology</td>
<td>3109 (46.00)</td>
<td>3178 (52.98)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>2500 (36.99)</td>
<td>1979 (32.99)</td>
</tr>
<tr>
<td>Primary care</td>
<td>1149 (17.00)</td>
<td>839 (13.99)</td>
</tr>
<tr>
<td>Not reported</td>
<td>1 (0.01)</td>
<td>2 (0.03)</td>
</tr>
<tr>
<td>Role, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician</td>
<td>4731 (70.00)</td>
<td>4138 (68.99)</td>
</tr>
<tr>
<td>Physician assistant</td>
<td>676 (10.00)</td>
<td>659 (10.99)</td>
</tr>
<tr>
<td>Nurse practitioner</td>
<td>1352 (20.00)</td>
<td>1199 (19.99)</td>
</tr>
<tr>
<td>Country, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>2064 (30.54)</td>
<td>1319 (21.99)</td>
</tr>
<tr>
<td>Philippines</td>
<td>1323 (19.57)</td>
<td>N/A</td>
</tr>
<tr>
<td>Italy</td>
<td>872 (12.90)</td>
<td>574 (9.57)</td>
</tr>
<tr>
<td>India</td>
<td>507 (7.50)</td>
<td>100 (1.67)</td>
</tr>
<tr>
<td>Bangladesh</td>
<td>466 (6.89)</td>
<td>N/A</td>
</tr>
<tr>
<td>Australia</td>
<td>242 (3.58)</td>
<td>148 (2.47)</td>
</tr>
<tr>
<td>Pakistan</td>
<td>222 (3.28)</td>
<td>N/A</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>161 (2.38)</td>
<td>1470 (24.51)</td>
</tr>
<tr>
<td>Spain</td>
<td>137 (2.03)</td>
<td>338 (5.64)</td>
</tr>
<tr>
<td>Ireland</td>
<td>107 (1.58)</td>
<td>N/A</td>
</tr>
<tr>
<td>Portugal</td>
<td>N/A</td>
<td>657 (10.95)</td>
</tr>
<tr>
<td>Canada</td>
<td>N/A</td>
<td>280 (4.67)</td>
</tr>
<tr>
<td>Netherlands</td>
<td>N/A</td>
<td>149 (2.48)</td>
</tr>
<tr>
<td>Mexico</td>
<td>N/A</td>
<td>116 (1.93)</td>
</tr>
</tbody>
</table>

aData collected on April 22, 2022, and at 203 and 190 days after the launch of touchIN CONVERSATION and touchMDT, respectively.

bCountry where the participant was based at the time of completing the activity. Data are reported for countries represented by ≥2% of the participants for at least one activity.

cN/A: not applicable.

Level 2—Satisfaction
The overall satisfaction scores were 84% (4.2/5) for touchIN CONVERSATION and 82% (4.2/5) for touchMDT. For touchIN CONVERSATION and touchMDT, respectively, satisfaction scores of 4.2 and 4.1 for the quality of the activity, 4.2 and 4.2 for meeting the stated learning objectives, 4.3 and 4.3 for the knowledge and effectiveness of the presenters, 4.4 and 4.1 for relevance to clinical practice, and 4.0 and 3.8 for impact on management strategies were achieved out of a maximum score of 5.0.

Levels 3 and 4—Knowledge and Competence
Before the launch of the touchIN CONVERSATION activity, 22% (11/50) of the respondents answered at least 6 of the 7 questions of the level 3 and level 4 questionnaires correctly, whereas after participating in the activity, this increased to 60% (30/50). There was a significant increase in the average number of correctly answered questions from before to after the activity (median, IQR 4.5, 3.0-5.0 vs 6.0, 4.75-6.0; mean, SD 4.36, 1.40 vs 5.42, 1.37; P<.001). These results are shown in Figure 1, where heat maps on the left show the proportion of respondents (n=50) and learners (n=50) who answered a specific number of questions correctly, as displayed by colors ranging from white (lowest proportion of respondents and learners) to dark red.
The box-and-whisker plots on the right show the distribution of the number of correctly answered questions by all respondents and learners. In both plots, the horizontal red line within the box indicates the IQR, the “x” symbol represents the mean, the boxes indicate the IQR, and the vertical lines (whiskers) extend to the range of values, excluding outliers. Outliers are defined as values that fall outside a distance of $1.5 \times$ the IQR from the upper and lower quartiles and are represented by empty circles. Respondents and learners were HCPs who completed the pre- and postactivity questionnaires, respectively.

There was also improved knowledge from before to after the activity in the selection of individualized glycemic targets for older patients (22/50, 44% answered correctly before the activity vs 35/50, 70% answered correctly after the activity) and of emerging therapies for T2D and obesity (31/50, 62% vs 44/50, 88%). In addition, improved competence in selecting individualized glycemic targets for younger patients (13/50, 26% vs 21/50, 42%) and in selecting an add-on therapy for patients at a high risk of cardiovascular disease (35/50, 70% vs 45/50, 90%) was observed (Figure 2, where the bar graphs show the percentage of respondents (n=50) and learners (n=50) who answered each question correctly. The numbers within the bars indicate their values. Respondents and learners were HCPs who completed the pre- and postactivity questionnaires, respectively).

For the touchMDT, there was no significant increase in the number of correct answers from before to after the launch of the activity (median, IQR 5.0, 4.0-5.0 vs 5.0, 4.0-5.0; mean, SD 4.36, 1.24 vs 4.58, 1.07; $P=.35$; Figure 1). Notably, 80% (40/50) of the respondents answered at least 4 of the 6 questions correctly before the activity was available, indicating high baseline knowledge and competence in this cohort (Figure 1). This increased to 86% (43/50) after the launch, with the greatest improvement observed in competence in treatment intensification after GLP-1 RA treatment (34/50, 68% vs 38/50, 76%; see Figure 3, where the bar graphs show the percentage of respondents (n=50) and learners (n=50) who answered each question correctly. The numbers within the bars indicate their values. Respondents and learners were HCPs who completed the pre- and postactivity questionnaires, respectively).

For both activities, the change in the mean number of questions answered correctly was similar across countries (touchIN CONVERSATION, $P=.36$; touchMDT, $P=.15$; Figures S1 and S2 in Multimedia Appendix 2) and years of experience (touchIN CONVERSATION, $P=.51$; touchMDT, $P=.90$; Figures S1 and S2 in Multimedia Appendix 2). For both activities, respondents specializing in primary care had the lowest mean scores at baseline (touch IN CONVERSATION: 3.62; touchMDT: 4.06; Figures S1 and S2 in Multimedia Appendix 2). Furthermore, a significant difference in the change in the mean number of questions answered correctly was observed between specialties (touchIN CONVERSATION, $P=.03$; touchMDT, $P=.03$), with primary care and diabetes specialists showing the largest increase in scores following the touchIN CONVERSATION and touchMDT activities, respectively (Figures S1 and S2 in Multimedia Appendix 2).

**Figure 1.** Summary of the number of correct responses for the level 3 and level 4 outcome questionnaires before and after the launch of touchIN CONVERSATION and touchMDT (touch MultiDisciplinary Team).
Figure 2. Summary of correct responses for individual topics for the level 3 and level 4 outcome questionnaires before and after the launch of touchIN CONVERSATION. CV: cardiovascular; T2D: type 2 diabetes.

Figure 3. Summary of correct responses for individual topics for the level 3 and level 4 outcome questionnaires before and after the launch of touchMDT. CV: cardiovascular; GLP-1 RA: glucagon-like peptide-1 receptor agonist; SGLT2i: sodium-glucose cotransporter-2 inhibitor; T2D: type 2 diabetes; touchMDT: touch MultiDisciplinary Team.
Level 5—Performance

Before the launch of the touchIN CONVERSATION and touchMDT activities, 32% (11/34) of the respondents selected the answer representing the best clinical option for all 4 questions. This increased to 59% (20/34) after participating in the activities (Figure 4; where the heat map on the left shows the proportion of learners (n=34) who answered a specific number of questions by selecting the best of 4 clinical options before and after viewing the activities, as displayed by colors ranging from white (lowest proportion of respondents and learners) to dark red (highest proportion of respondents and learners). The box-and-whisker plot on the right shows the distribution of the number of questions answered by selecting the best clinical option by all learners before and after viewing the activity. The horizontal red line within the box indicates the median, the “x” symbol represents the mean, the box indicates the IQR, and the vertical lines (whiskers) extend to the range of values, excluding outliers. Outliers are defined as values that fall outside a distance of 1.5× the IQR from the upper and lower quartiles and are represented by empty circles). Overall, a significant increase in the number of best clinical options selected from before to after the activity was observed (median, IQR 3.0, 2.0-4.0 vs 4.0, 2.5-4.0; mean, SD 2.65, 1.32 vs 3.15, 1.26; P=.03; Figure 4). Improved performance from before to after the activity was observed for all 4 questions; in particular, questions related to treatment intensification for patients not achieving their glycemic target (23/34, 68% of the respondents gave the best clinical option response before the activity vs 28/34, 82% who gave the best clinical option response after the activity) and eligibility criteria for treatment with a GLP-1 RA (21/34, 62% vs 26/34, 76%; see Figure 5, where the bar graph shows the percentage of learners (n=34) who answered each question by selecting the best clinical option before and after viewing the activities. The numbers within the bars indicate their values).

Figure 4. Summary of responses for the level 5 outcome questionnaire before and after the launch of touchIN CONVERSATION and touchMDT (touch MultiDisciplinary Team).

Figure 5. Summary of correct responses for individual topics for the level 5 outcome questionnaire before and after the launch of touchIN CONVERSATION and touchMDT. ASCVD: atherosclerotic cardiovascular disease; GLP-1 RA: glucagon-like peptide-1 receptor agonist; SGLT2i: sodium-glucose cotransporter-2 inhibitor; touchMDT: touch MultiDisciplinary Team.
The change in the mean number of best clinical options selected was similar across the years of experience (P=0.66; Figure S3 in Multimedia Appendix 2). There was a statistically significant difference in the change in the mean number of the best clinical options selected across different countries (P=0.03; Figure S3 in Multimedia Appendix 2). Learners from France and Germany showed the lowest and highest mean numbers of best clinical options selected at baseline, respectively; this did not increase following the touchIN CONVERSATION and touchMDT activities for participants from either country (Figure S3 in Multimedia Appendix 2). As in the level 3 and level 4 questionnaires, primary care specialists had a lower mean score at baseline compared with specialists in endocrinology (2.00 vs 2.96), and primary care specialists showed the largest increase in best clinical options selected following the touchIN CONVERSATION and touchMDT activities (P=0.01; Figure S3 in Multimedia Appendix 2).

**Discussion**

**Principal Findings**

This study evaluated outcomes following 2 faculty-led, CME-accredited, web-based educational activities on the management of patients with T2D and demonstrated that HCPs expressed high levels of satisfaction and improvements in their knowledge and competence, as well as self-reported performance in T2D management. By the 6- to 7-month postlaunch time point, each activity had been viewed by a global audience of approximately 6000 participants, of which most were specialist physicians. Learners’ satisfaction levels with the educational activities were high, and they considered them to be relevant to clinical practice, meet the stated learning objectives, and impact their future management strategies.

In the touchIN CONVERSATION activity, learners successfully improved their knowledge, competence, and performance, and the benefits of setting individualized glycemic targets were identified as key future educational needs. In addition, most participants identified as key future educational needs. In addition, most

**Table 2.** Unmet educational needs identified by the touchIN CONVERSATION and touch MultiDisciplinary Team (touchMDT) learners.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Questionnaire Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>touchIN CONVERSATION (level 2 to level 4 questionnaires)</td>
<td>touchMDT (level 2 to level 4 questionnaires)</td>
</tr>
<tr>
<td>1. Use of time-in-range metrics in continuous glucose monitoring to help optimize glycemic control</td>
<td>1. Strategies for achieving sustained weight loss in patients with T2D and obesity</td>
</tr>
<tr>
<td>2. Managing treatment regimens to avoid hypoglycemia in patients with T2D</td>
<td>2. Understanding the data from cardiovascular and renal outcomes trials for antihyperglycemic medications in patients with T2D at high cardiovascular or renal risk</td>
</tr>
</tbody>
</table>

aThe top 3 unmet educational needs are shown, as identified by learners who completed the level 2 to level 4 and level 5 questionnaires following the launch of the touchIN CONVERSATION and touchMDT activities. Learners were required to rank 4 predefined, potential educational gaps in response to the question “What do you think is the most important unmet educational need in this therapy area?”
bT2D: type 2 diabetes.
learners confirmed that they would change their practice in response to participation, highlighting the clinical value of the activity. Although an improvement in self-reported performance after participating in the touchMDT activity was reported, no significant increase in knowledge and competence was observed, and fewer learners indicated an intention to change practice compared with the touchIN CONVERSATION activity. This may reflect the relatively high baseline knowledge and competence levels observed in this cohort. However, competence in advising patients on the clinical benefits of SGLT2i therapy was relatively low, with little improvement observed after the activity. This may suggest a requirement for further education or may reflect a bias based on clinical experience with this treatment class. The high baseline scores and subsequent lack of significant increases in knowledge and competence following the touchMDT may also indicate that the activity successfully consolidated the knowledge and competence gained from the earlier touchIN CONVERSATION, which addressed, in part, similar topics. Thus, it is possible that the respondents of the level 3 and level 4 questionnaires for the touchMDT partially overlapped with the learners from the touchIN CONVERSATION, as the target audience and geographies were identical, and participants were reached through the same channels. This interpretation is consistent with the concept of spaced learning, according to which, re-exposing learners to information over time using temporal intervals results in more effective retention of information than if it was all provided at once [40].

Improvements in knowledge and competence were similar across countries and years of experience, indicating that education was beneficial for the entire range of HCPs. Some numerical differences were observed between countries for the level 3 and level 4 questionnaires, with improvements in knowledge and competence seen for learners in the United States, but not in France, Germany, Italy, or Spain following the touchMDT activity. When the impact of education on learners’ performance was assessed, there was a statistically significant difference in performance improvement across different countries, although learners from France and Germany showed no increase in the mean number of the best clinical options selected. Interestingly, although the lack of increase in performance in learners from Germany could be attributed to the high mean of the best clinical options selected at baseline, leaving little room for further increase, learners from France did not show any increase in performance despite showing the lowest score at baseline. Because of the small size of the subgroups, we cannot speculate on the potential reasons for this; further studies with larger groups of participants from these countries are required to obtain meaningful insights. Significant differences were also noted between specialties for the level 3, level 4, and level 5 questionnaires, with the largest improvements observed for primary care specialists. Overall, the HCPs in primary care demonstrated the lowest levels of knowledge, competence, and performance before both activities. This highlights the importance of ongoing education to ensure that primary care teams remain up to date with the rapidly evolving treatment and management landscape of T2D. The high participation numbers and satisfaction scores, combined with improvements in knowledge, competence, and performance observed following one or both activities, support the educational approach of (1) pre-canvasing prospective learners and using specific patient cases to ensure that the activity is immediately relevant to HCPs’ daily practice and (2) using multidisciplinary faculty and real-life patients to deliver the educational activity.

Most learners indicated that they would implement changes in clinical practice because of their educational activities. Nonetheless, a notable proportion of HCPs who participated in the activities stated that they would not change their practices at the current time. Although patient-centered care is increasingly becoming the focus of health care improvement, several barriers still exist that may prevent its application in daily clinical practice and on a larger scale across health care organizations [41]. These barriers include lack of organizational culture shift, flawed communication and leadership strategies, and practical limitations such as recently updated guidelines, accessibility to emerging drugs and practices, and cost-associated factors [41]. In addition, a recent study assessing factors that influence HCPs’ intention to put newly acquired knowledge into practice identified a lack of belief in one’s own capabilities (i.e., the belief that one is capable of performing the behavior) among the barriers to adopting changes in clinical practice. The consequences of adopting new clinical behaviors were also cited as a key barrier [42].

The results of this study indicated that several educational needs remain. As indicated by the questions that were frequently answered incorrectly after the activity, there appears to be a need for more education on setting individualized glycemic targets, particularly in younger patients, and a need to further understand the potential benefits of SGLT2i therapies. In addition, learners self-selected several educational needs, including the use of time-in-range metrics for continuous glucose monitoring, strategies to avoid hypoglycemia, how to achieve sustained weight loss, understanding data from cardiorenal outcome trials, efficacy data for emerging therapies, and managing the side effects of antihyperglycemic medications.

**Strengths and Limitations**

The strengths of this study were, first, the involvement of prospective learners in the development of the touchIN CONVERSATION and the provision of a multidisciplinary program, the touchMDT, delivered by and for an interdisciplinary team of HCPs. Both aimed to maximize HCP participation, engagement, and satisfaction with the program. Second, the activities were accredited; thus, physicians, physician assistants, and nurses could obtain CME credit through participation in education. Third, there were no barriers to accessing education, with both activities made freely available on the touchENDOCRINOLOGY website. Fourth, the outcome questionnaire data were collected using an independent sample model, and the questionnaire was fielded to a statistically representative sample. All data collection was carried out by an independent third-party vendor.

This study had several limitations. First, self-selection bias must be considered when assessing the impact of educational activities; thus, HCPs who consider their knowledge to be lacking in these topics are more likely to participate than those who consider their knowledge to be up to date. This bias...
generally affects medical education, irrespective of the format or delivery method used. To mitigate self-selection bias, we used a combination of different channels to reach the HCP target audience for the activities described here: not limited to clinicians actively seeking medical education, but extended to a heterogeneous population of HCPs, including social media subscribers and members of professional societies. Second, the long-term benefits of educational activities remain unknown. In the future, measuring the impact of education over a longer time frame (eg, at 12 and 24 months) may be beneficial, although the treatment paradigm for T2D is relatively fast-moving, and measuring the impact beyond 24 months may not prove insightful owing to changes in clinical practice. Rather, providing updates to the education based on feedback from learners and the results of the outcomes analysis would be more practical and would ensure that HCPs are kept up to date with information that is useful and relevant. Third, aggregated rather than matched data were used for the level 3 and level 4 questionnaires; however, a previous study of CME outcomes indicated that aggregated data are comparable with matched data and are therefore likely to be sufficiently accurate for many program evaluation purposes [43]. Fourth, as with any analysis of this type, subgroup analyses were limited by the small size of the subgroups, and as such, may not be generalizable to a larger population of HCPs. In future studies, a larger sample size may increase the statistical power of subgroup analyses. Fifth, when assessing the intention to change practice because of the education, a proportion of learners indicated practical limitations as an obstacle to applying changes to their daily practice; however, our questionnaire did not probe the nature of these practical limitations, but it would be beneficial to collect this information in future learning activities to assess whether any of these barriers can inform future education. Sixth, level 5 outcomes were measured subjectively (ie, they were based on self-reported performance rather than on observed changes in patient management). In future studies, a similar approach assessing self-reported performance could be paired with objective evaluations, such as the collection of anonymized patient records, to confirm whether self-assessment is predictive of objective improvement in HCPs’ performance and to provide a more detailed understanding of the impact of CME activities on HCPs’ performance and the health status of the population with diabetes.

Comparison to Prior Work
CME accreditors are placing increasing importance on the measurement of higher-level outcomes following participation in educational activities; however, data on outcomes from web-based CME activities in T2D are limited [44]. Several studies have demonstrated that traditional CME programs for diabetes can lead to improvements in clinical practice and patient outcomes [10,16,17,20]. However, these studies focused on more time-consuming activities, including face-to-face and live educational sessions, which may make it difficult for HCPs to fit into their schedules. In addition, HCP performance and patient outcomes are often assessed using costly and labor-intensive methodologies, such as objective structured clinical examination stations and patient chart reviews, which are not always practical for assessing outcomes from smaller-scale web-based education. The quality of CME is frequently measured using the Moore level of outcome framework, but studies assessing the impact of short web-based CME programs on knowledge, competence, and performance (Moore levels 3, 4, and 5) are limited [37]. There is growing interest in the use of short activities that encourage learner involvement to provide a more convenient and potentially more effective approach to ongoing education for HCPs, as supported by a small but growing body of evidence. For example, the use of short (15-minute) web-based educational sessions resulted in increases in physician knowledge and competence in studies on diabetes [19] and thrombotic thrombocytopenic purpura [45]. Similarly, the value of learner involvement in T2D education has been demonstrated in a randomized controlled trial, which reported that greater improvements in self-reported competence in diabetes management could be achieved through the use of an educational, case-based game compared with a series of face-to-face lectures and group discussions [15]. Furthermore, a pilot study demonstrated that physicians gained confidence and achieved improved performance in test diabetes cases following participation in an hour-long lecture combined with patient cases in a virtual-world setting [18]. The potential benefits of having learners contribute to the development of educational activities were shown in a pilot study, which demonstrated that a web-based educational activity in attention-deficit/hyperactivity disorder cocreated by PCPs was well received by an audience of other PCPs, although no evidence of its efficacy was available [46]. In this study, we further contributed to this growing body of evidence and demonstrated that web-based CME activities, which can be undertaken in short, easy-to-access sessions, can lead to improvements in HCP performance, as measured by a self-reported questionnaire.

Despite the limitations outlined earlier, the overall objectives of this analysis were met: the study demonstrated significant improvements in the knowledge, competence, and performance of HCPs in the management of T2D and obesity following participation in one or both activities; key outstanding educational gaps were identified; and areas for the improvement of educational outcomes assessment were highlighted, which, alongside the knowledge gained on key educational needs, may help to inform future activities that maximize HCP performance and ultimately patient outcomes.

Conclusions
This study demonstrated that short, case-based, patient-focused, and multidisciplinary team–led CME activities that HCPs can fit into their clinical schedules achieved high levels of satisfaction and improvements in HCP knowledge and competence, along with self-reported performance in T2D management. Ongoing educational needs identified included setting individualized glycemic targets, particularly in younger patients, and the potential benefits of SGLT2i therapies. These educational needs can be used to inform future educational activities in the diabetes HCP community. The activities described in this study reduce barriers to participation in CME activities, as they are convenient and easily accessible to learners and are free to access.
Acknowledgments

The authors would like to thank the faculty members Professor Dr med Michael Nauck (St. Josef Hospital [Ruhr-Universität Bochum], Bochum, Germany) and Dr Vincent Woo (University of Manitoba, Winnipeg, Manitoba, Canada) for contributing to the educational content of the touchIN CONVERSATION and touch MultiDisciplinary Team activities, as well as the touch Independent Medical Education (touchIME) audience outreach team (Joel Turner and Hannah Morton-Fishwick) for the collection and analysis of the level 1 data. Medical writing support was provided by Steph Carter, and editorial support was provided by Joanna MacDiarmid, both of touchIME, Stockport, Cheshire, United Kingdom. Educational activities were jointly provided by the University of South Florida Health and touchIME. Eli Lilly provided an independent medical education grant to fund educational activities and medical writing and editorial support but did not contribute directly to manuscript development, the decision to submit, or the submission process.

Data Availability

Data will be shared with bona fide researchers submitting a research proposal to touch Independent Medical Education Ltd. Access requests should be submitted to ADN (Alex.Noble@touchime.org). Data will be made available from the study publication date. Any individual participant data will be shared in data sets in a deidentified and anonymized format.

Authors’ Contributions

All authors contributed to study design and manuscript writing. ADN performed statistical analyses.

Conflicts of Interest

SBH reports advisory board or panel fees from Abbott, AstraZeneca, Bayer Inc, Eli Lilly, Janssen, Novo Nordisk, and Sanofi; consultancy fees from Abbott, AstraZeneca, Bayer Inc, Eli Lilly, Janssen, Novo Nordisk, and Sanofi; grants and research support from Abbott, AstraZeneca, Eli Lilly, Janssen, Novo Nordisk, and Sanofi; and salary and contractual services from Abbott, AstraZeneca, Bayer Inc, Eli Lilly, Janssen, Novo Nordisk, and Sanofi. AB, SQN, ADN, and HES are employees of touch Independent Medical Education Ltd. JV reports grants and research support from Eli Lilly, Eliem Therapeutics, Lexicon, Novo Nordisk, Pfizer, and Sanofi.

Multimedia Appendix 1
Images of the touchIN CONVERSATION and touch MultiDisciplinary Team activities.

Multimedia Appendix 2
For the touchIN CONVERSATION and touch MultiDisciplinary Team activities: Tables showing educational gaps and learning objectives, and questions included in the level 3, level 4 and level 5 outcome questionnaires; and figures showing a summary of responses to the questionnaires.

References


38. CME. touchTM Endocrinology. URL: https://touchendocrinologyime.org/ [accessed 2022-06-28]


42. Bergeron L, D


Abbreviations

ACCME: Accreditation Council for Continuing Medical Education
CME: continuing medical education
GLP-1 RA: glucagon-like peptide-1 receptor agonist
HbA1c: glycated hemoglobin
HCP: health care professional
PCP: primary care physician
SGLT2i: sodium-glucose cotransporter-2 inhibitor
T2D: type 2 diabetes
touchIME: touch Independent Medical Education
touchMDT: touch Multidisciplinary Team
Defining a Role for Webinars in Surgical Training Beyond the COVID-19 Pandemic in the United Kingdom: Trainee Consensus Qualitative Study

Emma Barlow¹*, BSc, MBChB, PGDip; Wajiha Zahra²*, MBBS, MSc; Jane Hornsby¹, BSc, MBChB; Alex Wilkins¹, BMBS, BMedSci; Benjamin M Davies³, BSc, MBChB, MPhil; Joshua Burke¹, BSc, MBChB

¹Association of Surgeons in Training, London, United Kingdom
²Department of Trauma and Orthopaedics, University Hospitals North Midlands, Stoke-on-Trent, United Kingdom
³Department of Neurosurgery, University of Cambridge, Cambridge, United Kingdom
*these authors contributed equally

Corresponding Author:
Benjamin M Davies, BSc, MBChB, MPhil
Department of Neurosurgery
University of Cambridge
Box 167, Cambridge Biomedical Campus
Addenbrooke’s Hospital
Cambridge, CB2 0QQ
United Kingdom
Phone: 44 207973030
Email: bd375@cam.ac.uk

Abstract

Background: The COVID-19 pandemic posed several challenges for surgical training, including the suspension of many in-person teaching sessions in lieu of webinars. As restrictions have eased, both prepandemic and postpandemic training methods should be used.

Objective: This study investigates trainees’ experiences of webinars during the COVID-19 pandemic to develop recommendations for their effective integration into surgical training going forward.

Methods: This project was led by the Association of Surgeons in Training and used an iterative process with mixed qualitative methods to consolidate arguments for and against webinars, and the drivers and barriers to their effective delivery, into recommendations. This involved 3 phases: (1) a web-based survey, (2) focus group interviews, and (3) a consensus meeting using a nominal group technique.

Results: Trainees (N=281) from across specialties and grades confirmed that the COVID-19 pandemic led to an increase in webinars for surgical training. While there were concerns, particularly around the utility for practical training (80.9%), the majority agreed that webinars had a role in training following the COVID-19 pandemic (90.2%). The cited benefits included improved access or flexibility and potential standardization of training. The majority of limitations were technical. These perspectives were refined through focus group interviews (n=18) into 25 recommendations, 23 of which were ratified at a consensus meeting, which was held at the Association of Surgeons in Training 2021 conference.

Conclusions: Webinars have a role in surgical training following the COVID-19 pandemic. The 23 recommendations encompass indications and technical considerations but also discuss important knowledge gaps. They should serve as an initial framework for ensuring that webinars add value and continue to evolve as a tool for training.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2200055325; http://www.chictr.org.cn/showprojen.aspx?proj=142802

(JMIR Med Educ 2022;8(4):e40106) doi:10.2196/40106

KEYWORDS
webinars; surgical training; trainee consensus; teaching; training; integration; trainee experience; user experience; perception; education; medical education; surgical skills
Introduction

The COVID-19 pandemic has significantly impacted many aspects of surgical training in the United Kingdom and the Republic of Ireland and has challenged the delivery of surgical training with the cancellation of examinations, training courses, and teaching programs, and a significant reduction in exposure to operative cases [1]. A study comparing surgical trainees’ operative logbook numbers in 2019 and 2020 reported an overall incident rate ratio (IRR) of 0.62, with exposure to elective surgery more affected (IRR 0.53) than emergency surgery (IRR 0.85) [2]. Subsequently, surgical training had to adapt, resulting in an accelerated shift in the use of digital learning environments by both trainees and trainers.

A webinar is defined as “a seminar conducted online” [3]. A seminar is a teaching method based on the Socratic dialogue of asking and answering questions with the word originating from the Latin *seminarium*, which means “seed plot” [4,5]. Webinars have evolved as a result of technological advancements leading to faster, more reliable internet connections and the use of video calls as a standard method of communication in today’s society. As a result, digital learning was slowly being integrated into pedagogical learning methods to create new blended learning methods [6]. However, the COVID-19 pandemic presented a new opportunity for the use of webinars to provide remote learning for surgical trainees, and, as a result, they have become a popular and increasingly prevalent training tool [4]. One review of the use of webinars for training in plastic surgery indicated an increase of 12,017% in the number of webinars relevant to this specialty post lockdown in the United Kingdom on the March 23, 2020 [4].

While the pandemic forced training to transition to the internet, the question remains if and how webinars should be integrated within surgical training going forward as the National Health Service’s services start to recover and surgical training adapts to the new norm. The Joint Committee on Surgical Training quality indicators for Higher Surgical Training states that trainees should have a minimum of 2 hours of facilitated formal teaching every week [7].

This study, led by the Association of Surgeons in Training (ASiT), aimed to identify the role of webinars in surgical training, both during and after the COVID-19 era, and consider what, in a best-case scenario, their potential integration into surgical training would look like in the future. To achieve this, the following objectives were set: (1) to investigate the strengths and weaknesses of webinar-based training for surgical specialties, (2) to establish the limitations of webinars and the barriers faced when integrating them into surgical training, and (3) to propose recommendations for its integration within surgical training.

Methods

Overview

This study was led by the ASiT via a steering committee of surgical trainees. The ASiT is an independent professional organization (registered charity number 1196477), with a membership of over 3500 surgical trainees, that promotes excellence in surgical training. The organization represents and supports trainees across all surgical specialties and training grades on a regional and national level in the United Kingdom and the Republic of Ireland and is the largest representative body for surgical trainees. The ASiT was originally founded in 1976 and is independent of, but works with, the National Health Service, General Medical Council, Surgical Royal Colleges, the Joint Committee on Surgical Training, and the Trainee Surgical Specialty Associations.

This study used mixed qualitative methods in accordance with COREQ (consolidated criteria for reporting qualitative research) guidelines [8] to consolidate the arguments for and against webinars and the drivers and barriers to their delivery in surgical training. This involved 3 phases: (1) a web-based survey, (2) focus group interviews, and (3) a consensus session at the annual ASiT 2021 international conference.

Phase 1: Web-Based Survey

A survey was developed to gather wider trainee perspectives on webinars in surgical training. It was built using Qualtrics (Qualtrics XM) in accordance with the published CHERRIES (Checklist for Reporting Results of Internet E-Surveys) guidelines on conducting web-based surveys with the aim of establishing broad opinion and key themes concerning the use of webinars before and during the COVID-19 pandemic [9-11]. The survey contained binomial, multiple-choice, Likert-scale questions and those with free-text responses (Multimedia Appendix 1) and was developed on the basis of consensus from the steering group. The survey was peer reviewed and piloted by the ASiT council prior to dissemination, where 1 question was omitted and 2 questions reworded, resulting in a final set of 26 questions. The survey was sent out to all ASiT members via the ASiT mailing list (MailChimp) and advertised via ASiT social media channels (Facebook and Twitter). Survey completion was voluntary and open to current and future surgical trainees of all grades and specialties. All responses were collected with informed consent provided by responders at the time of completion, and anonymized over a 6-week period (December 3, 2020, to January 14, 2021) without sampling. Thereafter, the survey was removed from the ASiT website. Descriptive statistics were used to aggregate data from predefined questions. Responses to a final free-text question—“Please feel free to write any other comments regarding your experience of teaching/training via webinars.”—were subjected to inductive thematic analysis by 2 authors (AW and JH) independently.

Phase 2: Focus Group Interviews

Trainees interested in participating in the interview stage were contacted via email with further study information and available interview slots between Monday, February 8, and Sunday, February 14, 2021. The interviews were arranged as 30-minute focus group sessions with up to 4 participants in each group and one member of the study group present to facilitate the session, prompt discussions, and answer queries. These sessions were conducted as web-based video calls using Zoom (Zoom Video Communications). Each session was recorded and transcribed with participants consent. Interviews were conducted...
using structured discussion. Interview transcripts and recordings were analyzed using thematic analysis, which produced 4 themes including strengths, limitations, drivers, and barriers, with focused questions generated through steering group consensus from the initial survey responses. All focus group participants were offered collaborative authorship [12].

**Phase 3A: Formation of Consensus Statements**

This project used a transparent consensus process [13]. Based on the findings of the survey and focus group interviews, a preliminary list of consensus statements (Multimedia Appendix 2) was formed by the steering committee (EB, WZ, JH, AW, BD, and JB), and refined with input from the ASiT council, into a final list for presentation during the consensus meeting (Multimedia Appendix 3). These were designed to capture the perspectives gathered so far and structured around the following core areas: the role, timing, conduct, content, opportunities, and knowledge gaps for webinar use.

**Phase 3B: Consensus Meeting**

The consensus meeting was held at the 2021 ASiT annual international conference, which was the first ASiT conference to be held remotely. It was advertised as free for any conference delegate via the conference platform (MedAll) and program, the ASiT website [14], and social media platforms. A Google Form (Google LLC) was shared among attendees and used to record the name, gender, specialty, grade, and deanery of all participants who attended the meeting. The session was held using Zoom, and discussion during the meeting took place both verbally and using the question-and-answer function. Slido (sli.do s. r. o.) was used during the consensus session to present each consensus statement and permit attendees to vote. Votes were binary (yes/no or agree/disagree). For a statement to be ratified, 70% or more of the consensus participants had to vote in agreement (yes/agree). The cutoff of 70% was based on the Grading of Recommendations Assessment, Development and Evaluation approach and agreed on by the study group prior to the session [15,16]. The Slido software automatically stored the results and facilitated their export for data analysis.

**Ethics Approval**

According to the Health Research Authority Guidance, ethics approval was not sought for this study.

**Results**

**Phase 1: Web-Based Survey**

**Demographics**

The survey was sent to 2790 ASiT members. The total number of survey responses was 281, with a response rate of 9.9%. In total, 278 complete responses were included in the final analysis, as incomplete responses were excluded.

All training regions in the United Kingdom and the Republic of Ireland were represented. In total, 56.7% of respondents were female. The majority of respondents were trainees in general surgery (43.2%), followed by orthopedics (21.6%), with responses from trainees in all other surgical specialties. There were responses from core surgical trainees (41.0%), specialist registrars (35.6%), foundation trainees and medical students (17.9%), specialty doctors or associate specialists (3.8%), and post–certificate of completion of training fellows 1.7%; Multimedia Appendix 4).

**Key Findings**

Respondents reported attending more training webinars in 2020 than in 2019 (Figure 1) and 96.5% of respondents agreed that this change was due to the COVID-19 pandemic. Overall, 98.9% of respondents agreed that webinars have become a standardized format for surgical teaching to replace face-to-face teaching during the COVID-19 pandemic. However, 80.9% agreed that webinars cannot fulfill the practical aspects of surgical training.

Specific reasons for attending more webinars included webinars being the only resource available (28.9%), more webinars being available (23.4%), to meet specific training requirements (15.8%), more awareness of webinars (15.6%), and webinars being an effective way to use time (14.3%). Only 10 respondents did not attend webinars, and the reasons for this included that there are too many webinars, so it is difficult to choose which ones to attend, and that they seem less effective for surgical training. The main factors that are considered when deciding to attend a webinar include the topic (27.6%), training requirements (20.6%), the speaker (15.4%), and cost (15.6%).

Regarding the delivery of webinars (Multimedia Appendix 5), the preferred duration of a webinar for 69.1% of respondents was 30 to 60 minutes. The most popular time of webinars was weekdays out of working hours. The majority of respondents had 1-2 hours available to attend webinars each week.

Respondents were asked to rate aspects about the effectiveness of webinars on Likert scales. Overall, the mean score for the webinar as a format for surgical training during the COVID-19 pandemic was 6.8 (SD 1.8) out of 10. The organization and structure of webinars had a mean score of 7 (SD 1.86) out of 10. The interaction between speakers and participants had a mean score of 5.2 (SD 2.22) out of 10, and technical aspects scored 6.4 (SD 2.29) out of 10. Meeting the demand of practical education scored poorly at 4.3 (SD 2.75) out of 10. Most people (44.5%) felt that the topic of the webinar determined whether it would be as effective as face-to-face teaching. Overall, 91.2% of respondents felt that they were more likely to ask questions during face-to-face teaching than during a webinar and that they were more likely to pay for face-to-face teaching.

Going forward, 90.2% of respondents were very likely, likely, or somewhat likely to attend training webinar after the COVID-19 pandemic. Respondents thought that webinars and web-based learning would be most useful for exam preparation, journal clubs, annual reviews of competency progression), and supervisor meetings (Multimedia Appendix 5).

In total, 106 respondents provided free-text answers when asked if they had any other comments about webinars. The themes from these responses were identified as demonstrated in Table 1.

https://mededu.jmir.org/2022/4/e40106

JMIR Med Educ 2022 | vol. 8 | iss. 4 | p.68

(page number not for citation purposes)
Table 1. Themes identified from the free-text answers in the initial web-based survey.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Example quotes</th>
</tr>
</thead>
</table>
| Webinars improve accessibility to teaching | • “I would watch a webinar of an interesting topic given by an institution very far from where I am based. In fact, I may be even more likely to do than I would to attend physically as the costs would be much lower.”  
• “World experts can be called upon to give a webinar on a given topic.”  
• “A great equaliser for those with kids etc to be able to attend without travelling.” |
| Webinars are useful for examination preparation | • “I think the training webinars focusing on required curriculum (exam/FRCS style preparation) and related topics have been very useful and are as good as in person teaching.” |
| Webinars are less engaging | • “I am more prone to distraction and lack of engagement in webinars than in face-to-face teaching.” |
| Webinars cannot replace teaching for practical skills | • “A lot of practical/surgical stuff involving hands on technique cannot be taught via webinars and therefore webinars are in no way a substitute for face-to-face training but merely an adjunct.” |
| Trainees have to use their spare time and own space for webinars | • “Too many mandatory/highly useful webinars/sessions held outside of working hours.”  
• “There is no down time now and people are expected to be available for teaching/meetings whether at work, or off work with illness/zero day/annual leave.” |

Phase 2: Focus Group Interviews

Demographics
In total, 18 trainees (outwith the steering committee facilitators) participated in 8 semistructured interviews conducted in February 2021. The interviewees were trainees in general surgery (n=7), trauma and orthopedics (n=5), otolaryngology (n=1), plastic surgery (n=1), urology (n=1), and vascular surgery (n=1). The foundation trainees (n=2) did not declare their intended surgical specialty. The training grades of the interviewees were specialist registrars (n=8), core surgical trainees (n=6), and senior house officer–level staff grade (n=1; see Multimedia Appendix 4).

Key Findings
Interview responses generally supported survey findings with regards to strengths, limitations, drivers, and barriers. Recurrent themes of accessibility, timing, and technical aspects were explored in more detail.

Accessibility and Cost
Avoiding travelling to remote locations was reported as beneficial for time, financial, and environmental reasons. Accessibility to worldwide experts offered an opportunity to access international speakers and panels usually only available at major conferences, while avoiding the associated cost and study leave. On a regional level, webinars allowed deanery teaching programs to be combined to maintain high-quality speakers despite reduced speaker availability. There was considerable heterogeneity in individual experience, a minority of interviewees reported formal protected time and spaces for locally arranged webinars, this may not be practical for national or international events.

Timing of Webinars
Interviewee preferences for timing of webinars varied with their individual circumstances. Some identified midweek evenings as a time with less conflicting demands. Concerns were raised that a reliance on out-of-hours webinars disenfranchised those with family or other commitments and could contribute to burnout. One trainee highlighted the diversity of views,
explaining that as she lived away from her family during the working week, she preferred to watch webinars during weekday evenings to keep her weekends and working days free. Some expressed a preference for watching webinars at home, others found it difficult to focus with multiple distractions. Access to study leave for webinars during working hours was reported as limited.

**Recordings**

The ability to view recordings later was welcomed as adding flexibility, with the reservation that recordings lose interactivity, particularly the opportunity to ask questions. Repeat recordings may not remain up to date with current evidence and guidelines and may need to be reviewed. Archiving of content is variable and could be improved.

**Technical Aspects**

Many interviewees described variation in technical fluency between presenters. Live polls, chat, and question-and-answer functions were highlighted as improving interactivity. Equipment and bandwidth problems were common; in one example, an anatomy session from a dissecting room had insufficient resolution to identify structures. This may represent the learning curve of adopting new teaching methods and could be addressed through targeted training on the use of web-based platforms, and appropriate technical support. When available, a session chair can have a valuable role to set expectations and manage features (including muting and unmuting, screen sharing, and reviewing audience chat) to allow the presenter to focus on presenting.

**Future Opportunities**

To counter the loss of social interaction, a suggestion was made that a hybrid approach with remote access to face-to-face teaching offers a solution for deanery teaching to reduce the need to travel. Small groups meeting locally to access remote teaching would allow some element of networking and social interaction.

**Phase 3: Consensus Session**

**Demographics**

The consensus session was attended by 33 delegates, of whom 32 completed the demographics questionnaire. In total, 20 (62.5%) participants were female and 12 (37.5%) were male. A range of training grades and surgical specialties were represented among the consensus session cohort with specialist registrars (ST3-8; 43.8%) the most represented grade and general surgery (37.5%) the most represented specialty (see Multimedia Appendix 4).

**Voting Results**

Of the 25 Statements, 23 obtained at least 70% approval from the consensus meeting participants (Table 2) and are included in the recommendations of this paper.
Table 2. List of consensus statements and the percentage of consensus votes received.

<table>
<thead>
<tr>
<th>Consensus statement</th>
<th>Consensus rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Webinars have a role in surgical training but should not completely replace face-to-face training</td>
<td>100</td>
</tr>
<tr>
<td>Webinars are considered a good option for theoretical knowledge</td>
<td>94</td>
</tr>
<tr>
<td>Webinars are considered a good option for examination preparation (mock interviews or viva vocecs)</td>
<td>88</td>
</tr>
<tr>
<td>Webinars are considered a good option for training administration (eg, work-based assessments, journal clubs, and annual reviews of competency progression)</td>
<td>82</td>
</tr>
<tr>
<td>Webinars are currently considered a poor option for practical skills</td>
<td>88</td>
</tr>
<tr>
<td>Webinars are currently considered a poor option for simulation training (eg, advanced trauma life support)</td>
<td>88</td>
</tr>
<tr>
<td>Webinars are currently considered a poor option for communication skills</td>
<td>42</td>
</tr>
<tr>
<td>The following is considered best practice for delivering surgical training webinars: they should be delivered live (not prerecorded)</td>
<td>70</td>
</tr>
<tr>
<td>The following is considered best practice for delivering surgical training webinars: they should be recorded (available for playback)</td>
<td>94</td>
</tr>
<tr>
<td>The following is considered best practice for delivering surgical training webinars: they should be supported by an information technology specialist for troubleshooting and support</td>
<td>97</td>
</tr>
<tr>
<td>The following is considered best practice for delivering surgical training webinars: they should incorporate interactive elements, including a chat box, polls, or breakout room</td>
<td>97</td>
</tr>
<tr>
<td>The following is considered best practice for delivering surgical training webinars: they should be effectively archived and easily retrievable for future review</td>
<td>94</td>
</tr>
<tr>
<td>The following is considered best practice for delivering surgical training webinars: a certificate of attendance should be issued to allow trainees to log professional development</td>
<td>100</td>
</tr>
<tr>
<td>The following is considered best practice for the timing of surgical training webinars: they should not exceed one hour</td>
<td>73</td>
</tr>
<tr>
<td>The following is considered best practice for the timing of surgical training webinars: they should be eligible for study leave and be delivered within protected teaching time</td>
<td>100</td>
</tr>
<tr>
<td>Webinars delivered during evenings may increase trainee engagement by avoiding clashing with clinical commitments; however, this may disadvantage trainees who have families, long commutes, or other extracurricular commitments</td>
<td>97</td>
</tr>
<tr>
<td>Payment for webinars, outside of core content delivered as part of their training program, is acceptable to surgeons in training, but it should reflect the fair costs of hosting the webinar</td>
<td>48</td>
</tr>
<tr>
<td>Surgical trainers should be provided with resources and training to develop their digital teaching skills</td>
<td>100</td>
</tr>
<tr>
<td>Webinars offer opportunities to improve access and equality of training for trainees (eg, through delivery regionally, nationally, or internationally), and this should be explored further</td>
<td>97</td>
</tr>
<tr>
<td>The mechanism by which webinar attendance is recognized or accredited should be clarified</td>
<td>97</td>
</tr>
<tr>
<td>The value of structured webinars (ie, a series of webinars) aligned with surgical curricula is uncertain but should be explored further</td>
<td>97</td>
</tr>
<tr>
<td>While access to learning resources can be improved with webinars, the lost opportunities for networking, team building, or socializing and their implications are uncertain and should be explored further</td>
<td>100</td>
</tr>
<tr>
<td>Increased participation in web-based training out of normal working hours and how this contributes to trainee burnout should be explored further</td>
<td>100</td>
</tr>
<tr>
<td>Adjuncts to support web-based or remote practical skills training are evolving; therefore, the role or value of webinars for practical skills training should be revisited</td>
<td>97</td>
</tr>
<tr>
<td>A hybrid approach, using both face-to-face and web-based methods, may be the future of surgical training and should be explored further</td>
<td>100</td>
</tr>
</tbody>
</table>

**Discussion**

**Principal Findings**

This initiative has produced the first trainee-led, consensus-based recommendations on the use of webinars in surgical training with unanimous agreement that webinars have a role in future surgical training. Both practical and communication training was considered poorly suited, although it was acknowledged that this may change as methods of remote training evolve. These 23 consensus recommendations should support their continued and effective implementation and iteration.

A role for webinars in surgical training was anticipated, in line with the positive experiences that have been increasingly published across health care disciplines globally [17]. Blythe and Thompson [18] report the successful implementation of a
novel web-based surgical teaching program in Northern Ireland with support from the Royal College of Surgeons Edinburgh. The program was designed to meet the needs of core surgical trainees during the COVID-19 pandemic owing to cancellation to face-to-face teaching and used both videoconferencing software and web-based webinars. As a result of the positive feedback received, the course has been further developed into the primary surgical teaching method for the core surgical training cohort of 2020-2021 in Northern Ireland. This shows the sustainability and value of remote education going forward [18].

Webinars allow the remote delivery of expert teaching to a large number of trainees simultaneously, and recording these sessions offers further opportunity to increase the audience to people unable to attend live. This also provides trainees the option of revisiting the webinar to consolidate their learning or revise it at a later date. As observed during the development of our ASiT conference, accessibility and convenience are the commonly cited reasons for their use. The Virtual ACCESS conference (a core-trainee–led web-based conference to enhance surgical succession) surveyed delegates and found that the major factors that attracted attendees to their web-based conference were that the conference was free (91/130, 70%), allowed an opportunity to present (81/130, 62.3%), and did not require travel (78/130, 60%) [19].

Limitations

However, as highlighted in our process, webinars elsewhere also have their limitations. This seems to be largely around variation in technical delivery [20]. This was a large focus of the feedback received during our initiative and subsequently recognized within the recommendations: for example, around timing, duration, technical support, and archiving. An expectation of (unpaid) availability out of hours has implications for the costs of surgical training [21]. Although not specifically for webinars, technical challenges are a commonly cited barrier in telemedicine and are more frequently studied [20]. The preference for live delivery was also noteworthy, given that many web-based education initiatives use recorded videos [22,23]. As the role of guidelines is to improve experience by standardizing practice with the best available evidence [24], it is hoped that these recommendations can therefore have an immediate and positive impact.

However, there are some important limitations to acknowledge in our process. First, this process only captured the perspective of trainees. Therefore, it did not incorporate evidence from medical education, such as the comparative effectiveness of webinars, nor did it incorporate the perspective of surgical trainers specifically [25]. We also acknowledge a low response rate. The collaborative coauthorship model used and the outlined study design were chosen to incentivize participation. While these must be considered, it is worth recognizing that response rates parallel those of similar surveys [26-28] and perspectives across disciplines and training stages were still represented. Further, trainees are typically involved in education themselves and will have a trainer’s perspective.

Second, while most recommendations reached near unanimous consensus, there were inconsistencies. For example, while the initial survey suggested that trainees did not feel that communication skills were well suited to webinars, this did not become a consensus recommendation. This discordance may reflect an evolving perspective on webinars and communication as the initial concerns centered around being unable to recognize nonverbal communication in a web-based environment. Given that clinical practice has also transitioned to many web-based consultations [29], it is likely that remote teaching will have a more relatable role. Similarly, while practical training was felt to be better suited to alternative teaching methods, this was not unanimous, with agreement that, particularly as training tools develop [30], the role of webinars in practical training be investigated further. Of note, Fehervari et al [31] reported the first web-based, large-scale, surgical skills course that compared different web-based teaching methods and was validated against the gold standard of face-to-face teaching. The published feedback showed the success of this web-based course as delegates felt that the event met the required standards of a high-quality surgical teaching course and did not feel that face-to-face teaching would have been more appropriate despite the course focusing on practical skills [31]. This differs from the results of our consensus session and shows that there is scope for further research into the utility of webinars and web-based teaching for practical training.

The inconsistencies likely reflect the rapidly changing landscape of webinars in surgical education [20], and the need to iterate these recommendations as evidence and experience changes [29]. Based on the uncertainties identified during this process, a number of key knowledge gaps for clarification were identified, which too should help to guide their optimization.

Therefore, while we consider these recommendations to have been formed through an inclusive and structured process, and of immediate value, as with any form of recommendation, they should be considered a starting point that will require update over time.

Conclusions

This is the first initiative to produce consensus-based recommendations on the role and use of webinars within surgical training. These recommendations were produced by trainees and should serve as an initial framework for ensuring that they add value to surgical training in the future.

Acknowledgments

We would like to thank the following collaborators for their contribution to the initial web-based survey, the focus group interviews, and the consensus group session:

Bushra Abdelqader, Edward Armstrong, Abdul Badran, Victoria Bevan, Srushti Bhat, Carly Bisset, Rachael Boardley, Siobhan Chien, Cristina Croitoru, Rory CutlheBer, Elorm Daketsy, Radhika Dua, Afsana Elanko, Ricky Ellis, Charlotte El-Sayed, Callum

References

[1] Barlow et al. (2022). JMIR Med Educ 2022 | vol. 8 | iss. 4 | e40106 | p.72
https://mededu.jmir.org/2022/4/e40106

(Barlow et al. 2022)

Conflicts of Interest
The authors are all current surgical trainees.

Multimedia Appendix 1
Initial online survey created using Qualtrics.
[DOCX File, 100 KB - mededu_v8i4e40106_app1.docx]

Multimedia Appendix 2
List of preliminary anticipated statements prior to review and finalisation.
[DOCX File, 16 KB - mededu_v8i4e40106_app2.docx]

Multimedia Appendix 3
List of finalised statements presented during the consensus meeting.
[DOCX File, 16 KB - mededu_v8i4e40106_app3.docx]

Multimedia Appendix 4
Table of participant demographics for each stage of the study.
[DOCX File, 130 KB - mededu_v8i4e40106_app4.docx]

Multimedia Appendix 5
The Delivery of Webinars a) Graph showing the number of hours trainees have available to attend webinars each week b) Graph showing trainees’ preferred timings of webinars c) Graph showing trainees’ preferred duration of webinars d) Graph showing the preferred setting for webinars/virtual learning to be used in the post pandemic surgical training environment.
[DOCX File, 195 KB - mededu_v8i4e40106_app5.docx]

References
7. JCST Quality Indicators. Joint Committee on Surgical Training. URL: https://www.jcst.org/quality-assurance/quality-indicators/ [accessed 2021-04-11]


Abbreviations

ASiT: Association of Surgeons in Training
CHERRIES: Checklist for Reporting Results of Internet E-Surveys
COREQ: consolidated criteria for reporting qualitative research
Barlow E, Zahra W, Hornsby J, Wilkins A, Davies BM, Burke J
Defining a Role for Webinars in Surgical Training Beyond the COVID-19 Pandemic in the United Kingdom: Trainee Consensus Qualitative Study
JMIR Med Educ 2022;8(4):e40106
URL: https://mededu.jmir.org/2022/4/e40106
doi:10.2196/40106
PMID:36542431

©Emma Barlow, Wajiha Zahra, Jane Hornsby, Alex Wilkins, Benjamin M Davies, Joshua Burke. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.12.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Effects of Practicing With and Obtaining Crowdsourced Feedback From the Video-Based Communication Assessment App on Resident Physicians’ Adverse Event Communication Skills: Pre-post Trial

Andrew A White^1, MD; Ann M King^2, MA; Angelo E D’Addario^2, MA; Karen Berg Brigham^3, MPH, JD; Suzanne Dintzis^4, MD, PhD; Emily E Fay^5, MD; Thomas H Gallagher^1, MD; Kathleen M Mazor^6, EdD

^1Department of Medicine, University of Washington School of Medicine, Seattle, WA, United States
^2National Board of Medical Examiners, Philadelphia, PA, United States
^3Collaborative for Accountability and Improvement, University of Washington, Seattle, WA, United States
^4Department of Pathology, University of Washington School of Medicine, Seattle, WA, United States
^5Department of Obstetrics and Gynecology, University of Washington School of Medicine, Seattle, WA, United States
^6Meyers Primary Care Institute, University of Massachusetts Medical School, Worcester, MA, United States

Corresponding Author:
Andrew A White, MD
Department of Medicine
University of Washington School of Medicine
Box 356429
1959 Pacific St
Seattle, WA, 98195
United States
Phone: 1 2066161447
Fax: 1 206 221 8732
Email: andwhite@uw.edu

Abstract

Background: US residents require practice and feedback to meet Accreditation Council for Graduate Medical Education mandates and patient expectations for effective communication after harmful errors. Current instructional approaches rely heavily on lectures, rarely provide individualized feedback to residents about communication skills, and may not assure that residents acquire the skills desired by patients. The Video-based Communication Assessment (VCA) app is a novel tool for simulating communication scenarios for practice and obtaining crowdsourced assessments and feedback on physicians’ communication skills. We previously established that crowdsourced laypeople can reliably assess residents’ error disclosure skills with the VCA app. However, its efficacy for error disclosure training has not been tested.

Objective: We aimed to evaluate the efficacy of using VCA practice and feedback as a stand-alone intervention for the development of residents’ error disclosure skills.

Methods: We conducted a pre-post study in 2020 with pathology, obstetrics and gynecology, and internal medicine residents at an academic medical center in the United States. At baseline, residents each completed 2 specialty-specific VCA cases depicting medical errors. Audio responses were rated by at least 8 crowdsourced laypeople using 6 items on a 5-point scale. At 4 weeks, residents received numerical and written feedback derived from layperson ratings and then completed 2 additional cases. Residents were randomly assigned cases at baseline and after feedback assessments to avoid ordinal effects. Ratings were aggregated to create overall assessment scores for each resident at baseline and after feedback. Residents completed a survey of demographic characteristics. We used a 2×3 split-plot ANOVA to test the effects of time (pre-post) and specialty on communication ratings.

Results: In total, 48 residents completed 2 cases at time 1, received a feedback report at 4 weeks, and completed 2 more cases. The mean ratings of residents’ communication were higher at time 2 versus time 1 (3.75 vs 3.53; P<.001). Residents with prior error disclosure experience performed better at time 1 compared to those without such experience (ratings: mean 3.63 vs mean 3.46; P=.02). No differences in communication ratings based on specialty or years in training were detected. Residents’ communication was rated higher for angry cases versus sad cases (mean 3.69 vs mean 3.58; P=.01). Less than half of all residents...
Communication after medical harm typically falls short of patients’ and families’ needs [1]. Patients often feel abandoned, confused, and uncertain about how to obtain information and support [2]. Because many physicians are unprepared for these conversations [3,4], in 2017, the US Accreditation Council for Graduate Medical Education (ACGME) required all residents to receive training on how to disclose adverse events and participate in the “real or simulated” disclosure of harm events to patients and families [5]. Unfortunately, many US residency programs do not assure that graduates achieve competency in these skills. In a 2021 survey of over 11,000 US respondents, 28% of respondents reported receiving no training in error disclosure. Of those who did, only 9.2% underwent simulation training [6]. Instead, most received lecture-based training or informal training—techniques with critical limitations for developing communication skills that require practice and feedback.

Although some error disclosure curricula may improve knowledge and attitudes, published interventions lack rigorous assessments of skill acquisition [7]. Residents can learn through clinical practice, but high-stakes disclosures are infrequent, are problematic to observe, and are seldom accompanied by formative feedback from supervisors or harmed patients [8,9]. Lectures and multiple-choice questions do not assure skill acquisition [7], and simulations with standardized patients are logistically demanding, are expensive to implement at scale, and are lacking in statistical reliability [10-14]. It is also unknown whether standardized patients and faculty physician raters adequately approximate the viewpoints of patients who have experienced medical injuries. In particular, physicians’ viewpoints about ideal disclosure content and performance differ from those of patients, potentially limiting faculties’ ability to assess and provide coaching on performance [9,15]. To fully meet ACGME requirements and patient expectations, educators need tools for residents to practice error disclosure and receive formative feedback that is patient-centered, reliable, prompt, affordable, and practical.

The US National Board of Medical Examiners created the Video-based Communication Assessment (VCA) app to allow physicians and trainees to practice and receive specific individual feedback on their verbal communication [16,17]. The VCA app presents brief videos of case vignettes and prompts users to audio-record what they would say to the patient. Laypeople recruited on Amazon’s Mechanical Turk (MTurk) rate the responses as if they were the patient in the vignette [18]. MTurk has a very large, diverse population for crowdsourcing, along with extensive proof for being an inexpensive, rapid, and high-quality data source [19-21]. VCA users receive feedback reports with ratings, comparative data on their cohorts, learning points derived from raters’ comments, and sample audio recordings of highly rated peer responses. In a study of primary care communication scenarios, the VCA app provided high-quality, actionable feedback [22]. We adapted this tool for harmful medical error cases and found that crowdsourced laypeople provide reliable assessments of physician error disclosure and that adequately sized panels of crowdsourced laypeople can serve as reliable surrogates for panels of patients with past involvement in harmful errors [23]. However, the effects of VCA practice and feedback on residents’ error disclosure skills are unknown.

This paper describes a pre-post trial of using VCA practice and feedback to improve residents’ error disclosure skills. This approach involved using the VCA app as the vehicle for both training and assessment as residents sequentially practiced communication skills, received the results of the assessments and feedback to guide improvement, and then responded again to assess skill growth. We hypothesized that residents’ error disclosure skills, as rated by laypeople, would improve with personalized feedback. We also sought to evaluate 2 secondary research questions. First, we hypothesized that residents’ performance would vary with patient affect. Second, based on prior literature [24,25], we hypothesized that residents’ confidence with error disclosure would not correlate with laypersons’ ratings of their communication skills.

Methods

Overview

This pre-post efficacy study of the VCA app for error disclosure was conducted to determine whether laypersons’ ratings of resident error disclosure improve with practice and feedback. With input from experienced attending physicians, we made 12 cases, which included 4 cases specific to the following three fields: internal medicine, pathology, and obstetrics and gynecology (OBGYN). Each case consisted of 3 or 4 vignettes
depicting sequential stages in a conversation, such as initially sharing information about a mistake or responding to a patient’s emotional response (Multimedia Appendix 1). We recruited resident physicians at an academic center to complete each of the four cases (ie, 2 cases at 2 time points). Participating residents were randomly assigned to 1 of 2 groups within their specialty that differed only by the order of case assignment (Figure 1). Residents completed 2 cases at time 1; feedback was provided via the app after approximately 4 weeks, and residents were asked to complete the remaining two cases at time 2. The time points were spaced by 4 weeks, so that residents could be recruited while they were available during teaching conferences. We asked residents to not discuss the cases with colleagues until after study completion. We collected crowdsourced ratings of residents’ performances and surveys from residents.

Figure 1. The crossover study design for 48 residents who used the Video-based Communication Assessment app to practice error disclosure and receive feedback.

VCA Software and Content

The VCA app was described previously [16,23], and this project used the same software for presenting vignettes, recording residents’ responses, and delivering feedback. For each specialty, 2 cases portrayed an angry patient response, and another 2 cases portrayed a sad emotional response. Professional actors portrayed the patients. We designed cases to be of equivalent error severity. Further, 2 attending physicians and 3 senior residents pilot-tested the cases in their specialty and provided feedback via structured interviews to improve the cases’ relevance, quality, and believability.

Feedback reports, which were viewed in the VCA app, presented users with their overall communication scores for each vignette, the average scores of their peers, and written advice derived from the comments of raters who responded to the following question: “what would you want the provider to say if you were the patient in this situation?” Users could replay their own responses and listen to a highly rated response from a peer (Figure 2).

Figure 2. A screenshot from the Video-based Communication Assessment app displaying feedback for a single vignette within a case and the user controls for replaying the vignette video, replaying the user’s response, replaying a highly rated response, and expanding the field containing advice regarding what laypeople wish the physician would say.

Jorge Gonzalez (1 of 4) - Pathology
A 52 yo man undergoes surgery to remove a parotid gland nodule. Upon reviewing the slides, you rendered the diagnosis of pleomorphic adenoma, a common benign salivary gland neoplasm. You dictated a report, and the transcriptionist mistakenly enters the diagnosis of “pleomorphic adenocarcinoma.” You did not notice the transcription error and finalized the transcribed report with the incorrect diagnosis of pleomorphic adenocarcinoma (high grade, undifferentiated adenocarcinoma.). At his post-operative visit one week later, the surgeon discusses the unexpected diagnosis of high grade cancer with the patient and schedules appointments with an oncologist and radiation oncologist. He is discussed at head and neck tumor board. Upon review of the pathology at tumor board, the reporting error is discovered. The pathology report is amended to reflect the correct benign diagnosis. The surgeon informs the patient, he then asks you to speak to you to understand the error. He says:

Learning Points:
Be transparent; provide a truthful explanation of what occurred, take a...

https://mededu.jmir.org/2022/4/e40758

JMIR Med Educ 2022 | vol. 8 | iss. 4 | e40758 | p.78
(page number not for citation purposes)
Setting and Participants
We recruited residents from March to June 2020 at all training levels (postgraduate years 1-4) from 3 departments at the University of Washington. The departments represented a procedural specialty (OBGYN), a diagnostic specialty (pathology), and a nonprocedural specialty (internal medicine). In the United States, pathology and OB/GYN residency training typically lasts 4 years, and internal medicine residency training lasts 3 years. We invited all 237 OB/GYN (n=28), internal medicine (n=183), and pathology (n=26) residents. Recruitment for the pathology and OB/GYN specialties occurred at program-wide teaching conferences, with protected time for VCA use. The COVID-19 pandemic disrupted the in-person internal medicine conferences that were planned for recruitment; most internal medicine residents were recruited via email invitation. Each conference included a 10-minute orientation to VCA and 30 minutes for completing 2 cases. Residents received a US $50 gift card after completing all 4 VCA cases. At the time, the residencies did not have required program-wide training for error disclosure, and this trial did not provide didactic training.

Ethical Considerations
The University of Washington Institutional Review Board determined that this study was exempt from a review of resident, layperson, and patient advocate participants based on its policies, procedures, and guidance (case identifier: STUDY00008246) [26]. Risks and benefits were explained verbally; consent was implied by voluntary participation in the VCA exercise.

Outcomes Measured
Residents provided audio responses to each vignette through the VCA software. Audio responses were bundled into rating tasks for MTurk layperson raters who met the following criteria: US residents, those aged 18 years or older, and those able to speak and read English. Raters for OB/GYN vignettes were further restricted to women. To participate, raters completed an audio check, answered demographic questions, read a description of the vignette that was written in lay language, viewed the patient video, and listened to the first resident response. The raters then responded to 6 items that covered domains related to accountability, honesty, apology, empathy, caring, and overall response (Multimedia Appendix 2) before advancing to the next response. The items used a 5-point scale that was labeled with “Poor,” “Fair,” “Good,” “Very good,” and “Excellent.” The instrument was developed by the investigators because the core competency assessment tool developed by the US ACGME only measures residents’ disclosure of patient safety events as “participate,” “disclose,” or “models disclosure,” rather than assessing disclosure quality or the patient experience of disclosure [27]. Although a very limited number of tools exist for rating residents’ error disclosure to standardized patients [7], they include questions about body language or are intended for faculty raters; therefore, they were not appropriate for incorporation in the VCA app. After rating the last response in the set, the raters were prompted to enter free text in response to the following question: “What would you want the provider to say if you were the patient in this situation?” Crowdsourced raters received variable compensation amounts based on a rate of US $0.20 per rating.

Residents were surveyed before and after each video-based communication assessment with the questionnaires that were built into the VCA app. The initial survey asked about the residents’ sex, their level of training, and whether they had personally disclosed medical harm to a patient. Residents answered the following two items, using a 5-point scale ranging from “strongly agree” to “strongly disagree”: “I am confident in my ability to share information with a patient after medical harm” and “I am confident in my ability to respond to patient emotions after medical harm.” After assessment completion, residents were asked to rate the VCA app’s ease of use and the relevance of the cases. They were also asked to rate their performance in the domains of accountability, clear explanation, apology, empathy, caring, and overall response (Multimedia Appendix 2). The items were rated on a 5-point Likert scale that was labeled with “Poor,” “Fair,” “Good,” “Very good,” and “Excellent.” Before time 2, residents were asked if they had “incorporated the feedback into how [they] communicate with patients generally (not just communication about medical harm events).” Due to a technical error, this second survey was not shown to OB/GYN residents.

Analysis
We sought at least 8 raters per response after removing raters with indications of inattention or low contributions to reliability [28]. Ratings were aggregated across raters, items, and vignettes to create an overall score for each resident at time 1 and at time 2. Scores were created by averaging multiple responses to 6 Likert scale questions that were designed to assess general communication skills; thus, the continuous scores presented in this paper were derived from ordinal approximations of continuous variables [29,30]. The reliability and generalizability of the representative cases were analyzed and reported separately [23]. We used a 2x3 split-plot ANOVA to test the effects of time (pre-post) and specialty on communication ratings.

We used paired samples 2-tailed t tests to examine whether residents’ self-confidence in their ability to share information and respond to emotions increased from time 1 to time 2. To determine if self-confidence was related to actual ratings, data were subjected to multiple linear regression analyses wherein residents’ self-confidence in their abilities predicted pre- and postfeedback ratings. To determine if changes in such confidence from time 1 to time 2 were associated with changes in actual ratings, difference scores were created; one score represented the difference in residents’ ratings between rounds, and the other represented the difference in residents’ reported self-confidence in their abilities between rounds. Difference scores were subjected to a Pearson correlation analysis. To determine if self-reports of experiences with personally disclosing a harmful error to a patient (“yes” vs “no”) before time 1 were associated with time 1 performance, we used an independent samples 2-tailed t test. To determine if years in training were associated with ratings, we performed a Pearson correlation analysis. To determine if physicians’ specialty was associated with prior disclosure experience, we used a chi-square test.
To determine if communication ratings varied with the emotional affects of patients, we created 2 affect scores for each resident. One score represented a resident’s average rating across all vignettes for the two angry affect cases, and the other represented a resident’s average rating for the two sad affect cases. Scores were subjected to a paired samples 2-tailed t-test.

Results

Demographics and User Experience
Of the 238 residents from all specialty departments who were contacted to volunteer for this study, 62 (26%) completed the first two VCA cases (time 1) and received feedback (Table 1). Of these 62 residents, 48 (77%) completed the postfeedback cases (time 2). Less than half of all residents (27/62, 44%) reported prior experience with disclosing medical harm to patients; experience differed significantly among specialties ($P<.001$). Further, 1 of 17 (6%) pathology residents reported previously participating in such conversations, whereas 15 of 23 (65%) OB/GYN residents and 11 of 22 (50%) internal medicine residents reported prior experience with such conversations.

![Table 1](https://mededu.jmir.org/2022/4/e40758)

Table 1. Characteristics and study completion of the 62 resident physicians who participated in this pre-post study of crowdsourced Video-based Communication Assessment (VCA) app ratings.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Specialty, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obstetrics and</td>
</tr>
<tr>
<td></td>
<td>gynecology (n=23)</td>
</tr>
<tr>
<td>Sex (female)</td>
<td>19 (83)</td>
</tr>
<tr>
<td>Postgraduate year</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6 (26)</td>
</tr>
<tr>
<td>2</td>
<td>6 (26)</td>
</tr>
<tr>
<td>3</td>
<td>5 (22)</td>
</tr>
<tr>
<td>4</td>
<td>6 (26)</td>
</tr>
<tr>
<td>Completed the VCA at time 2</td>
<td>15 (65)</td>
</tr>
</tbody>
</table>

Prior to time 2, of the 48 returning residents, 23 completed a survey about incorporating the VCA feedback in general communications with patients. About half (11/23, 48%) agreed or strongly agreed that they had incorporated the feedback, 39.1% (9/23) neither agreed nor disagreed, and 13% (3/23) disagreed that they had incorporated the feedback.

Of the 62 residents, 38 (61%) completed surveys about the VCA app. A majority (32/38, 84%) found the cases to be “very much” or “completely” relevant to their practice. Additionally, 71% (27/38) found the VCA app to be “very much” or “completely” easy to navigate. We achieved a mean of 8.63 crowdsourced raters per case after removing raters with low contributions to reliability, resulting in an average cost of US $8.90 to rate the responses from 1 resident.

Changes in Resident Communication Ratings From Time 1 to Time 2
Based on the ANOVA for examining changes in mean resident communication ratings from time 1 to time 2, we found that residents were rated significantly higher at time 2 (mean 3.75, SD 0.16) than at time 1 (mean 3.53, SD 0.25; $P<.001$).

Self-confidence and Communication Ratings
Among the 30 residents who completed surveys at both time points, confidence in the ability to share information about medical harm increased from time 1 (mean 2.87, SD 0.73) to time 2 (mean 3.47, SD 0.63; $P<.001$). Residents’ confidence in their ability to respond to patients’ and families’ emotions after medical harm events also increased from time 1 (mean 3.20, SD 0.71) to time 2 (mean 3.60, SD 0.72; $P=.005$). The multiple linear regression analysis showed no relationship between confidence in such abilities and performance on either the prefeedback ratings or postfeedback ratings. Based on the difference scores evaluated with the Pearson correlation analysis, we found no relationship between increases in confidence over time and increases in ratings over time.

Self-reported Disclosure Experience, Specialty, and Years in Training
No differences in communication ratings based on specialty were detected. We found no significant relationship between residents’ years in training and overall communication ratings ($P=44$). However, residents who reported prior experience with disclosing medical harm to patients performed better at time 1 (mean 3.63, SD 0.23) compared to those without prior disclosure experience (mean 3.46, SD 0.25; $P=.02$).

Ratings by Patient Affect
Residents’ communication was rated significantly higher for angry cases (mean 3.69, SD 0.21) versus sad cases (mean 3.58, SD 0.21; $P=.01$).

Self-reported Performance in Error Disclosure Domains
After their first VCA use, residents’ mean self-rating of their overall response was 3.82 (SD 0.80). Residents reported their performance in sincerely expressing regret to patients (mean 4.05, SD 0.77), acknowledging and validating patients’ emotions (mean 4.00, SD 0.66), showing that they care about the patient (mean 3.79, SD 0.70), expressing accountability for the harmful event (mean 3.68, SD 0.70), and explaining things in a way that the patient could understand (mean 3.45, SD 0.80).
Discussion

Principal Findings

We measured laypeople’s assessment of residents’ error disclosure skills before and after they received numerical and written feedback. Residents’ mean ratings on a 5-point scale improved from 3.53 at baseline to 3.75 after feedback (P<.001). This finding provides novel evidence that simulated practice and feedback from laypeople can improve resident physicians’ error disclosure communication. The VCA app represents a novel, scalable, and statistically reliable tool for educators seeking to satisfy the ACGME mandate that residents “participate in the disclosure of patient safety events, real or simulated” [5]. Our findings indicate that the VCA app can be used across multiple specialties. This tool could particularly support educators and residents in diagnostic specialties, such as pathology, who may have fewer opportunities to participate in real-life error disclosure.

Consistent with prior literature about the accuracy of physician self-assessment [24,31], we found that residents’ confidence in their error disclosure skills did not correlate with laypeople’s ratings for these skills. Although physicians should still reflect on their performance, educators can emphasize the use of crowdsourced ratings as a more patient-centered and reliable way to assess error disclosure preparedness. This finding aligns with recommendations by disclosure experts that physicians should seek advice before discussing harmful events with patients [32,33]. Coaches often use brief practice and feedback to help clinicians recognize ineffective communication habits and phrasing before actual disclosures.

The higher ratings of residents when addressing patients with an angry affect versus patients with a sad affect warrant further study. This finding runs counter to predictions that residents might react defensively to confrontational, angry patients, which would be expected to result in lower communication ratings. One possible explanation for this finding is that angry patients’ challenging comments precipitated the explicit acknowledgement of their anger, whereas sad, withdrawn patients did not prompt a direct expression of acknowledgment or support. This would exacerbate existing physician tendencies to avoid discussing patient emotions, as indicated by observations that only 55% of attending surgeons who perform simulated error disclosure attempt to acknowledge or validate patients’ emotions [34]. A second hypothesis is that residents feel less shame when causing anger rather than sadness and respond readily with supportive expressions. Alternatively, the residents in this study may have been exposed to different instructions (ie, outside of this study) for handling patient emotions. Future research could pair an analysis of VCA response content with novel surveys about raters’ expectations for emotional support to characterize effective approaches for specific patient emotions. Error disclosure curricula should prepare trainees to tailor their communication approach to different patient and family emotional responses [35].

Strengths and Limitations

Our work has limitations. First, the VCA app assesses the skills needed for effective communication after harm but excludes other important areas, such as nonverbal communication. This limitation is offset by the strength that participants received actionable and focused feedback about their phrasing, which addresses a top concern of physicians and helps with not overwhelming them with advice across multiple domains [1,3]. Second, residents did not receive just-in-time coaching or a lecture on error disclosure—practices that might improve performance. Third, this study did not assess long-term skill maintenance. Fourth, the educational significance of the effect size is unknown, as we did not establish a threshold for either competence or mastery. Fifth, the limitations of the study population include recruitment at a single academic center and heterogeneity in the cohort’s training levels. A minority of eligible internal medicine residents participated (22/183, 12%), and not all participants completed all cases, which may have introduced selection bias. To address this, subsequent studies should be embedded in mandatory curricula rather than be based on volunteer participation. Sixth, we did not collect data that would allow us to explain why some residents reported that they did not incorporate the feedback into their communication practices or why some did not complete the cases at the second time point; these findings warrant further study. The efficacy and real-world implementation of this work remain unknown; we tested the VCA app as a stand-alone intervention instead of incorporating it into a longitudinal curriculum with a lecture. Lastly, we did not simultaneously measure faculty ratings of residents. Although this limited our ability to make comparisons between faculty and crowdsourced laypeople, we do not believe that this is a weakness. Rather, it highlights 2 key strengths of the VCA app. First, laypeople directly represent the ultimate arbiters of physician communication effectiveness—patients themselves. Second, crowdsourced laypeople can be recruited rapidly for statistically reliable sample sizes and at lower costs when compared to faculty.

Conclusion

The VCA app for error disclosure allows users to practice in a safe environment, provides formative feedback, and appears to facilitate skill acquisition. If these findings are replicated, the VCA app will likely offer a scalable way for residency leaders to meet ACGME mandates and assess residents’ skills. Yet, important questions remain about how best to incorporate the VCA app in graduate medical education. Multisite confirmatory studies should examine the effectiveness of using the VCA app in conjunction with didactic sessions on error disclosure and coaching from teaching faculty, as well as longitudinal skill development. Additional work could determine the efficacy of the VCA app for other challenging communication scenarios and for other learner groups, including practicing physicians. Finally, the viewpoints of residency leaders and residents about VCA acceptability, feasibility, and appropriateness will be needed to ensure adoption and sustainable use.
Acknowledgments
We thank Anders Chen, MD, and Kevin Blau, MD, for their assistance with study enrollment. This study was funded by the University of Washington’s Patient Safety Innovations Program. The funder had no role in the study design or manuscript review.

Data Availability
All data generated or analyzed during this study are included in this published article and its supplementary information files (Multimedia Appendices 3 and 4).

Authors' Contributions
All authors contributed to the study’s conception and design. AAW, KBB, AED, and AMK were responsible for the acquisition of data. AAW, AED, AMK, and KMM analyzed and interpreted the data. AAW, KBB, and AED drafted the manuscript. All authors critically revised the manuscript.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Text of the 12 cases used for the error disclosure video communication assessment.
[DOCX File, 32 KB - mededu_v8i3e40758_app1.docx ]

Multimedia Appendix 2
Survey instruments used in a trial of crowdsourced ratings and feedback about resident physicians' adverse event communication skills.
[DOCX File, 13 KB - mededu_v8i3e40758_app2.docx ]

Multimedia Appendix 3
Deidentified data of average ratings from crowdsourced laypeople for residents’ error disclosure by patient affect (angry vs sad).
[XLSX File (Microsoft Excel File), 10 KB - mededu_v8i3e40758_app3.xlsx ]

Multimedia Appendix 4
Deidentified data of survey responses from residents who used the Video-based Communication Assessment app for error disclosure practice.
[XLSX File (Microsoft Excel File), 13 KB - mededu_v8i3e40758_app4.xlsx ]

References


27. Internal medicine milestones. Accreditation Council for Graduate Medical Education. 2020. URL: https://www.acgme.org/globalassets/PDFs/Milestones/InternalMedicineMilestones.pdf [accessed 2022-09-12]


Abbreviations

ACGME: Accreditation Council for Graduate Medical Education
MTurk: Mechanical Turk
OBGYN: obstetrics and gynecology
VCA: Video-based Communication Assessment

©Andrew A White, Ann M King, Angelo E D’Addario, Karen Berg Brigham, Suzanne Dintzis, Emily E Fay, Thomas H Gallagher, Kathleen M Mazor. Originally published in JMIR Medical Education (https://mededu.jmir.org), 03.10.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
A Fork in the Road for Emergency Medicine and Critical Care Blogs and Podcasts: Cross-sectional Study

Michelle Lin¹, MD; Mina Phipps²; Yusuf Yilmaz³,4, PhD; Christopher J Nash⁵, MD, EdM; Michael A Gisondi⁶, MD; Teresa M Chan³,7, MD, MHPE

¹Department of Emergency Medicine, University of California San Francisco, San Francisco, CA, United States
²Department of Emergency Medicine, Stanford University, Stanford, CA, United States
³Education Research, Innovation, and Theory Program, McMaster University, Hamilton, ON, Canada
⁴Department of Medical Education, Faculty of Medicine, Ege University, Izmir, Turkey
⁵Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA, United States
⁶Precision Education and Assessment Research Lab, Department of Emergency Medicine, Stanford University, Stanford, CA, United States
⁷Division of Emergency Medicine, Department of Medicine, McMaster University, Hamilton, ON, Canada

Corresponding Author:
Michelle Lin, MD
Department of Emergency Medicine
University of California San Francisco
1001 Potrero Avenue, Suite 6A
UCSF Department of Emergency Medicine
San Francisco, CA, 94110
United States
Phone: 1 415 206 5758
Fax: 1 415 206 5818
Email: michelle.lin@ucsf.edu

Abstract

Background: Free open-access meducation (FOAM) refers to open-access, web-based learning resources in medicine. It includes all formats of digital products, including blogs and podcasts. The number of FOAM blog and podcast sites in emergency medicine and critical care increased dramatically from 2002 to 2013, and physicians began to rely on the availability of these resources. The current landscape of these FOAM sites is unknown.

Objective: This study aims to (1) estimate the current number of active, open-access blogs and podcasts in emergency medicine and critical care and (2) describe observed and anticipated trends in the FOAM movement using the Theory of Disruptive Innovation by Christensen as a theoretical framework.

Methods: The authors used multiple resources and sampling strategies to identify active, open-access blogs and podcasts between April 25, 2022, and May 8, 2022, and classified these websites as blogs, podcasts, or blogs+podcasts. For each category, they reported the following outcome measures using descriptive statistics: age, funding, affiliations, and team composition. Based on these findings, the authors projected trends in the number of active sites using a positivist paradigm and the Theory of Disruptive Innovation as a theoretical framework.

Results: The authors identified 109 emergency medicine and critical care websites, which comprised 45.9% (n=50) blogs, 22.9% (n=25) podcasts, and 31.2% (n=34) blogs+podcasts. Ages ranged from 0 to 18 years; 27.5% (n=30) sold products, 18.3% (n=20) used advertisements, 44.0% (n=48) had institutional funding, and 27.5% (n=30) had no affiliation or external funding sources. Team sizes ranged from 1 (n=26, 23.9%) to ≥5 (n=60, 55%) individuals.

Conclusions: There was a sharp decline in the number of emergency medicine and critical care blogs and podcasts in the last decade, dropping 40.4% since 2013. The initial growth of FOAM and its subsequent downturn align with principles in the Theory of Disruptive Innovation by Christensen. These findings have important implications for the field of medical education.

(JMIR Med Educ 2022;8(4):e39946) doi:10.2196/39946
KEYWORDS
open educational resource; free open-access meducation; FOAM; meducation; open-access; internet based; web based; website; social media; medical education; disruptive innovation; blog; podcast; emergency; critical care

Introduction

Many health care professionals abandoned their use of traditional medical textbooks over the last decade in favor of open-access, web-based, digital learning resources [1-5]. These resources and the trends to use them are collectively termed FOAM, for free open-access meducation [6,7]. We define FOAM as medical education resources produced in any format that are freely available online and across many different platforms without a required log-in. Examples include medical blogs, podcasts, subscription-free websites, and crowdsourced materials that constantly evolve. As the popularity of FOAM resources grew, so did the number and quality of these products [6]. In 2014, Cadogan et al [8] reported that 183 free blogs and podcasts were active in the fields of emergency medicine and critical care alone. This represented an approximately 90-fold increase in prevalence of these learning resources over the previous 10-year period.

The FOAM movement democratized and disrupted medical education by providing rapid, convenient, and open access to a breadth of clinical content on demand [9,10]. Globally, providers in practice and trainees both use FOAM for personal learning and in formal health professions curricula [11-13]. However, recent evidence shows that trainees seem to primarily use trusted blogs and podcasts when building their foundational medical knowledge [5,14]. Especially with the COVID-19 pandemic driving education toward more web-based content, the integration of FOAM in formal curricula has become more commonplace [4,15].

It is unclear if the current landscape of open-access blogs and podcasts can meet the demand and needs of modern learners [16]. The Theory of Disruptive Innovation by Christensen provides a conceptual framework to map the growth of these resources, anticipate roadblocks, and forecast new developments [17,18]. Christensen suggests that innovation by smaller entities (disruptors) in a market can challenge and may displace established organizations (incumbents). Incumbents may initially overlook disruptors because they initially seem trivial, underresourced, or too niche to succeed. Disruptive innovations typically start as either low-end or new-market footholds, catering to a low-cost idea or a new segment of an existing market to reach underserved customers. Over time, when the disruptor crosses a mainstream acceptability threshold, it begins to overtake all or part of the incumbent’s market share, and eventually profitability, because it is less expensive or more accessible.

In this framework, FOAM is a market-creating, disruptive type of innovation in medical education [19-21]. New-market innovations seek to fulfill a need in an unserved population—sometimes to provide goods or services offered by incumbent entities to those populations that could not afford or could not access those products. Although the new-market innovation targets a nonconsumption market (providing a good or service that was previously nonexistent to its audience), it still directly challenges the incumbent’s market space by potentially attracting away some of the incumbent’s customers with its new, affordable, or accessible option. The FOAM movement followed this new-market disruption model as free blogs and podcasts grew to establish a foothold in the market of knowledge translation and information dissemination. They were created in response to the incumbents (print textbooks, journals, live conferences, and professional societies) to aid health care providers and trainees who previously could not have easily accessed paywalled literature [12,22]. Prior to the FOAM movement, clinicians would access information by reading the original research from journal articles, attending conferences, or reading subscription-based references. Emergency medicine and critical care blogs and podcasts entered the market in 2002, and continued improvement cycles by bloggers and podcasters made these open-access resources more appealing and acceptable for learning and clinical practice. This was evidenced in 2002-2013 when there was an exponential growth in the number of blogs and podcasts in emergency medicine and critical care. Although FOAM would likely never entirely replace incumbent resources [7], it still attracted many end users and thus siphoned market share.

As medical education continues to evolve, one might anticipate that open-access blogs and podcasts would at some point supplant incumbent sources of content. However, many of the blogs and podcasts identified in 2014 by Cadogan et al [8] were small in scale and relied on the volunteerism efforts and financial resources of dedicated individuals [6]. Which of these blogs and podcasts remain active today is unknown, and that gap has far-reaching implications for medical education and knowledge translation. Therefore, this study aims to (1) estimate the current number of active, open-access blogs and podcasts in emergency medicine and critical care and (2) describe observed and anticipated trends in the FOAM movement using the Theory of Disruptive Innovation by Christensen as our theoretical framework [17,18].

Methods

Study Design

Using a positivist paradigm (a form of hypothesis testing through observation and measurements to build explanatory associations [23]) and Theory of Disruptive Innovation by Christensen as a theoretical framework [17,18], we conducted a cross-sectional study that identified active, open-access emergency medicine and critical care blogs and podcasts.

Data Collection

We identified active websites between April 25, 2022, and May 25, 2022, using a similar methodological approach as that described in 2014 by Cadogan et al [8]. More specifically, we created a roster of blogs and podcasts by combining a list created by the Life in the Fast Lane organization in 2019 with
Feedspot’s “80 Best Emergency Medicine Blogs and Websites” list in 2022 [24]. Subsequently, we identified additional sites through a purposeful snowball sampling technique, personal communications, social media, and a self-report form published as a blog post solicitation on the Academic Life in Emergency Medicine website on May 4, 2022. Lastly, we performed a Google search during May 1-8, 2022, to identify any overlooked sites, using a Boolean search strategy with the following terms: (“emergency medicine” OR “critical care” OR “intensive care”) AND (“podcast” OR “blog”).

Inclusion and Exclusion Criteria
We included emergency medicine and critical care websites if they published at least one post in the previous 6 months, were free and open-access, were composed in English, did not require a log-in or subscription, served health care professionals as the audience, and published content intermittently in dated posts. We then classified these websites into blogs (content was primarily text), podcasts (content was primarily audio, which may include text-based show notes), or blogs+podcasts. We excluded websites if they covered a broad range of specialties, of which emergency medicine or critical care comprised the minority of topics. Two study investigators (ML and MP) adjudicated uncertainties regarding website inclusion and classifications.

Outcome Measures
We collected the following data for each website: date of first published post, commercial ads used on the home page, other funding strategies (such as sales of books, web-based courses, or merchandise), institutional affiliation (sponsored by a professional society, journal, or external nonprofit or for-profit organization), and the personnel composition of the website (number of administrators, professional identity, and clinical practice setting affiliation).

Statistical Analysis
We used descriptive statistics to summarize our results. We tallied the total number of active emergency medicine and critical care websites and noted the number of blogs, podcasts, and blogs+podcasts. For each of these 3 categories, we report the median age of the website and each outcome measure.

Ethical Considerations
This study does not involve human participants. Therefore, we did not require institutional review board approval for this study because the information was publicly available.

Results
As of May 2022, we report 109 emergency medicine and critical care websites, comprised of 50 (45.9%) blogs, 25 (22.9%) podcasts, and 34 (31.2%) blogs+podcasts, as well as their characteristics (Table 1). The educational sources ranged in age from less than a year to nearly 18 years. Sponsorship of these sites varied, including advertisements (n=20, 18.3%) and institutional sources (n=48, 44%). Many sites (n=30, 27.5%) used nontraditional sources of revenue, such as the selling of merchandise, books, web-based courses, and premium podcast content. For some sites (n=30, 27.5%), a funding source was not immediately apparent. Team sizes ranged from single-person authors to large teams of at least 5 individuals, with most (n=60, 55%) comprised of at least 5 people. Physicians led most sites (n=97, 88.1%).

Table 1. Characteristics of active, open-access emergency medicine and critical care websites in 2022.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Blogs (n=50)</th>
<th>Podcasts (n=25)</th>
<th>Blogs+podcasts (n=34)</th>
<th>All sites (n=109)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years, median (range)</td>
<td>8.3 (0.8-17.7)</td>
<td>4.3 (1-10.6)</td>
<td>8.8 (4.6-15.3)</td>
<td>7.1 (0.8-17.7)</td>
</tr>
<tr>
<td>Sponsorship, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advertisements</td>
<td>8 (16.0)</td>
<td>2 (8.0)</td>
<td>10 (29.4)</td>
<td>20 (18.3)</td>
</tr>
<tr>
<td>Other funding strategies</td>
<td>13 (26.0)</td>
<td>3 (12.0)</td>
<td>14 (41.2)</td>
<td>30 (27.5)</td>
</tr>
<tr>
<td>Institutional</td>
<td>21 (42.0)</td>
<td>14 (56.0)</td>
<td>13 (38.2)</td>
<td>48 (44.0)</td>
</tr>
<tr>
<td>No ads, funding strategies, or institutional affiliation</td>
<td>16 (32.0)</td>
<td>8 (32.0)</td>
<td>6 (17.6)</td>
<td>30 (27.5)</td>
</tr>
<tr>
<td>Team composition</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of individuals, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11 (22.0)</td>
<td>7 (28.0)</td>
<td>8 (23.5)</td>
<td>26 (23.9)</td>
</tr>
<tr>
<td>2-4</td>
<td>8 (16)</td>
<td>10 (40)</td>
<td>4 (11.8)</td>
<td>22 (20.2)</td>
</tr>
<tr>
<td>≥5</td>
<td>30 (60.0)</td>
<td>8 (32)</td>
<td>22 (64.7)</td>
<td>60 (55.0)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1 (2.0)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>1 (0.9)</td>
</tr>
<tr>
<td>≥1 member in academia</td>
<td>34 (68.0)</td>
<td>20 (80.0)</td>
<td>24 (70.6)</td>
<td>78 (71.6)</td>
</tr>
<tr>
<td>Physician-led</td>
<td>45 (90.0)</td>
<td>21 (84.0)</td>
<td>30 (88.2)</td>
<td>97 (88.1)</td>
</tr>
</tbody>
</table>
Discussion

Principal Findings
The attrition rate of FOAM websites should alarm medical educators. We found that the total number of free blogs and podcasts in emergency medicine and critical care plummeted by 40.4% when compared to the 2014 report by Cadogan et al [8]. Figure 1 depicts the trends of FOAM blogs and podcasts over time, highlighting the deviation of the present-day landscape from extrapolations based on previous trends. This decline in FOAM is an unforeseen shift in the medical education marketplace that is particularly concerning given the vast number of learners who depend on the stable availability of these resources.

Figure 1. Number of active blogs+ (defined as blogs or blogs with podcasts) and podcasts during 2002-2013 compared to 2022. The dotted lines represent predicted trends if growth had continued at a consistent rate.

![Graph showing trends of FOAM blogs and podcasts](image)

The Christensen Theory of Disruptive Innovation helps us understand why fewer FOAM sites currently exist [18]. We posit that incumbents, such as professional societies and medical journals, launched their own educational blogs and podcasts, leveraging existing administrative staff and resources unavailable to the individual blogger or podcaster. This hybridization strategy by incumbents follows the disruptive innovation theory, which states that when the incumbent entity recognizes that a disruption is encroaching on its market space (and profitability), co-opting the disruptive innovation can retain market dominance and relevance [25]. Such incumbent activity may have lessened the need and urgency for new sites and thus reduced the scale of any continued market disruption by those smaller entities.

Financial Sustainability
Sustained success requires a foundation of infrastructure and resources. The disruptive innovation framework confirms this financial reality and explains why existing FOAM sites may be dissolving. The framework’s premise is that successful disruptors eventually become more financially solvent as they encroach on the incumbents’ market and customers. The FOAM movement, however, is grounded in a social good mission by volunteers providing free education to all who wish to learn. This volunteerism comes at the expense of opportunity costs and may have been unsustainable for many sites that no longer exist. Some sites likely had to increase their team size with more authors, editors, and administrators willing to donate their time and efforts. This is demonstrated by our finding that most sites now have at least 5 team members. Interestingly, fewer open-access FOAM sites may have resulted from previously free blogs or podcasts opting into a paid subscription business model to help maintain financial stability. Indeed, the increasing financial pressures on FOAM producers have caused some to shift their business model entirely to ensure the viability of their outlets. Our analysis found that a substantial number of FOAM sites revised their business plans. Of the 109 active sites, only about one-quarter (n=30, 27.5%) do not generate funds from advertisements, sales, or sponsorship. More specifically, 18% (n=20) embed ads; 27.5% (n=30) use a sales model that asks for donations or payment for merchandise, continuing medical education credits, books, or web-based courses; and 44% (n=48) are affiliated with a sponsoring institution, such as a professional society, hospital, or journal.

Academic Sustainability
Similar to financial opportunity costs, there are also academic opportunity costs. Traditionally, academicians are promoted and rewarded for their scholarly efforts based on grant funding, peer-reviewed publications, and national reputations. Digital scholarship in the form of blogs and podcasts has less academic value and is therefore less beneficial to career promotions and advancement [26,27]. This may play a role in the diminishing contributions of academicians, who play roles in 71.6% of existing FOAM sites, in favor of endeavors of higher academic
value. Academic benefit, however, does not explain the decline in FOAM sites created by educators not based in academia.

**Higher Expectations**

FOAM consumers have become accustomed to well-designed websites, rigorously produced content, up-to-date recommendations, and means for curricular integration of resources, and their expectations continue to rise. The volunteer-driven enthusiasm of many sites may have waned because of these production pressures placed upon FOAM producers over time. As another disincentive toward creating one’s own FOAM site, several existing sites allow guest authors to publish on their platform, providing educators with an outlet to disseminate their work with less investment of their time and resources. Consequently, the FOAM movement may be a victim of its own success—as the resources used most frequently by health care professionals and trainees matured, there was a natural imperative for some sites to grow or close. Our findings suggest a volatile educational marketplace with resources coming and going, threatening the reliance on and longitudinal use of these resources by stakeholders.

Taken altogether, this turnover in enduring FOAM sites is concerning for trainees and educators who currently rely on the stability of blogs and podcasts for planned learning and curriculum use. Looking forward, the reliance on the volunteerism and the financial investments of a few individuals seems not to be a sustainable model for the FOAM movement. It thus seems fated to follow one of two paths, based on our adapted version of the Christensen construct (Figure 2). One is to ultimately be co-opted by organizations with the financial means and infrastructure to support ongoing content production, such as journals, publishers, professional societies, academic institutions, and for-profit organizations (option 1). The FOAM movement will still technically exist but will instead report to a partisan parent organization. Examples include the following: *EM News*, which hosts a blog under its publisher Wolters Kluwer; the podcast *Annals of Emergency Medicine*, which highlights its own journal findings; the podcast *EM Cases*, which is sponsored by the not-for-profit organization Schwartz/Reisman Emergency Medicine Institute; the blog and podcast platform of *Core EM*, which is hosted by the NYU/Bellevue emergency medicine residency program; and the podcast *EM Over Easy*, which serves as the official podcast for the American College of Osteopathic Emergency Medicine. The other path is that FOAM sites dissolve over time (option 2). While some attrition or business remodeling has occurred, we still have not seen most blogs and podcasts choose a path one way or the other. We believe the next several years will be crucial as most sites will commit themselves to a particular path. One can envision a third theoretical path (option 3) resulting in a new-market model of an independent FOAM site with continued open-access content and alternative revenue streams to maintain financial stability.

**Figure 2.** Disruptive innovation model of free educational blogs and podcasts for knowledge translation, adapted from the Christensen Theory of Disruptive Innovation, illustrating its three potential futures for the FOAM (free open-access meducation) movement: joining existing incumbents as a sustaining innovation (option 1), disappearing (option 2), or remaining nonpartisan and sustainable as a new-market incumbent (option 3). The gray dashed line signifies the rising threshold of product quality by the audience before they will use it.
The Future of FOAM

At its core, the FOAM movement is a social-good, disruptive innovation that provides high-quality, open-access educational blogs and podcasts for health care professionals. The FOAM movement has made significant inroads into the medical education and digital scholarship landscape. For instance, some sites have incorporated learning management platforms [28,29], curricular lesson plans for trainees [30-32], virtual communities of practice [33-37], and digital object identifiers for blog posts or podcasts, making them eligible for Altmetric Attention Scores [38]. However, without more infrastructure and added financial and academic security, the FOAM movement seems destined for either assimilation into traditional incumbent institutions or disbandment.

We predict that if the FOAM sites are to achieve independence and sustainability (Figure 2, option 3), the solution will rest upon finding successful business models. The web-based Khan Academy may serve as such a model. As a nonprofit organization, it provides open-access educational videos and a learning management system for any educator or learner to use for free as an adjunct to traditional classroom teaching. It remains operational through philanthropic donations and sales of niche educational services. By contrast, however, the FOAM movement is a fluid, decentralized, virtual community. If FOAM similarly adopts a philanthropic means for sustainability, we anticipate that funders will likely gravitate toward a limited number of the highest-quality or the most impactful blogs and podcasts to start. This future would require a means to objectively measure the quality of these sites in an equitable and transparent manner, such as with the Social Media Index [39,40].

Acknowledgments

We wish to thank Mike Cadogan and the Life in the Fast Lane blog team for their generous sharing of their 2019 list of blogs and podcasts in emergency medicine and critical care.

Conflicts of Interest

ML is the founder and editor-in-chief of the educational blog Academic Life in Emergency Medicine (ALiEM). MP is funded by the IntroSems Plus program at Stanford University. MAG is the podcast host for the ALiEM EM Match Advice series. CJN is the Associate Director of Growth for ALiEM and the ALiEM University Chief Technology Officer. YY is funded by the Scientific and Technological Research Council of Turkey Postdoctoral Fellowship grant. TMC is the cofounder for the ALiEM Faculty Incubator and Chief Strategy Officer for the blog organization CanadiEM. Furthermore, she also reports honoraria from McMaster University for her education research work with the McMaster Education Research, Innovation, and Theory (MERIT) group and administrative stipend for her role of associate dean via the McMaster Faculty of Health Sciences Office of Continuing Professional Development. She has received various unrelated research grants, teaching honoraria, and speakership fees from academic institutions (Baylor University/Texas Children’s Hospital, Catholic University of Korea, Taiwan Veteran’s General Hospital, Prince of Songkla University, Harvard Medical School, International Association of Medical Sciences Educators, Ontario College of Family Physicians, Northern Ontario School of Medicine, University of British Columbia, University of Northern British Columbia, Holland Bloorview, nonprofit organizations (PSI Foundation), physician organizations (Association of American Medical Colleges, Canadian Association of Emergency Physicians, Society of Academic Emergency Medicine, the Royal College of Physicians and Surgeons of Canada, Medical Council of Canada), and governmental sources (Government of Ontario, Virtual Learning Strategy eCampus Ontario program).

References

https://mededu.jmir.org/2022/4/e39946 JMIR Med Educ 2022 | vol. 8 | iss. 4 | e39946 | p.90

(page number not for citation purposes)


29. ALiEM University: Be Free to Learn. ALiEM. URL: https://aliemu.com [accessed 2022-05-26]


Abbreviations

- ALiEM: Academic Life in Emergency Medicine
- FOAM: free open-access meducation

**Edited by N Zary; submitted 31.05.22; peer-reviewed by T Davis, T Tarchichi; comments to author 01.09.22; revised version received 02.09.22; accepted 12.10.22; published 28.10.22.**

**Please cite as:**
Lin M, Phipps M, Yilmaz Y, Nash CJ, Gisondi MA, Chan TM
A Fork in the Road for Emergency Medicine Blogs and Podcasts: Cross-sectional Study
JMIR Med Educ 2022;8(4):e39946
URL: https://mededu.jmir.org/2022/4/e39946
doi: 10.2196/39946
PMID: 36306167

https://mededu.jmir.org/2022/4/e39946 | JMIR Med Educ 2022 | vol. 8 | iss. 4 | e39946 | p.92 (page number not for citation purposes)
Software-Based Simulation on a 3D Environment for Vaccination Teaching and Learning: Design Science Research

Daniel Domingueti¹, MSc; Darlinton Barbosa Feres Carvalho¹, PhD; Diego Roberto Colombo Dias¹, PhD; Valéria Conceição Oliveira², PhD

¹Departamento de Ciência da Computação, Universidade Federal de São João del-Rei, Sao Joao del Rei, Brazil
²Campus Centro Oeste Dona Lindu, Universidade Federal de São João del-Rei, Divinopolis, Brazil

Corresponding Author:
Darlinton Barbosa Feres Carvalho, PhD
Departamento de Ciência da Computação
Universidade Federal de São João del-Rei
Campus Tancredo Neves
BR-494, s/n
Sao Joao del Rei, 36301-360
Brazil
Phone: 55 32 3379 4935
Email: darlinton@acm.org

Abstract

Background: Student training requires specific laboratories for vaccination practice, which are usually limited, and even professionals' continuing education regularly lacks proper care. Thus, new methodologies, concepts, and technologies, such as software-based simulations, are in highly demand.

Objective: This work aims to develop a 3D virtual environment to support teaching activities in the vaccination room. The software-based simulation must contribute positively to teaching considering a variable set of scenarios.

Methods: We applied the design science research method to guide the work. First, the concepts and opportunities were raised, which we used to build the simulation (ie, the proposed technological artifact). The development was assisted by a specialist, in which we sought to create a vaccination room according to Brazilian standards. The artifact evaluation was achieved in 2 stages: (1) an evaluation to validate the design with experts through the Delphi method; and (2) a field evaluation with nursing students to validate aspects of usability (System Usability Scale [SUS]) and technology acceptance and use (Unified Theory of Acceptance and Use of Technology version 2).

Results: We built the simulation software using the Unity game engine. An additional module was also developed to create simulation scenarios and view the students' performance reports. The design evaluation showed that the proposed solution is adequate. Students' evaluations confirm good usability (SUS score of 81.4), besides highlighting Performance Expectation as the most positively influential factor of Behavioral Intention. Effort Expectancy is positively affected by younger users. Both evaluation audiences cited the high relevance of the proposed artifact for teaching. Points for improvement are also reported.

Conclusions: The research accomplished its goal of creating a software-based simulation to support teaching scenarios in the vaccination room. The evaluations still reveal desirable improvements and user behavior toward this kind of technological artifact.

(KEYWORDS software simulation; vaccination room; immunization; teaching; training; evaluation; virtual world; Unity3D; SUS; UTAUT2)

Introduction

There are still few technologies to support teaching in the vaccination room. As it needs a specific laboratory, the education is typically related to managing the room and vaccine administration. Unfortunately, this is a common situation in Brazil. Despite having laboratories in (Brazilian) universities, learning usually happens in a professional environment with a trained nurse. Besides, in classrooms, the students may not be aware of the variety of situations that they will be encountering when interacting with a patient in the vaccination room.
In Brazil, as defined by the National Program of Immunizations (PNI), the team responsible for performing all the activities in the vaccination room comprises a nurse and nursing technicians or auxiliary nurses. PNI also lists the nurse’s tasks, such as preparing the vaccination room, performing the vaccination process, writing reports, providing the team’s continuing education process, and others. To sum up, the nurse is responsible for the overall supervision of the vaccination room [1].

Given the complexity observed in the tasks and the higher turnover rates of professionals, we need to highlight the importance of the team’s continuing education process. Although highly relevant, it is happening in scarcity and irregular ways. Nursing technicians or auxiliary nurses usually only receive training when they start their activities and when offered to senior professionals, mainly to update the vaccination schedule [2].

In this context, the proposal of new education and training methodologies based on integrating and participatory teaching and learning models is in high demand [3]. For instance, digital games and simulations allow the creation of virtual environments where players can learn about educational content. Games can generally motivate and engage students in learning activities [4]. More specifically, a simulation can benefit from the modern digital games tools and concepts to create more immersive educational experiences.

Although they share similar tools and ideas, digital games and simulations differ slightly. Games usually focus on the fun and competition among players, whereas simulations might focus on other objectives, such as those shared with an educational process. In a simulation, the essential elements from a process or event are replicated in a digital environment to promote an educational experience [5,6]. In this case, all elements essential to being represented in the simulation are considered: an application built to mirror—more accurately as possible—the object’s life cycle, process, or event [7].

Hoping to contribute to this sense, we evaluated the software development process for simulation to assist nursing students’ educational process in Brazil. Our hypothesis was that a 3D virtual environment, following standards and allowing the simulation of relevant scenarios, will be a viable alternative and positively contribute to teaching in the vaccination room. Therefore, our objective was to create a software-based simulation to assist nursing students’ educational process in Brazil. The software uses the concepts and tools of simulations and digital games applied to the vaccination room.

Methods

Study Design

We developed this research using the design science research (DSR) method. DSR is a method in which a designer answers relevant research questions through artifact creation. Essentially a problem-solving paradigm, DSR purposes innovative artifacts in which information systems become more efficient at solving relevant problems through a rigorous scientific process [8].

DSR starts by setting the basic requirements, problems, and opportunities. Later, the process follows the project research, in which the artifacts are designed, created, and their design further evaluated. The artifacts should be based on well-defined support theories or professional expertise. The process goes on with evaluations to assess the solution created. The researcher must communicate their results contributing to the scientific knowledge base at the end of the process [8,9].

The instantiation of the DSR methodology in this research is depicted in Figure 1, following the template proposed by Pimentel et al [10]. Accordingly, its definition follows the 2 main types of research (ie, project research and behavioral science research). The results confront the study’s central hypothesis through both evaluations (ie, design and field). We executed only 1 DSR cycle in this research, which started with project research and concluded with the field evaluation.
Project Research

Based on the premises presented in the “Introduction” section, the project research objective was to devise a solution, that is, a 3D environment software simulation, to support teaching activities in vaccination rooms. The design and development were performed by including a multidisciplinary team of software and vaccination experts, who cleared the essential requirements and evaluated the design iteratively.

Given the similarity of current simulations with digital games, research about tools used to specify and develop digital games formed the base of this work. Based on relevant and classical models available in the literature, the performed software specification follows the unified model from Hira et al’s work [11] and the educational elements based on Leite and Mendonça [12]. The inapplicable elements to this project (eg, business related) were removed from the model, while educational elements were added. Mock-ups and user interaction diagrams were used to define the virtual vaccination room and the other interface elements.

The simulation was implemented in the Unity game engine (Unity Technologies) due to its popularity and suitability for the intent. The multimedia resources (such as sounds, images, and 3D models) were (Creative Commons 0) licensed or created for the occasion. Furthermore, we developed a desktop system in which the instructor can create and manipulate instances to be executed further on the simulation. The desktop system was developed with the framework Avalonia UI.

The performed evaluation was defined based on previous research [13]. It comprised a design evaluation, still part of the project research, and a field evaluation as part of the behavioral research. Both evaluations used the same software version.

Regarding the design evaluation, a group of experts (with experience in vaccination or related areas, mostly university professors) was invited to compose an expert panel by the end of 2020. The experts evaluated the artifact through the Delphi method [14] to validate its content and general design. Only 1 round was performed in this research, bearing in mind that an expert has regularly assessed the artifact during its creation.

The group tested the 2 simulation scenarios and answered a questionnaire related to the following artifact attributes: objective, structure, presentation, and relevance. For each attribute, the experts classified a set of affirmatives to show their agreement level according to the following status: (1) inadequate, (2) partially adequate, (3) adequate, and (4) fully adequate. Besides, the experts could write down further considerations in a specific space for each attribute or say them aloud during the evaluation.

The consensus achieved by all the experts was measured according to the content validity index (CVI) [15] and the content validity ratio (CVR) [16]. We calculated both variables for each question, the mean of each category, and the mean of the total, considering all questionnaire items.

Behavioral Research

Overview

Behavioral science research seeks an assessment of the solution proposed regarding the hypothesis. The performed evaluation pursued an assessment regarding the software usability and acceptance by students.
Scenarios and Questionnaire Sections

We conducted the field evaluation with nursing students, who executed the software simulation running 2 scenarios. Later, the students answered a questionnaire composed of 4 sections: (1) demographic questions; (2) usability-related questions; (3) technology use and acceptance questions; and (4) open questions, in which the participants could express their opinion about the simulation.

The assessment considered 2 typical scenarios. The first scenario depicts the case of a 7-month-old child whose previous vaccines were not administered but are expected in the sixth month of life. The second describes a pregnant woman who needs the Tdap (tetanus, diphtheria, pertussis) vaccine. Both scenarios describe an ordinary real-life situation, demanding apprentices’ analysis of all patient conditions to identify and provide proper care.

As the evaluation objective was not to measure the participants’ knowledge, help was provided as requested. Notwithstanding, we asked the participants to send their performance reports for further analysis. We also considered annotations from the participants’ perspective about the experience in the analysis.

The contact with the participants was realized by convenience through email during the first semester of 2021. All the evaluations were executed individually or in groups of up to 3 participants.

Because of the COVID-19 pandemic, the participants interacted through a web conference. Thus, the participants also used their equipment to run the developed software. First, the participants were informed about the research, shown the simulation, and asked to share their computer screens during the experiment to assist when needed.

Usability Evaluation

The artifact usability was measured with the System Usability Scale (SUS) [17]. Data collection used a 10-item questionnaire, in which participants must define their concordance level according to a 5-point Likert scale. Following the method assessment calculation, the usability measurement of a given tool/artifact score ranges from 0 to 100. According to Bangor et al [18], the SUS score relates to adjectives, grades, and acceptance ranges.

Use and Technology Acceptance Evaluation

The Unified Theory of Acceptance and Use of Technology (UTAUT) and its further extension, Unified Theory of Acceptance and Use of Technology version 2 (UTAUT2) [19], are popular tools to analyze the use and acceptance of technology. Those methods were broadly extended and translated into many languages, including a fully adapted Brazilian version [20]. Nishi [20] translated the questionnaire and added a few more moderating variables to the UTAUT2 model. In this work, we used a more suitable version of this modified model (Figure 2).

We removed the constructs “Habit” and “Use Behavior” given the novelty of the artifact. Besides, the “Social Influence” construct was removed because the artifact does not have any social interactions between the participants (and we contacted the participants individually). We also removed the moderating variables “Experience,” “Schooling,” and “Marital status” due to the low variation observed (and expected) in the demographic data. Regarding Price Value (PV), participants considered the artifact under a free software license.

This way, it is possible to analyze the following hypothesis from Figure 2:

- H1 (+): Performance Expectancy (PE) affects the Behavioral Intention (BI) positively;
- H2 (+): Effort Expectancy (EE) affects the BI positively;
- H3 (+): Facilitating Conditions (FC) affects the BI positively;
- H4 (+): Hedonic Motivations (HM) affects the BI positively;
- H5 (+): PV affects the BI positively;
- H6a (+): Household Income (HI) acts as a positive moderating effect on FC;
- H6b (+): HI acts as a positive moderating effect on HM;
- H6c (+): HI acts as a positive moderating effect on PV;
- H7a (−): Sex acts as a negative moderating effect on HM;
- H7b (−): Sex acts as a negative moderating effect on HM;
- H7c (−): Sex acts as a negative moderating effect on EE;
- H8a (−): Age acts as a negative moderating effect on EE;
- H8b (−): Age acts as a negative moderating effect on PE.

We suppose the variable “Age” negatively affects the constructs EE and PE. We believe that young people can learn modern technologies more easily than older people. We also established that the variable “Sex” negatively affects the constructs EE, PV, and HM, knowing that most of the health and welfare course graduates in Brazil (73.8%) are females [21], and this way supports the assessment of its effect on the model. To conclude, we suppose a positive effect of the variable HI on the constructs FC, HM, and PV. This decision assumed that people with higher HI (and thus better social conditions) might be better acquainted with recent technologies. The selection of each hypothesis’s positive or negative effect is related to its interpretation, as displayed in Table 1.

The analytical model used in this research was evaluated using the partial least squares structural modeling equation method and analyzed through the methodology suggested by Benitez et al [22] and Hair et al [23]. We executed a bootstrap process with 100 samples and evaluated each hypothesis according to its relation effect, either positive (ie, supported) or negative (ie, unsupported); only P value <5% was considered significant. We developed scripts to automate the data processing, solve the model through the partial least squares path modeling estimation engine SemInR, and report results [24].
Figure 2. The Unified Theory of Acceptance and Use of Technology version 2 analytical model used in this research.

Table 1. Interpreting the effects of moderating variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Positive effect</th>
<th>Negative effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>Age</td>
<td>Older</td>
<td>Younger</td>
</tr>
<tr>
<td>Household income</td>
<td>Higher</td>
<td>Lower</td>
</tr>
</tbody>
</table>

Ethics Approval
The evaluation was approved by the Brazilian ethics committee Certificado de Apresentação de Apreciação Ética (CAAE 30545820.2.0000.5151).

Results

Overview of Outcomes From the Simulation
The proposed solution is a software simulation suitable for scenario-based training in a standard vaccination room modeled as a 3D environment. In this environment, a nurse avatar in a typical workday represents the user apprentice. A typical scenario assumes all required procedures before general servicing was performed, including preparing and opening the vaccination room.

Upon starting the simulation, apprentices must move their avatar toward the virtual patient in the room. Figure 3 illustrates the apprentice’s view when starting the simulation in a scenario with a mother and her child. The avatar moves through the room when pressing the keys W, A, S, and D (or the keyboard arrows), and mouse clicks provide interaction with some elements in the environment.

The tasks emulate the real-life process, except for checking the postvaccination adverse event (PVAE). While checking the PVAE, the nurse should observe whether the patient shows any immediate unexpected reaction. In the positive case, the nurse needs to react to them accordingly. We decided not to implement it because each event can be unique and vary for each patient, thus representing a higher complexity to replicate in the simulation.

Thus, the stages and respective tasks to be performed by the apprentice in the simulation are follows:

- Vaccine screening: analyze the patients’ health conditions, analyze their vaccination card, define the vaccines to be administrated, and register the vaccines in the information system.
- Hand hygiene: use the liquid soap dispenser, the paper towel, and the alcohol-based hand disinfectant dispenser.
- Preparation of the vaccines: select the vaccine administration route, select the needle size and dose, and remove the vaccine from the thermal box.
- Vaccination: apply the vaccine and dispose of the materials in the correct bin.
- Process finalization: set the return date and tell the patient to leave.

The simulation allows scenarios with the possibility of applying more than 1 vaccine. In this case, the apprentice repeats stages 3 and 4 until the administration of all vaccines.

The virtual room is divided into 3 sectors to consolidate the interaction style. The sectors gather a collection of related
elements in a similar context: screening, hand hygiene, and vaccine preparation. As the apprentice selects 1 of these sectors, the camera view changes to a fixed position from the sector. Thus, the apprentice can interact with all needed elements and execute tasks properly.

Multimedia Appendix 1 presents a comprehensive report with many screenshots and detailed descriptions of the proposed software simulation.

The thermal box contains most vaccines presented in the Brazilian 2020 schedule. Thus, in a standard scenario, the apprentice must select a vaccine from 18 options (Multimedia Appendix 1). The vaccines presented by default are available through typical (ie, actual/ordinary) scenarios. However, an instructor can set up new vaccines in custom scenarios, allowing simulation considering vaccines with unique characteristics.

The software actively records user actions through the simulation process, in which the user can export the respective report as an external file. Considering that an incorrect action prevents the simulation flow, recorded as a wrong choice, it is worth mentioning that all vaccine combinations are accepted as a valid input. As the correct administration varies according to each vaccine requirement, considering the patient’s age, muscle state, and even unique medical conditions, some scenarios might demand a detailed assessment of the report data.

The educational assessment is related to how well the apprentice executed the vaccination process, considering the number of incorrect interactions. Thus, when concluding the simulation, the performance report displays the following items:

- The last stage performed.
- List of vaccines administered.
- List of vaccines to be administered in the future as per the patient vaccines card.
- The number of incorrect selections or inputs when:
  - Defining a vaccine to be administrated.
  - Defining a return date.
  - Interacting with the information system.
  - Interacting in the hygiene sector.
  - Defining the administration route (total and specific by vaccine).
  - Defining the needle size (total and specific by vaccine).
  - Defining the dose (total and specific by vaccine).
  - Selecting a vaccine flask from the thermal box (total and specific by vaccine).

Figure 3. Starting view of the simulation environment (i.e., vaccination room).

Auxiliary Module for Instructors—Scenario Management

The auxiliary system allows the instructors to specify and manage parameters to create (or edit) a simulation scenario for their students. This system also aids in examining the performance report. The scenario variable options, organized into 3 categories, are presented in Textbox 1.

The system provides 7 avatars (ie, 3D character models) and 5 different vaccination cards to grant more realism and educational possibilities. Besides the current vaccination card, the system provides 4 other card models previously used in Brazil. Figures 4 and 5 show the current vaccination card: the real-life card and the implemented version in the simulation, respectively.
Textbox 1. Scenario variable options for simulation.

**Basic parameters**
- Patient’s name
- Patient’s birth date
- Patient’s 3D model
- Companion’s 3D model
- Scenario description
- Consultation date
- Opening dialog text

**Patient’s health conditions**
- Preexisting diseases and allergies
- Medication being used
- Reactions to previous vaccine administrations

**Vaccination history**
- Type of vaccination card
- Expected return date
- Vaccines administrated previously
- List of possible vaccines to be administrated by the apprentice
- Permission to use special fields in the vaccination card (the apprentice can set any vaccine in this field)
- Permission to apply a vaccine that is not listed in the current schedule

Figure 4. Current vaccination card (first page) used nowadays in Brazil.

![Current vaccination card](image-url)
Figure 5. Implementation of the current vaccination card in the simulation (first page).

Design Evaluation
A total of 9 experts participated in the evaluation. The age of 5 experts is between 30 and 39 years, 1 between 40 and 49 years, and 3 between 50 and 59 years. All experts have graduated in nursing; besides, one holds a master's degree and 8 a PhD. Table 2 presents their overall professional experience in nursing. Further, we noted 5 specialization areas and all experts work as professors in 4 Brazilian universities.

Table 2. Experts’ characterization according to their professional experience.

<table>
<thead>
<tr>
<th>Years since graduation</th>
<th>Number of experts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-9</td>
<td>2</td>
</tr>
<tr>
<td>10-19</td>
<td>2</td>
</tr>
<tr>
<td>20-29</td>
<td>4</td>
</tr>
<tr>
<td>30 or more</td>
<td>1</td>
</tr>
</tbody>
</table>

Some experts pointed out that the hand hygiene process was inadequate, mainly because the alcohol and liquid soap dispensers were displayed swapped according to new recommendations. Besides, some simulation elements displayed poor representation, such as the syringe and the flask used to represent the dose selection, and the avatars’ 3D models that were not fully matching their description.

One expert also mentioned that the error messages were confusing and occasionally could not recognize their cause. To conclude, many participants took a while to grasp the “Advance Stage” button—used to advance the vaccine definition task after the user selected all vaccines to be administered—and suggested changing its place on the screen.

Regarding their performance reports, Tables 4 and 5 summarize the results. Notably, the experts struggled more with the hand hygiene interaction, selecting a vaccine flask from the thermal box, and in the needle selection task. The return date was inserted incorrectly by 1 expert. Besides, the remarkable difference between the number of errors caught in both scenarios suggests that the experience acquired in the first scenario resulted in fewer interaction errors in the following scenario. Finally, 2 experts did not send their reports, while E6 and E7 performed only 1 scenario.
Table 3. Results of the Delphi questionnaire for the design evaluation.

<table>
<thead>
<tr>
<th>Attribute/item</th>
<th>Range</th>
<th>SD</th>
<th>Content validity ratio</th>
<th>Content validity index</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objectives</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. The information/content is consistent with the educational needs of the target audience (undergraduate students).</td>
<td>3-4</td>
<td>0.31</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2. The information/content is important for the quality of vaccination education.</td>
<td>3-4</td>
<td>0.31</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>3. It invites or instigates changes in the behavior and attitude of the students (future professionals).</td>
<td>2-4</td>
<td>0.67</td>
<td>0.78</td>
<td>89.88</td>
</tr>
<tr>
<td>4. It can be circulated in the scientific/educational environment of the nursing field.</td>
<td>3-4</td>
<td>0.42</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>5. It meets the objectives of undergraduate nursing courses.</td>
<td>4-4</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td><strong>Structure and presentation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. The material looks attractive.</td>
<td>3-4</td>
<td>0.50</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2. The content is adequate.</td>
<td>3-4</td>
<td>0.42</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>3. The information presented is scientifically correct.</td>
<td>2-4</td>
<td>0.67</td>
<td>0.78</td>
<td>88.89</td>
</tr>
<tr>
<td>4. There is a logical sequence of the proposed content.</td>
<td>3-4</td>
<td>0.48</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>5. The information is well structured in concordance and spelling.</td>
<td>3-4</td>
<td>0.31</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>6. The writing style corresponds to the level of knowledge of the target audience.</td>
<td>4-4</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>7. The illustrations are expressive enough.</td>
<td>3-4</td>
<td>0.47</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td><strong>Relevance</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. The themes reinforce aspects that should be reinforced.</td>
<td>4-4</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2. The material covers the subjects needed for vaccination knowledge.</td>
<td>4-4</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>3. It proposes the construction of knowledge.</td>
<td>4-4</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>4. The material allows the transfer and generalization of vaccination learning.</td>
<td>3-4</td>
<td>0.42</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>5. It is suitable for use in teaching vaccination.</td>
<td>3-4</td>
<td>0.31</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 4. Interaction errors (7-month-old child) present in the experts’ performance reports.

<table>
<thead>
<tr>
<th>Items</th>
<th>Participant</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
<th>E6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interacting in the hygiene sector</td>
<td>E1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Set a wrong vaccine in the card</td>
<td>E1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Interacting with the computer (ie, information system)</td>
<td>E1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>Select the administration route</td>
<td>E1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Select the needle</td>
<td>E1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>Select the dose</td>
<td>E1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Select a vaccine flask from the thermal box</td>
<td>E1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>17</td>
<td>1</td>
<td>1</td>
<td>22</td>
</tr>
<tr>
<td>Defining a return date</td>
<td>E1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

aE: expert.
Table 5. Interaction errors (pregnant woman scenario) present in the experts’ performance reports.

<table>
<thead>
<tr>
<th>Items</th>
<th>Participant</th>
<th>Participant</th>
<th>Participant</th>
<th>Participant</th>
<th>Participant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interacting in the hygiene sector</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Set a wrong vaccine in the card</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Interacting with the computer (ie, information system)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Select the administration route</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Select the needle</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Select the dose</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Select a vaccine flask from the thermal box</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Defining a return date</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

aE: expert.

Field Evaluation

The field evaluation was conducted with 20 participants (14 females and 6 males). All participants study in the Universidade Federal de São João del-Rei (undergraduate or graduate students); 7 are nurses. They are all less than 30 years of age, with the majority being single; however, 1 was married and 1 had a common-law marriage relation. The number of students with HI grouped by the Brazilian minimum salary (BMS; around US $200/month) was as follows: 5 with income below 2 BMS, 9 with income over 2 BMS and below 4 BMS, and 9 over 4 and below 10 BMS. Only 3 participants related a previous experience with virtual reality applications, games, or similar.

The overall SUS score obtained from the mean of all students’ scores was 81.4. Thus, the software simulation usability is acceptable, associated with the adjective good and grade B.

Table 6 presents the descriptive statistics of the UTAUT2 questionnaire: the mean (SD) and the range for each question.

Multimedia Appendix 2 presents the measurement model evaluation results, in which all observed values comply with the standard criteria. Therefore, this assessment grants internal data validity to the measurement model.

The discriminant was validated by the Fornell-Larcker criterion [25]. Establishing discriminant validity implies that a construct is truly distinct from others by empirical standards and captures phenomena not represented by other constructs in the model. The results, available in Multimedia Appendix 2, show a valid relation when the value in the main diagonal of the table is greater than any other in the same column. Thus, the discriminant validity was achieved by all relations between the constructs, except for FC and EE. Nevertheless, we established the model for this research according to the UTAUT2 analytical model.

According to the structural model evaluation results (Table 7), we can only state hypothesis H8a as supported. The other results from the structural model evaluation, showing the indirect effects and the constructs’ \( R^2 \) value, are provided in Multimedia Appendix 2.

Table 8 shows the results of the 15 submitted performance reports. The highest number of errors in the field simulation occurred when interacting in the hygiene sector, selecting a vaccine flask from the thermal box, and interacting with the computer (ie, simulated health information system).

Finally, the answers to open questions about the simulation’s positive and negative points and suggestions or critics about it and the whole experiment are summarized inTextbox 2.
Table 6. Descriptive statistics of the UTAUT2\(^a\) questionnaire.

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE1(^b): I find the tool useful in my daily life.</td>
<td>6.55 (0.60)</td>
<td>5-7</td>
</tr>
<tr>
<td>PE2: Using the tool increases my chances of achieving things that are important to me.</td>
<td>6.30 (1.03)</td>
<td>3-7</td>
</tr>
<tr>
<td>PE3: Using the tool helps me to accomplish things more quickly.</td>
<td>6.10 (0.79)</td>
<td>5-7</td>
</tr>
<tr>
<td>PE4: Using the tool increases my productivity.</td>
<td>6.20 (0.83)</td>
<td>5-7</td>
</tr>
<tr>
<td>EE1(^c): I find the tool easy to use.</td>
<td>6.20 (1.26)</td>
<td>3-7</td>
</tr>
<tr>
<td>EE2: It is easy for me to become skillful at using the tool.</td>
<td>6.50 (1)</td>
<td>3-7</td>
</tr>
<tr>
<td>EE3: My interaction with the tool is clear and understandable.</td>
<td>6.30 (0.92)</td>
<td>4-7</td>
</tr>
<tr>
<td>EE4: Learning how to use the tool is easy for me.</td>
<td>6.45 (0.83)</td>
<td>5-7</td>
</tr>
<tr>
<td>FC1(^d): The tool is compatible with other technologies I use.</td>
<td>6.25 (1.20)</td>
<td>4-7</td>
</tr>
<tr>
<td>FC2: I have the resources necessary to use the tool.</td>
<td>6.35 (0.99)</td>
<td>4-7</td>
</tr>
<tr>
<td>FC3: I can get help from others when I have difficulties using the tool.</td>
<td>6 (1.29)</td>
<td>2-7</td>
</tr>
<tr>
<td>FC4: I have the knowledge necessary to use the tool.</td>
<td>6.35 (0.93)</td>
<td>4-7</td>
</tr>
<tr>
<td>HM1(^e): Using the tool is fun.</td>
<td>6.60 (0.60)</td>
<td>5-7</td>
</tr>
<tr>
<td>HM2: Using the tool is enjoyable.</td>
<td>6.45 (0.68)</td>
<td>5-7</td>
</tr>
<tr>
<td>HM3: Using the tool is very entertaining.</td>
<td>6.25 (0.97)</td>
<td>4-7</td>
</tr>
<tr>
<td>PV1(^f): The tool is reasonably priced.</td>
<td>6.45 (1.10)</td>
<td>4-7</td>
</tr>
<tr>
<td>PV2: The tool is a good value for the money.</td>
<td>6.55 (0.89)</td>
<td>4-7</td>
</tr>
<tr>
<td>PV3: At the current price, the tool provides a good value.</td>
<td>6.40 (1.10)</td>
<td>4-7</td>
</tr>
<tr>
<td>BI1(^g): I intend to continue using the tool in the future.</td>
<td>6.35 (0.99)</td>
<td>4-7</td>
</tr>
<tr>
<td>BI2: I will always try to use the tool in my daily life.</td>
<td>5.75 (1.16)</td>
<td>3-7</td>
</tr>
<tr>
<td>BI3: I plan to continue to use the tool frequently.</td>
<td>5.65 (1.35)</td>
<td>3-7</td>
</tr>
</tbody>
</table>

\(^a\)UTAUT2: Unified Theory of Acceptance and Use of Technology version 2.

\(^b\)PE: Performance Expectancy.

\(^c\)EE: Effort Expectancy.

\(^d\)FC: Facilitating Conditions.

\(^e\)HM: Hedonic Motivations.

\(^f\)PV: Price Value.

\(^g\)BI: Behavioral Intention.
Table 7. Total direct effects and hypothesis validation.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Relation</th>
<th>Total direct effects</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>PE → BI</td>
<td>0.593</td>
<td>.58</td>
</tr>
<tr>
<td>H2</td>
<td>EE → BI</td>
<td>0.549</td>
<td>.74</td>
</tr>
<tr>
<td>H3</td>
<td>FC → BI</td>
<td>-0.356</td>
<td>.60</td>
</tr>
<tr>
<td>H4</td>
<td>HM → BI</td>
<td>-0.321</td>
<td>.42</td>
</tr>
<tr>
<td>H5</td>
<td>PV → BI</td>
<td>0.489</td>
<td>.85</td>
</tr>
<tr>
<td>H6a</td>
<td>HI → FC</td>
<td>-0.024</td>
<td>.91</td>
</tr>
<tr>
<td>H6b</td>
<td>HI → HM</td>
<td>-0.136</td>
<td>.54</td>
</tr>
<tr>
<td>H6c</td>
<td>HI → PV</td>
<td>0.100</td>
<td>.62</td>
</tr>
<tr>
<td>H7a</td>
<td>Sex → PV</td>
<td>0.297</td>
<td>.34</td>
</tr>
<tr>
<td>H7b</td>
<td>Sex → HM</td>
<td>0.273</td>
<td>.57</td>
</tr>
<tr>
<td>H7c</td>
<td>Sex → EE</td>
<td>0.320</td>
<td>.27</td>
</tr>
<tr>
<td>H8a</td>
<td>Age → EE</td>
<td>-0.385</td>
<td>.05</td>
</tr>
<tr>
<td>H8b</td>
<td>Age → PE</td>
<td>-0.322</td>
<td>.19</td>
</tr>
</tbody>
</table>

Table 8. Interaction errors caught in the field evaluation.

<table>
<thead>
<tr>
<th>Item</th>
<th>Total, n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interacting in the hygiene sector</td>
<td>24</td>
</tr>
<tr>
<td>Set a wrong vaccine in the card</td>
<td>0</td>
</tr>
<tr>
<td>Interacting with the computer</td>
<td>11</td>
</tr>
<tr>
<td>Select the administration route</td>
<td>0</td>
</tr>
<tr>
<td>Select the needle</td>
<td>5</td>
</tr>
<tr>
<td>Select the dose</td>
<td>6</td>
</tr>
<tr>
<td>Select a vaccine flask from the thermal box</td>
<td>19</td>
</tr>
<tr>
<td>Defining a return date</td>
<td>3</td>
</tr>
</tbody>
</table>

Textbox 2. Positive and negative points regarding the simulation.

Positive points
1. Ease of use.
2. It is fun (ludic element).
3. The tool has proximity to the reality and experience in the vaccination room, according to the safety rules of the Brazilian National Program of Immunizations.
4. It allows the user to visualize situations that minimize errors.
5. It allows the user to remember the vaccination process step-by-step—one participant emphasized the hand hygiene process.
6. Easy to learn how to use.

Negative points
1. The avatar and the camera system are challenging to handle.
2. The interaction with some aspects of the vaccination room is tricky—mainly with items from the hygiene sector.
3. The patient interaction in the scenario with a child, not the adult, negatively affects realism.
4. The patient does not move to the stretcher to receive the vaccine administration, which negatively affects realism.
5. Because of a lack of knowledge in the use of computers, learning how to use the simulation is challenging.
6. Equipment with high processing power is required because one male participant had difficulties in using the simulation with his equipment.
Discussion

Principal Findings

This paper presents a novel software-based simulation providing a 3D environment where an apprentice must complete the vaccination process according to a variable set of scenarios and Brazilian standards. An additional module was also developed to manage simulation scenarios and view performance reports. The creation considered the identified problem and opportunities along with its development, following the DSR method. The main results and in-depth discussion regarding the evaluations are presented in the following sections.

The proposed simulation was created to support teaching in the vaccination room, using innovative methods and technological resources. The design assessment concluded with activities regarding the project research, while the behavioral science research concluded with field evaluation. Thus, we conducted 2 evaluations to answer the research questions related to the project research (ie, design evaluation) and the behavioral research (ie, field evaluation).

Regarding the design evaluation, the experts considered the artifact approved according to the results of the design evaluation. The artifact modeling and its implementation are adequate (with possible minor improvements, such as a better 3D representation of some objects, inclusion of the PVAE’s procedure, and the information screen). The experts stated its high relevance to teaching and learning. That was seen not only in the “Relevance” attribute but also in the qualitative feedback.

In the field evaluation, nursing students assessed the artifact’s relevance. This evaluation aimed to evaluate the simulation’s usability and its use and acceptance by the students (and potential future users).

Usability was evaluated using the SUS. The results indicate a final score of 81.4 points, and an acceptable usability grade, which also can be described as good.

The students also indicated a remarkable acceptance level of the technology through the UTAUT2-based evaluation. PE was the factor that most influenced the students’ BI to continue using the technology. The simulation was easy to learn, and the SUS final score was reflected by the opinions of various participants (students and experts).

Both evaluations pointed at the ease of use and learning as positive aspects, although some participants struggled with the interaction system. The main difficulties reported were moving the avatar through the virtual environment, accessing the hand hygiene sector, and understanding the task list system. Remarkably, all participants completed the virtual vaccination process during the experiment.

Design Evaluation

The results of the “Objectives” attribute (Table 3) allow us to infer that the simulation meets its primary goal. The content and information displayed in the simulation meet the goals of nursing courses (CVR=1; SD 0). Moreover, the content and information displayed in the simulation are coherent with the target audience (CVR=1; SD 0.32). Still, except for item 5, not all experts present the same agreement level, given the SD variation identified and the qualitative results.

Specifically, item 3 achieved the minimum acceptable value according to the CVR criterion. The reason may be related to the hand hygiene process, with the dispensers’ position switched, as pointed out by some experts. They also noted an analogous situation in the “the information presented is scientifically correct” item from the “Structure and Presentation” section.

One expert said that the hand hygiene process was inaccurate regarding the most recent recommendation. A nurse may prefer washing hands with liquid soap rather than alcohol gel, but the 2 products should not be used simultaneously according to the most recent standard recommendation. In addition, the expert proposed the removal of the sink from the simulation, allowing only 1 way of washing hands.

However, it is worth mentioning that using both products together is not a bad practice in all cases. Besides, there are circumstances where a nurse cannot clean their hands with either soap or alcohol. To sum up, the professional must judge the situation and choose the best alternative. Thus, we dropped the suggestion of removing the sink and changing the hand hygiene sector.

We decided not to implement the PVAE due to its idiosyncrasies. Nevertheless, it is a required step. Many experts suggested adding at least a reminder at the end of the process as an educational feature.

The “Structure and Presentation” assessment shows that the writing style corresponds to the target audience’s knowledge. No problems related to grammar and spelling were noticeable in the simulation, and it uses the specific technical terms appropriately. The experts approved the simulation’s visual quality, and the variation between the agreement levels (adequate and fully adequate) is related to the qualitative feedback.

Because of the experts’ possible lack of experience in computer applications (such as digital games or simulations) and the adopted interaction style, we expected hurdles concerning recognition and interaction within the 3D simulation environment. The performed assessment reflects it through several items. It is worth mentioning that the interaction style applied in the proposed software simulation is a well-consolidated pattern in the software industry, which is present in many modern digital games.

Some experts struggled to move the avatar, which also handicapped their interaction with particular objects in the environment. When the interaction context changes—from the avatar view to a given sector view—selecting items from that sector can be impossible due to an overlay of the avatar mesh with the sector objects. The most prominent errors observed during the experiments in the hand hygiene sector (Tables 4 and 5) are related to this interaction issue.

Several experts tried to set the patient’s return date on the computer instead of initially interacting with the virtual patient. According to the real-life procedure, a nurse confirms the returning date calculated by the standard information system,
as long as the patient updates the vaccination schedule. Otherwise, the nurse must input the new return date directly into the system based on the patient’s vaccination card.

We chose to set the return date at the end of the simulation through interaction with the patient. The reason was the high complexity involved in replicating the automatic process executed by the standard information system in the simulation. However, it can indicate a simulation mistake because the current metaphor drifts apart from real life. It is also possible to insert the return date directly into the patient’s vaccination card. Still, it was not a required step, and only 1 participant used this option during the field and design evaluations.

Some experts tried to register the vaccine in the system more than once. That happened because of an initial misunderstanding about stage transition and the overall task list comprehension. Despite minor interaction problems observed and few experts’ remarks to add a higher degree of realism and trustworthiness to the simulation, the whole artifact was approved. The final mean of both CVI and CVR was above the expected value.

**Field Evaluation**

The usability assessment obtained from the SUS score substantiates the participants’ opinions regarding ease of use. Among other positive points mentioned by the students, the ludic learning factor and artifact relevance were also highlighted.

Furthermore, according to the participants, “the tool presents similarity with the PNI safe practices,” whereas the professionals noted that “use matches the reality lived in the vaccination room.” Thus, it is an adequate tool to support teaching and learning experiences. The user “has the vision of how it is like to be in a vaccination room,” it allows to “practice before interacting with a real patient,” and thus the simulations allow to “visualize situations that diminish missteps.”

Similar to the situation with some experts, some students also struggled to handle the avatar movement and interaction with other elements. The interaction style defined for the movement is standard in the digital games industry. Nevertheless, it is not a natural interaction style for all. Unexperienced users need directions and time to properly learn how to handle the avatar.

The errors related to computer interaction can be explained by the lack of comprehension regarding the changing stages through the simulation. As observed in the design evaluation, the students also tried to register the vaccine on the computer more than once. One student attempted to write the return date as well. The reasons are the same as those discussed in the “Design Evaluation” section.

Students also clicked more than once in vaccine flasks from the thermal box, and a few just unthinkingly clicked on random vaccines. The interaction with the thermal box was perceived as simple to be achieved and understood. Yet, the lack of attention or experience with similar software increases the number of errors.

Whereas the previous mistakes are understandable and tamed mainly by experience, those in the hand hygiene sector indicate a significant issue. Only 4 students from the 15 who sent the report did not record missteps in this area. This sector was also extensively discussed in the experts’ evaluation.

Regarding the UTAUT2 evaluation, we can only generally support hypothesis H8a due to the statistical significance (Table 7). The conclusion is that being younger has a positive influence on the construct EE.

The UTAUT2 results are complementary and supported by the SUS score and qualitative feedback. Although the UTAUT2 method is unsuitable for a small number of participants because it may lack statistical reliability, it is valuable as a theoretical framework to guide technology assessment and provide relevant results.

Fully achieving statistical significance was not our goal nor the expected result, bearing in mind the limited number of participants we could gather. Thus, despite its limitations, the results are still valid and represent the participants’ perceptions well. Consequently, UTAUT2 is used as a theoretical framework to understand the simulation through the participants’ view and not as a final and general assessment.

Therefore, we noted that the PE positively affects the BI of the students in continuing to use the simulation (H1). In this case, the users feel compelled to continue using the simulation because they perceive its benefits. Regarding the scenario performance, all students completed the vaccination process and pointed out most factors as positive regarding their experience. Their performance is also reflected in Table 8, in which their number of missteps is remarkably smaller than that of the experts’ (Tables 4 and 5).

EE also contributes positively to BI (H2), according to the participants. Thus, the simulation’s challenge is adequate for the audience. Although few experts had the impression of a lack of logical sequence for the simulation’s tasks, the arrangement was perceived as positive from the students’ perspective. Users need to comprehend the environment and the task to accomplish its purpose by moving the avatar and interacting with the suitable element. This process presents a reasonable challenge and resembles the real-life process. As 1 male participant noted in his qualitative feedback, the simulation tool is close to the reality and experience in vaccination rooms.

By contrast, the influence of FC diminishes the user’s BI (H3). Few students complained about the lack of knowledge, experience, and training regarding computer systems, besides a steep learning curve associated with the simulation interaction procedure. Few experts also noted the lack of clear instructions in the simulation, which can be a problem. However, the learning needs seem not to be an issue because the struggling participants could use the simulation with only few initial instructions. Moreover, the FC construct assessment achieved a high score, and it is worth noting the overall high SUS score and the high EE and PE values.

The HM also does not confer positive effects in students’ BI (H4). The artifact was built according to simulation concepts but not as a game, and some students also emphasized it. However, we expected a positive influence considering the target audience (young adults and students aged between 19 and 27 years) and the similarity of the artifact with a digital
game. Nonetheless, some students mentioned that the simulation was fun as a positive aspect.

The PV contributes positively to the BI (H5). The participants considered the artifact available under a free software license. Thus, the artifact was recognized as having a good cost-benefit because it is free and benefits the participants.

Regarding the influence of sex, hypotheses H7a, H7b, and H7c positively affected the constructs PV, HM, EE, and BI. Accordingly, female participants are less influenced by the characteristics of price, HM, and effort expectation. Furthermore, sex directly affects the students’ BI of the continued use of the artifact, with men being more favorable than women.

Regarding the variable age, we can note that it has adverse effects on EE (H8a) and PE (H8b) and on BI according to our sample. This finding indicates that younger users have better effort and performance expectations regarding simulation use. It matches our observation regarding the experiment and the discrepancy in the number of errors from experts and students.

According to our sample, the $R^2$ value indicates that the variable age explains 10.3% of the PE variance. In association, 19.4% of the EE variance is explained through age and sex. Besides, hypothesis H8a was supported and achieved statistical significance in our model. Therefore, younger users have higher BI in using the simulation than older ones.

To conclude the structural model analysis, all other constructs predict the users’ BI on the continued use of the artifact in 73.3% of the cases, according to the $R^2$ value in our sample. Moreover, the analysis shows the influence of variable age with statistical significance favorable toward young users on the EE. PE, EE, and PV are the influential positive factors of the BI in the structural model built. Although the PV (ie, artifact considered free software) affects BI, HI is neglectable.

The B11, B12, and B13 items from Table 6 assess the participants’ BI, reaching an overall average of 5.92. Despite the smaller score compared with other constructs, the general intention is optimistic. Although some participants are neutral, most declare the intention to continue using the simulation (B11). Most students will also try to use the tool as much as possible (B12). The same argument can be extended to item B13, in which the students intend to use the simulation frequently. The low SD compared with the average indicates a high agreement level.

Although there is still considerable debate regarding the potential usefulness of serious games, previous research shows that such digital approaches appear to be at least as effective as controls and, in many studies, more effective for improving knowledge, skills, and satisfaction [26]. Further rigorous and theory-driven research is required and could promote better understanding, leading to enhanced design processes and outcomes.

Both evaluation audiences cited the high relevance of the proposed simulation for teaching. Points for improvement were also reported.

**Strengths and Limitations**

The lack of end users in the design process is a notable limitation concerning the collaborative co-design team, but the evaluations revealed their perceptions and points for improvement. The evaluations were conducted entirely online, using participants’ computers. Students required special assistance and sometimes demanded brief interventions during the experiment. Before the evaluation stage, a brief introduction about the interaction with the simulation may have added a bias to the participants’ answers. Besides, participants had different assumptions about simulation performance because each used a different computer.

Regarding the UTAUT2 evaluation, the sample size is also seen as a limitation of this study. Although PLS-PM has many advantages over other methods, being more reliable when the sample size is small [23], its performed analysis is limited. According to Hair et al [27], each construct demands at least five participants, while 15-20 will be ideal, yet without convergence assurance. Notwithstanding, the method and the sample allowed a limited but still solid analysis.

Nonetheless, the model assessment granted internal validity to all criteria and mostly achieved the discriminant validity. The discriminant validity was not established between the constructs FC and EE, but an analogous situation was reported in the UTAUT2 model proposition [19]. Further, the model has not achieved statistical significance in almost all the hypotheses. However, the model was kept as theoretically proposed because the results obtained from the other evaluations complement and support the results found in the UTAUT2 analysis. It is still important to mention that the results were validated by a statistic professional who also assisted in the initial analysis of the results to ensure higher reliability for the study. Moreover, results obtained from the other evaluations complement and support the results found in the UTAUT2 analysis.

**Conclusions**

The research accomplished its goal of creating a software-based simulation to support teaching scenarios in the vaccination room. The evaluation results showed that the proposed simulation is adequate, with good usability and student acceptance.

The design evaluation indicates that the artifact allows transferring, sharing, and generalizing the knowledge. Therefore, the created simulation is suitable to be used in vaccination education.

Given all the assessed elements that influence the users’ BI and the qualitative feedback provided, students approved the artifact: the simulation presents good usability, and its users accept it well.

From the participants’ point of view, the simulation had a more significant focus on the educational experience. At the same time, HM was seen as a secondary element. This result meets our theoretical foundations in simulations and digital games. However, a positive influence was expected given the target audience of this evaluation: young adults. Both HM and FC are detractor factors to the BI.

We propose as future work adding different vaccination schedules to be selected by the instructor and exploring the
simulation with other tasks, such as opening the vaccination room and those related to the vaccine conservation. Besides, it is necessary to validate the UTAUT2 with more participants to validate the entire model used in this study.

Acknowledgments
This study was financed in part by FAPEMIG.

Data Availability
The data sets generated during or analyzed during this study are not publicly available due to compliance with ethical standards on research with humans but are available from the corresponding author on reasonable request.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Details of the proposed software simulation.

[PDF File (Adobe PDF File), 657 KB - mededu_v8i4e35712_app1.pdf ]

Multimedia Appendix 2
Details of the use and technology acceptance evaluation.

[PDF File (Adobe PDF File), 203 KB - mededu_v8i4e35712_app2.pdf ]

References


Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BI</td>
<td>Behavioral Intention</td>
</tr>
<tr>
<td>BMS</td>
<td>Brazilian minimum salary</td>
</tr>
<tr>
<td>CVI</td>
<td>content validity index</td>
</tr>
<tr>
<td>CVR</td>
<td>content validity ratio</td>
</tr>
<tr>
<td>DSR</td>
<td>design science research</td>
</tr>
<tr>
<td>EE</td>
<td>Effort Expectancy</td>
</tr>
<tr>
<td>FC</td>
<td>Facilitating Conditions</td>
</tr>
<tr>
<td>HI</td>
<td>Household Income</td>
</tr>
<tr>
<td>HM</td>
<td>Hedonic Motivations</td>
</tr>
<tr>
<td>PE</td>
<td>Performance Expectancy</td>
</tr>
<tr>
<td>PNI</td>
<td>National Program of Immunizations (in Brazil)</td>
</tr>
<tr>
<td>PV</td>
<td>Price Value</td>
</tr>
<tr>
<td>PVAE</td>
<td>postvaccination adverse event</td>
</tr>
<tr>
<td>SUS</td>
<td>System Usability Scale</td>
</tr>
<tr>
<td>UTAUT2</td>
<td>Unified Theory of Acceptance and Use of Technology version 2</td>
</tr>
</tbody>
</table>
Emojis can improve health communication, especially when incorporating emojis into traditionally word-only texts. Beyond improving communication, emojis also offer greater access to health care, especially for vulnerable and marginalized populations with limited health literacy. A recent study found that 94% of patients with limited health literacy preferred health reports with emojis. Moreover, health officials are considering adding emojis to cardiopulmonary resuscitation (CPR) guidelines as well as public health guidelines for handwashing. As the world evolves with new technology and new methods of communication, we must also evolve the language and method we use to communicate health information to patients. In this viewpoint, we aim to discuss the methods health care professionals can use to develop novel communication methods using emojis and the benefits of their incorporation into health care communication.

A major obstacle for increasing the use of emojis in health care is the lack of diversity in health care–related emojis. Of the 3521 emojis in the Unicode Standard, which is the organization that maintains text and communication standards across electronic devices, only 30 of them were related to health care. These mainly were generic body parts (eg, ears, hands, or legs) or general symbols such as “pill” or “syringe.” There is a large deficiency of emojis that communicate detailed aspects of health care, such as CPR, drawing blood, and getting an injection. As such, we believe that the current generation of upcoming medical students have the potential to fill this gap with new emojis.

As members of a new generation that sends over 10 billion emojis a day, we have the greatest experience in knowing how to balance the artistic features of an emoji with the necessary detail to convey the information correctly. Researchers have proposed the development of new health-related emojis such as a liver emoji and kidney emoji. However, tech-savvy medical students have the ability to increase the use and accessibility of emojis by just increasing the use of emojis in

---

**Abstract**

Emojis can improve health communication, especially when incorporating emojis into traditionally word-only texts. Beyond improving communication, emojis also offer greater access to health care, especially for vulnerable and marginalized populations with limited health literacy. A recent study found that 94% of patients with limited health literacy preferred health reports with emojis. Moreover, health officials are considering adding emojis to cardiopulmonary resuscitation (CPR) guidelines as well as public health guidelines for handwashing. As the world evolves with new technology and new methods of communication, we must also evolve the language and method we use to communicate health information to patients. In this viewpoint, we aim to discuss the methods health care professionals can use to develop novel communication methods using emojis and the benefits of their incorporation into health care communication.

(JMIR Med Educ 2022;8(4):e39059) doi:10.2196/39059

**KEYWORDS**

emoji; medical education; technology, education; medical students; creativity; student; health awareness; health; awareness; medical; society; innovation; communication; medical communication; electronic; artistic; representation

---

1Department of Chemistry and Chemical Biology, Harvard College, Cambridge, MA, United States
2Massachusetts General Hospital, Boston, MA, United States
*these authors contributed equally

**Corresponding Author:**
Jarone Lee, MD
Massachusetts General Hospital
55 Fruit Street
Boston, MA, 02114
United States
Phone: 1 617 726 2760
Email: jarone.lee@mgh.harvard.edu
common health media. Using common applications such as inTextMoji, Bitmoji, Avatoo, and many more, we can place colorful, creative, and inviting representations onto traditionally text-heavy guides in medicine. Most applications for emoji-making involve converting pictures or drawings into Unicode pictures that you can send through text, so students can use their talents in digital arts to establish new symbols and representations [9]. Applications such as EmojiRequest allow users to submit their designed emojis to a public contest, where the most popular emojis become publicly available on their phone app and website. Eventually, popularly requested emojis will be proposed to the Unicode Standard—acceptance into which will cause these emojis to be available on handheld devices worldwide.

Anyone can submit an application for a new emoji design to the Unicode Standard. There is a submission window from April 4 to July 31 every year for unique emoji designs. For the emoji application, a new design has to have a descriptive name, a category that it fits into, and a reason that necessitates its inclusion into the standard. For making decisions, the Unicode Consortium focuses on if novel emojis fit its selection factors, which are its metrics for deciding if there is a public need for the novel emoji design. These factors include its distinctiveness from other emojis, its expected use levels based on internet search analytics from similar topics, and the cultural universality of the emoji design [10]. The most important metric is the expected use levels, which is demonstrated by comparing the search popularity of the emoji’s topic with the term elephant. The emoji’s topic will be compared to elephant using Google Trends, Bing Trends, and general search analytics to show there is a public interest in the topic depicted in the emoji. This is because the elephant emoji is not the most popular or unpopular emoji, so it will be an indicator of the potential popularity of the novel emoji. Potential emojis must be submitted in a 18 x 18 pixel size, which is the size of emojis in a phone, and a 72 x 72 pixel size.

After initial submission, applicants will wait 2 to 6 weeks to see if their emoji has passed the basic review and if it would be presented to the Unicode Technical Committee (UTC) for full consideration. Emojis that do not make it past this stage cannot be reconsidered for 2 years. At the UTC meeting, around 50 to 70 new emojis will be discussed [11], and the final emoji list will be released around early March of the following year. It will usually take several months for emoji vendors such as Apple, Twitter, and Google to approve new emojis from the UTC and release them on their platforms.

Even though it might be a lofty goal, previous grassroots-based movements have successfully advocated for new emojis to be added to the Unicode Standard based on popular request. Most notably this was the development of a wide array of skin colors for emojis to allow for greater representation of different peoples. Most applications for emojis to allow for greater representation of different peoples. With using health care emojis, we are able to increase the accessibility of health care information to a larger audience. Emojis have been used in diverse patient populations such as older adult patients, non-English–speaking patients, and young children [12,13]. All of these groups reported having an increased understanding of the health care information being conveyed through emojis rather than through traditional communication methods [5,14,15]. We want the readers to use their creativity and experience to kick-start the future and much needed evolution of health care communication.

**Conflicts of Interest**

None declared.

**References**


**Abbreviations**

- **CPR**: cardiopulmonary resuscitation
- **UTC**: Unicode Technical Committee

**Please cite as:**

Marzouk S, He S, Lee J. Emoji Education: How Students Can Help Increase Health Awareness by Making Emojis
JMIR Med Educ 2022;8(4):e39059
URL: https://mededu.jmir.org/2022/4/e39059
doi: 10.2196/39059
PMID: 36367758

©Sammer Marzouk, Shuhan He, Jaronee Lee. Originally published in JMIR Medical Education (https://mededu.jmir.org), 11.11.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Incorporating Paid Caregivers Into Medical Education to Enhance Medical Student Exposure to This Essential Workforce

Samir Kamat¹, BA; George Danias¹, BS; Aneesh Agarwal¹, MBA; Sumanth Chennareddy¹, BA; Joseph Han¹, BS; Samuel Lee¹, BS

Department of Medical Education, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Samir Kamat, BA
Department of Medical Education
Icahn School of Medicine at Mount Sinai
1 Gustave L Levy Pl
New York, NY, 10029
United States
Phone: 1 2122416096
Email: samir.kamat@icahn.mssm.edu

Abstract

The implications of the COVID-19 pandemic underscored the utility of home-based health care due in part to social distancing requirements, curtailment of elective hospital procedures, and patient apprehension of the health care setting. The pandemic particularly accentuated the integral role of paid caregivers (eg, home health aides, personal care attendants, and other home care workers) in caring for patients with chronic health conditions. Given the paradigm shift toward community- and value-based health care models, paid caregivers are likely to play an even greater role as care team members. Despite the increasingly prominent role paid caregivers are assuming in health care, especially for patients who are chronically ill, in our experience as medical students, we have very little exposure to these care team members, with most interactions occurring in brief, chance encounters. Specifically, we advocate for increased medical student exposure to paid caregivers to facilitate their recognition as valuable care team members. We propose to achieve this through (1) classroom-based module learning with live paid caregivers and (2) plain language communication training to enhance reciprocal engagement.

(JMIR Med Educ 2022;8(4):e38329) doi:10.2196/38329

KEYWORDS
medical education; education; student; communication; perspective; medical student; paid caregiver; caregiver; health care model; home-based health care; patient care; health care provider; medical student; student experience; training; care team; integration; clinical decision

Introduction

Although the COVID-19 pandemic underscored the critical role of the many frontline workers who support the general public daily, our ambulatory care rotation exposed us to another unacclaimed part of the care team: paid caregivers. Paid caregivers include home health aides, personal care attendants, and other home care workers who care for people with functional impairments at home [1]. This categorization excludes relatives, friends, and spouses, even those with health care experience, who are unpaid caregivers. Paid caregivers provide essential support to various populations, although they primarily care for geriatric and homebound geriatric patients. The number of homebound adults in the last decade has more than doubled. Among adults aged over 70 years, 4.2 million were homebound in 2020, compared to 1.6 million in 2019, likely due to the results of the pandemic; this further highlights the increasing need for paid caregivers to provide support in the home [2]. By characterizing their roles in clinical and community environments and sharing anecdotes from our educational experiences, we hope to make the case that medical education should more deeply expose students and trainees to paid caregivers.

Our Experiences With Paid Caregivers

Underscore Their Integral but
Underrecognized Contributions to Patient Care

As a part of our Geriatric Medicine clerkship, we rotated through the Mount Sinai Visiting Doctors program, which provides home-based primary care for homebound patients in New York [3]. Many patients we encountered had functional impairments that limited their quality of life, and many others relied on paid caregivers to support their clinical care. The most striking finding was the extent of paid support; from well-appointed apartments to government-subsidized units, patients of all socioeconomic statuses routinely relied on paid caregiver assistance in the home environment. In one encounter with a patient with a severe disability, the physician relied on the home health aide for information on the patient’s functioning, including getting out of bed, bowel movement regularity, sleep quality, and medication adherence. Although providers are only present for brief periods with their patients, paid caregivers fill the temporal and informational gaps in providing a more complete picture for improved care delivery.

In general, paid caregivers help individuals with activities of daily living (eg, eating and toileting) and instrumental activities of daily living (eg, housekeeping and meal preparation) [4]. Trained paid caregivers with more specialized skills may be able to provide additional assistance for patients with dementia, behavioral health issues, and palliative care needs [4]. One study found that half of community-residing patients with advanced dementia received paid care, with 30% having part-time paid care and 18% having full-time paid care [5]. Although paid caregivers help with daily activities, they also perform a wide variety of other tasks, including maintaining physical living conditions, participating in family dynamics, identifying emergent clinical changes, and assisting with patient self-care of chronic medical issues [1,4]. Notably, the literature suggests that paid caregivers provide significant mental and emotional health support for patients as well [1,4]. For one homebound patient with dementia, beyond providing physical care, it was evident through the paid caregiver’s interactions and attitude toward the patient that they considered the patient to be similar to family. In addition to characterizing the patient’s symptoms, mood, and life events during our visit, the paid caregiver also demonstrated involvement in supporting the patient’s hobbies, interests, and family obligations. During our interview, the patient indicated that they considered the paid caregiver an integral part of their family. The paid caregiver consistently went “above and beyond” by demonstrating interest in the health of other family members and providing emotional support, as well as through smaller actions such as bringing the patient’s favorite foods. This indicated the strength of the patient–paid caregiver relationship and how paid caregivers may have roles that impact broader dimensions of health and well-being. Due to the profound role they play in care, paid caregivers may be able to improve patient outcomes and support high-value care from long-term care funding programs such as Medicaid [1].

Current Medical Education Lacks Intentional Student Exposure to Paid Caregivers

Although enlightening, our brief experience with paid caregivers in the home only captures a small fraction of the paid caregiver workforce and their services. Overall, 3 million paid caregivers furnish important services in the community for patients with chronic and acute conditions [6]. Considering the substantial role paid caregivers play in the lives and health of patients, as we characterize through our experiences, we strongly believe that medical education should include deliberate and sufficient exposure to these essential care team members.

As part of the paradigm shift toward value-driven interventions for better chronic condition management, several initiatives have successfully incorporated paid caregivers, recognizing their potential to improve health for certain patient populations. For example, the Mount Sinai Hospital Home Care Collaboration Solutions trained home care staff and home care aides to keep track of changes in their patients’ health statuses [4]. Another study from St. John’s Well Child and Family Center presented pilot data showing a 40% improvement in medication adherence and a decrease in patients’ unhealthy days by over 38% (25.3 to 15.6 months) when trained aides did additional chronic care management tasks and were involved with the medical and social care teams [4]. Interventions involving paid caregivers appear to be on an upward trend, particularly as health policy emphasizes community-based care models [1]. Exposure to these initiatives, whether in direct interactions during clinical rotations or through case-study didactic modules, would provide valuable insight for future health professionals into how a paid caregiver may be helpful for a patient and how to best integrate paid caregivers into the health care team.

As a formative stage of training, medical education is uniquely positioned to highlight the increasingly important roles of paid caregivers. Although we were fortunate to get exposure through rotations, more widespread interactions with paid caregivers are needed, potentially through family medicine or dedicated ambulatory care rotations, so that medical trainees can observe and experience the importance of paid caregivers who care for certain patient populations.

We Propose Classroom-Based Learning and Plain Language Skills Training to Enable Future Physicians to Better Involve Paid Caregivers in Patient Care

Although experience in the clinical environment with a paid caregiver is ideal, such exposure may not always be possible. Therefore, we propose that a preclinical clinical skills module detailing the role and contributions of paid caregivers in patient care is useful if clinical opportunities are limited. Low clinical exposure can be overcome through classroom-based exposure to paid caregivers, in-person or remote video teleconferencing, and communications training, perhaps through standardized patients. Furthermore, the clinical skills module should include...
communication skills needed for physician–paid caregiver interactions and video simulations. Errors in patient care may arise due to inadequate health literacy among paid caregivers, which can potentially be prevented by improving physician–paid caregiver communication [7]. Thus, the module should teach learners, as future physicians, how to communicate effectively with paid caregivers in plain language. Intentionally developing plain language communication skills will make physicians more accessible and reduce communication barriers with paid caregivers and patients [8]. Additionally, through module-based learning, experienced paid caregivers can voice concerns, questions, and knowledge of the patient’s care to future physicians, creating bilateral communication between providers. Thus, such a module will emphasize paid caregivers’ broad roles outside of just personal care to build a reciprocal clinician–paid caregiver partnership.

As students will undoubtedly come across paid caregivers in their clinical rotations and careers, early recognition of these critical players will contribute to a more comprehensive medical education. Physicians may not appreciate a paid caregiver’s value to the patient [1]. Specific education in medical training on how the paid caregiver can contribute to the patient’s care, presentation of evidenced-based outcomes on how paid caregivers improve patient quality of life, and allowing students to talk to or hear from paid caregivers in the classroom setting will help raise future physician awareness surrounding the presence and critical roles of paid caregivers for certain patient populations. Early exposure to paid caregivers and how to effectively involve them can afford trainees a more holistic approach to patient care in their future clinical practice.

Acknowledgments
We thank Dr Jennifer Reckrey and Dr Ravishankar Ramaswamy for providing critical feedback.

Conflicts of Interest
None declared.

References


Perceptions of US Medical Students on Artificial Intelligence in Medicine: Mixed Methods Survey Study

David Shalom Liu¹, MS, BS; Jake Sawyer¹, BS; Alexander Luna¹, BS; Jihad Aoun¹, MS, BS; Janet Wang¹, BS; Lord Boachie¹, BS; Safwan Halabi², MD; Bina Joe³, PhD

¹College of Medicine and Life Sciences, University of Toledo, Toledo, OH, United States
²Pediatric Radiology, Ann & Robert H Lurie Children's Hospital of Chicago, Chicago, IL, United States
³Department of Physiology and Pharmacology, College of Medicine and Life Sciences, University of Toledo, Toledo, OH, United States

Corresponding Author:
Bina Joe, PhD
Department of Physiology and Pharmacology
College of Medicine and Life Sciences
University of Toledo
3000 Arlington Ave
Toledo, OH, 43614
United States
Phone: 1 419 383 4144
Email: bina.joe@utoledo.edu

Abstract

Background: Given the rapidity with which artificial intelligence is gaining momentum in clinical medicine, current physician leaders have called for more incorporation of artificial intelligence topics into undergraduate medical education. This is to prepare future physicians to better work together with artificial intelligence technology. However, the first step in curriculum development is to survey the needs of end users. There has not been a study to determine which media and which topics are most preferred by US medical students to learn about the topic of artificial intelligence in medicine.

Objective: We aimed to survey US medical students on the need to incorporate artificial intelligence in undergraduate medical education and their preferred means to do so to assist with future education initiatives.

Methods: A mixed methods survey comprising both specific questions and a write-in response section was sent through Qualtrics to US medical students in May 2021. Likert scale questions were used to first assess various perceptions of artificial intelligence in medicine. Specific questions were posed regarding learning format and topics in artificial intelligence. Specific questions were posed regarding learning format and topics in artificial intelligence.

Results: We surveyed 390 US medical students with an average age of 26 (SD 3) years from 17 different medical programs (the estimated response rate was 3.5%). A majority (355/388, 91.5%) of respondents agreed that training in artificial intelligence concepts during medical school would be useful for their future. While 79.4% (308/388) were excited to use artificial intelligence technologies, 91.2% (353/387) either reported that their medical schools did not offer resources or were unsure if they did so. Short lectures (264/378, 69.8%), formal electives (180/378, 47.6%), and Q and A panels (167/378, 44.2%) were identified as preferred formats, while fundamental concepts of artificial intelligence (247/379, 65.2%), when to use artificial intelligence in medicine (227/379, 59.9%), and pros and cons of using artificial intelligence (224/379, 59.1%) were the most preferred topics for enhancing their training.

Conclusions: The results of this study indicate that current US medical students recognize the importance of artificial intelligence in medicine and acknowledge that current formal education and resources to study artificial intelligence–related topics are limited in most US medical schools. Respondents also indicated that a hybrid formal/flexible format would be most appropriate for incorporating artificial intelligence as a topic in US medical schools. Based on these data, we conclude that there is a definitive knowledge gap in artificial intelligence education within current medical education in the US. Further, the results suggest there is a disparity in opinions on the specific format and topics to be introduced.

(JMIR Med Educ 2022;8(4):e38325) doi:10.2196/38325
Introduction

Artificial intelligence (AI) is the science of simulating human intelligence with machines for a variety of applications in all sectors, including medicine. Rapid advances in computational capabilities and cloud-based data systems, especially for the machine learning and deep learning subtypes, have led to innovative applications of AI in both clinical medicine and medical research [1-9]. For example, the CheXNeXt algorithm for chest radiograph diagnosis was found to perform at a level similar to radiologists [1]. AI algorithms can also predict future adverse medical events much better than traditional methods. One example is predicting aneurysms [4]. Even in the realm of psychiatry, AI algorithms can help detect subtle, yet key information about patients, such as speech patterns, that can predict subsequent psychosis onset [3]. For medical research applications, AI has been able to recognize complex patterns in large amounts of data (eg, gene expression and gut microbiota) to classify clinical conditions, such as cardiomyopathies [10], inflammatory bowel disease [11], and cardiovascular diseases [12]. According to Topol [6], this is beginning to have an impact at 3 levels: for clinicians, predominantly via rapid, accurate image interpretation; for health systems, by improving workflows and potentially reducing medical errors; and for patients, by enabling them to process their own data to promote health. Thus, the practice of clinical medicine is poised to drastically change with the inevitable infusion of AI.

Given the changing landscape of medical practice, a critical question is whether current medical students are being prepared during their training to effectively understand and work with AI. The intent to promote such training is evident. In the United States in 2018, the American Medical Association made it an official policy (H-480.940) to encourage medical students to expand their knowledge of AI as it pertains to medicine. Given the changing landscape of medical practice, a critical question is whether current medical students are being prepared during their training to effectively understand and work with AI. The intent to promote such training is evident. In the United States in 2018, the American Medical Association made it an official policy (H-480.940) to encourage medical students to expand their knowledge of AI as it pertains to medicine. 

Methods

Survey Design

The survey was designed using the online app Qualtrics (Qualtrics International Inc); all the survey questions are detailed in Multimedia Appendix 1. The survey participants provided informed consent at the beginning of the survey, which had 2 main components. The first section of the survey gathered the demographics and medical education of the participants. The second section of the survey aimed to assess perceptions by medical students regarding AI in medicine, and familiarity of US medical students regarding AI and its application in medicine. The informed consent form described the survey length (5 minutes), the investigator, the purpose of the study, and the privacy policy. To limit unauthorized access, only 2 researchers could access the data. This mixed methods survey consisted of 24 multiple-choice questions on a 5-point Likert scale and a write-in section. Survey questions were generated by referring to previous, similar studies on the perceptions by medical students of AI in other countries [19-22]. Further, novel questions were added on the preferred formats for an AI curriculum in medical school. Finally, a write-in section was added that allowed for students to voice any thoughts on AI education in medicine that were not captured in the survey. There was no randomization of the order of survey questions. The premise for the length of the survey (5 minutes and 24 questions) was based on previous, similar published surveys [19-22]. Questions were presented equally on 6 webpages. Respondents could review and change their answers through a back button. Respondents were required to answer all questions, though a “not applicable” option was provided for some questions.

Survey Distribution

Medical students across the United States were the target audience. To distribute the survey to the target audience, the deans of student affairs at all 169 US allopathic and osteopathic medical schools were contacted via email for participation. Fourteen allopathic and 3 osteopathic medical schools agreed to distribute the survey. Next, a link to the Qualtrics survey was sent to the relevant school faculty to be distributed to their students. There were no financial incentives given to the survey participants. 

KEYWORDS

artificial intelligence; eHealth; digital health; integration; medical education; medical curriculum; education; medical student; medical school; elective course
participants; their participation was voluntary. Overall, the survey was distributed to 11,248 students, of whom 390 responded (response rate 3.5%). A unique response ID was created on the Qualtrics survey page for each respondent to ensure that respondents did not submit the survey twice. Once completed, opening the survey link again with the same computer showed the completion page of the survey.

Statistical Analyses
Some respondents did not answer all the questions, so we correspondingly reduced the value of N for questions that were omitted. Responses that were completed in less than 1 minute, which was determined to be the fastest time someone could realistically complete our pilot survey, were excluded. We first compared the Likert responses between respondents who had received formal AI training and those who had not. Next, we determined whether the responses on how and what the medical students wanted to learn differed between those who wanted to spend less than 3 hours per month studying AI and those who wanted to spend more than 3 hours per month. The Pearson chi-square or Fisher exact test were used depending on the context. A P value of <.05 was considered significant.

Ethics Approval
This study and its anonymous online Qualtrics survey (IRB protocol 300975) was reviewed by a Social, Behavioral, and Educational Institutional Research Board Committee member at the University of Toledo. The committee member determined this study to meet criteria for exemption per 45 CFR 46.104 (d)(2)(i) or (ii).

Results
Demographics
Survey responses from 390 US medical students with an average age of 26 (SD 3) years were collected from 17 of the 169 US medical schools contacted (the schools are listed in Textbox 1). A total of 390 responses were received. Table 1 summarizes the demographic data of our surveyed sample. The response rate was 3.5% (390/11,248). It was not possible to calculate how many students opened the recruitment email, so we cannot report viewing or participation rates. A total of 250 of 390 students (64.1%) were from allopathic (MD) programs while 140 of 390 students (35.9%) were from osteopathic (DO) programs. The 390 participants included US medical students from all 4 years of medical education: first year (142, 36.4%), second year (94, 24.1%), third year (77, 19.7%), and fourth year (64, 16.4%). Additionally, 8 of 390 medical students (2.7%) were in the PhD component of a dual MD/PhD program. Only 34 of 390 students (8.7%) reported having received a formal education in AI topics via college courses. The median response time for the survey was 5 minutes and 18 seconds.

Textbox 1. Names of schools surveyed (n=17).
- Chicago Medical School at Rosalind Franklin University of Medicine & Science
- Hofstra Northwell School of Medicine
- Medical College of Wisconsin
- Ohio State University College of Medicine
- Stony Brook University School of Medicine
- University of Central Florida College of Medicine
- University of Colorado School of Medicine
- University of Hawaii John A. Burns School of Medicine
- University of Kentucky College of Medicine
- University of Toledo College of Medicine and Life Sciences
- Virginia Commonwealth University School of Medicine
- Warren Alpert Medical School of Brown University
- Washington University in St. Louis School of Medicine
- Kansas City University College of Medicine—Joplin Campus
- Kansas City University College of Medicine—Kansas City Campus
- Ohio University Heritage College of Medicine
- West Virginia School of Medicine
Table 1. Demographics of survey participants (N=390).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Current year in medical school, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>First year</td>
<td>142 (36.4)</td>
</tr>
<tr>
<td>Second year</td>
<td>94 (24.1)</td>
</tr>
<tr>
<td>Third year</td>
<td>77 (19.7)</td>
</tr>
<tr>
<td>Fourth year</td>
<td>64 (16.4)</td>
</tr>
<tr>
<td>MD/PhD</td>
<td>8 (2.7)</td>
</tr>
<tr>
<td><strong>Age, years</strong></td>
<td>25.8 (3.4)</td>
</tr>
<tr>
<td>Range</td>
<td>20-50</td>
</tr>
<tr>
<td><strong>Prior formal AI education, n (%)</strong></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>356 (91.3)</td>
</tr>
<tr>
<td>Yes</td>
<td>34 (8.7)</td>
</tr>
</tbody>
</table>

Attitudes Toward AI in Medicine

The survey assessed the general attitudes US medical students had toward AI in medicine (Figure 1). For example, 351 of 390 (90%) students agreed that AI will be a significant feature in medicine during their lifetime. Furthermore, 308 of 388 students (79.4%) were excited about using AI technology as a future physician. Despite this excitement, 238 of 388 respondents (61.3%) were broadly worried about the ethics of using AI in medicine.

The participants were prompted to select the 3 medical subspecialties they believed would be most affected by AI integration. The respondents selected diagnostic radiology (278/390, 71.3%), pathology (167/390, 42.8%), and interventional radiology (95/390, 24.6%) (Multimedia Appendix 2). Furthermore, we found that 70 of 386 (18.1%) students were less likely to enter specialties they thought would be affected by the anticipated integration of AI into that specialty (Multimedia Appendix 2).

Figure 1. Attitudes toward and familiarity with AI in medicine of US medical students. Values indicate the number of responses (corresponding to the legend) to each statement, shown as n (%). N=388 for all statements except “AI will take on a significant role in my lifetime” (N=390) and “It’s hard to understand and approach AI because of media sensationalism” (N=387). AI: artificial intelligence; AUC: area under the curve; CNN: convolutional neural network; ROC: receiver operating characteristics.
Knowledge of and Familiarity With AI in Medicine

Next, this survey assessed the familiarity of medical students with the application of AI in medicine (Figure 1). Only a small portion of students (54/388, 13.9%) indicated they had knowledge of core AI concepts (eg, cross validation and deep learning). Our findings further indicated that the medical students were unfamiliar with current clinical applications of AI, with only 89 of 388 (22.9%) agreeing that they could “list some examples of recent clinically-relevant AI research,” whereas 267 (68.8%) disagreed. Less than half of respondents (152/388, 39.2%) agreed that they could “separate ‘hype’ AI articles vs. clinically relevant AI articles,” whereas 162 of 388 (41.8%) disagreed. Moreover, 167 of 387 (43.2%) agreed that it was “hard to understand and approach AI because of media sensationalism,” while 124 of 387 (32%) neither agreed nor disagreed. Only 96 of 388 (24.8%) disagreed that media sensationalism made approaching and understanding AI more difficult. This survey further assessed the sources that the students had used to learn about AI in medicine; these were found to include media (263/386, 68.1%), family and friends (134/386, 34.7%), online forums (98/386, 25.4%), and professors or doctors (89/386, 23.1%) (Multimedia Appendix 2).

Perspectives on AI in Current Medical Education Curricula

Next, the survey assessed medical students’ perspectives on AI in current medical school education. Most students (347/388, 89.4%) agreed that they wanted to “learn what medical students should know about AI in medicine” (Figure 2). A portion of students (60/388, 15.5%) agreed that learning relevant AI topics (eg, ethics or pros and cons of AI) could significantly detract from their medical school education, while a majority of the surveyed students (258/388, 66.5%) disagreed. Despite overwhelmingly positive responses toward this topic, only 34 of 387 (8.8%) medical students agreed that their respective medical schools offered resources to explore the topic of AI in medicine. Finally, most students (355/388, 91.5%) agreed with the statement “some training in AI concepts and related topics during medical school can be useful for my future career.”

Figure 2. Perspectives by US medical students of AI in current medical education. Values indicate the number of responses (corresponding to the legend) to each statement, shown as n (%). N=388 for all statements except "My school offers resources if I want to explore the topic of AI in medicine" (N=387). AI: artificial intelligence.
Preferred AI Resources and Topics

We also assessed opinions regarding AI in medical education. Over half the students (197/379, 52%) reported that 1 or 2 hours per month would be their preferred maximum amount of time spent on learning AI in medical school, whereas 123 of 379 (32.5%) students preferred that more than 3 hours per month be spent on exploring this topic and 11 of 379 (2.9%) preferred no time at all (Table 2). When the students were asked to select the resources or formats that would be most useful to learn AI in medicine, their 3 most-selected choices were short lectures (264/378, 69.8%), formal preclinical electives (180/378, 47.6%), and Q and A panels (167/378, 44.2%) (Figure 3). The medical students reported that the AI-related topics they were most interested in were “fundamental concepts of AI” (247/379, 65.2%), “when to use AI in medicine” (227/379, 59.9%), “strengths and weaknesses of using AI in medicine” (224/379, 59.1%), “ethics of AI” (211/379, 55.7%), and “what aspects of a physician’s job can be replaced with AI and which can’t” (203/379, 53.6%) (Figure 4).

Table 2. Time per month US medical students would prefer to study artificial intelligence topics (N=379).

<table>
<thead>
<tr>
<th>Time preferred</th>
<th>Responses, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>11 (2.9)</td>
</tr>
<tr>
<td>30 minutes</td>
<td>48 (12.7)</td>
</tr>
<tr>
<td>1 hour</td>
<td>100 (26.4)</td>
</tr>
<tr>
<td>2 hours</td>
<td>97 (25.6)</td>
</tr>
<tr>
<td>3 hours</td>
<td>43 (11.4)</td>
</tr>
<tr>
<td>4 hours</td>
<td>41 (10.8)</td>
</tr>
<tr>
<td>5 hours or more</td>
<td>39 (10.3)</td>
</tr>
</tbody>
</table>

Figure 3. Media preferred by US medical students to explore artificial intelligence topics. Responses were classified based on how many hours a respondent had stated they would like to spend per month studying artificial intelligence in an earlier question. The red bars represent those who answered 2 hours or less and teal bars represent those who answered 3 hours or more. *P<.05, **P<.01, ***P<.001. Exact P values can be found in Multimedia Appendix 2. AI: artificial intelligence.
Figure 4. Artificial intelligence topics preferred by US medical students. Responses were classified based on how many hours a respondent had stated they would like to spend per month studying artificial intelligence in an earlier question. The red bars represent those who answered 2 hours or less and teal bars represent those who answered 3 hours or more. *P < .05, **P < .01, ***P < .001. Exact P values can be found in Multimedia Appendix 2. AI: artificial intelligence.

When comparing the responses of those who preferred to spend ≤2 and ≥3 hours per month to learn AI with a chi square analysis, we found significantly different responses. Compared to the ≤2 hours per month group, the ≥3 hours per month group was more interested in short lectures (76.1% vs 66.1%, P = .06), programming workshops (P < .001), AI research symposia (P = .01), interdisciplinary research teams (P = .1), and national conferences on AI (P < .001). Compared to the ≤2 hours per month group, the ≥3 hours per month group was more interested in “AI models in the clinic” (P = .01), “types of models in AI” (P < .001), “AI in medical research” (P = .01), “recent AI/medical innovations” (P = .03), “using python” (P < .001), and “multidisciplinary AI research teams” (P = .03).

Write-in Responses
A total of 23 free responses were collected (Multimedia Appendix 2, Table S4). The responses were collected and grouped into 3 categories: generally positive, academic concerns, and ethical concerns.

Several respondents emphasized the need for increasing medical students’ awareness of the role of AI in medicine. Examples include the following: “This is an extremely important topic that needs more focus,” “I honestly have heard very little about the subject,” and “I feel like I really didn’t know at all about AI in medicine and hope there will be educational opportunities in the future.”

Other respondents expressed their concerns about incorporating this topic into the medical curriculum: “I am against adding more components to preclinical medical education...” and “I don’t think medical students have enough computer science and engineering background to learn much about AI.” Another student noted, “It isn’t terribly necessary for medical students to fully grasp all the fundamentals of AI nor for them to have programming workshops...[However,] it would be a great disservice for people walking into the field to be unfamiliar with the implications and applications of AI.”
Discussion

Perceptions by US Medical Students of AI in Medicine

The current study was conducted to document perceptions by current US medical students of AI and the implementation of AI knowledge into medical education. In agreement with similar survey reports from other countries [19-22], our study found that 89.4% (347/388) of surveyed US medical students wanted to learn about AI in medicine and agreed that AI would play a significant role in medicine during their future professional lives as physicians. These views did not significantly differ between those who received formal AI training and those who did not. Overall, our study supports the conclusion that current medical education in the US lags behind the enthusiasm of medical students to learn about AI with appropriate learning resources. As AI becomes continually integrated into medicine, our survey indicates that US medical education for future physicians would benefit from the addition of educational components on AI in medicine. Overall, our findings are consistent with similar survey reports from other countries. Previous survey studies showed that 70% of German, 83% of South Korean, and 78% of UK medical students agreed that AI should be part of medical training [19,21].

Although our survey suggests that 89.4% (347/388) of US medical students would like to learn about AI, only 13.9% (54/388) indicated that they understood fundamental AI terms and concepts (Figure 1). This discrepancy has also been observed in medical students in other countries [19,21]. A potential reason for this is the lack of relevant AI resources and expertise in medical education [28]. It is evident that health care in the 21st century will continue to evolve into an interdisciplinary and integral partnership between physicians, engineers, and computer scientists [26,30]. Therefore, it would be beneficial for future physicians to learn the fundamentals of AI in medical applications to comfortably work with AI technologies and meaningfully apply incoming technological innovations in medicine. It should be noted that certain residency programs, most prominently in radiology, have emphasized the requirement that their trainees have a knowledge of AI fundamentals [14,31-36]. This could be due to the higher prevalence of the use of imaging in AI relative to, for example, genomic data, which is only beginning to be studied with AI in medicine.

Our survey also showed that 91.5% (355/388) of surveyed US medical students agreed that training on AI concepts in medical school was important for their future careers (Figure 2) and 79.3% (308/388) were excited to use AI technologies (Figure 1). These opinions were not significantly affected by whether the respondents had prior formal AI training and indicate that current US medical students not only realize the need for incorporating AI topics into medical education as a “checkbox” to better prepare for future technological revolutions in medicine but are also enthusiastic about embracing such changes. This point is further reflected by our finding that 18.1% (70/386) of US medical students expressed hesitance in pursuing 1 of their top 3 desired specialties due to the incorporation of AI in that specialty (Multimedia Appendix 2). In contrast, other survey studies have found that 54% and 49% of medical students in Germany and Canada, respectively, were less likely to choose certain specialties due to the future incorporation of AI [19,22]. Most surveyed US medical students did not consider the inclusion of AI in medical education as a distraction but were instead excited to learn about AI in medicine, which was further demonstrated by their strong eagerness to explore AI in medical topics (Figure 1).

Furthermore, there is a lack of a structured approach to teaching AI systems in medical education. One of the questions our survey aimed to answer was what methodologies medical students prefer for learning about AI in health care (Figure 3). Our data showed that students preferred more medical student–directed and flexible opportunities to learn about AI in medicine, such as short lectures, formal preclinical electives, Q and A panels, and programming workshops. Currently, there is a lack of such opportunities; we found that students had obtained information on AI from other sources, including the media (263/386, 68.1%), family and friends (134/386, 34.7%), and online forums (98/386, 25.4%) (Multimedia Appendix 2). This finding is consistent with previous reports on where German medical students obtained their exposure to AI in medicine [21]. One example of curricular integration of AI is the University of Toronto Faculty of Medicine’s 14-month Computing for Medicine course, which began in 2019 [37]. However, the pace of change in medical education in adding AI-related topics is relatively slow compared to the pace at which the application of AI in medicine is currently progressing [38]. Thus, some leading experts have pushed for more radical changes in medical education or more extracurricular opportunities for students [17,26,29]. With the current shift of US medical education from strictly in-class learning to increased reliance on external resources (such as popular online learning platforms from Pathoma, Boards & Beyond, and Osmosis), as well as the advent of massive open online courses as a primary source of self-directed AI education, current medical students may be more receptive to self-directed learning based on extracurricular resources [39,40]. Thus, although most surveyed medical students preferred learning about AI through formal media, either formal curricular changes to incorporate AI should ramp up in pace or, as a potential alternative, online, freely accessible resources should be created for medical students to learn about AI [29].

Finally, it is important to distinguish between 2 sets of AI competencies. The first includes “core” competencies that most future physicians should know for their day-to-day practice, and the second includes “advanced” competencies for future physicians who intend to drive research and innovation in the field of AI in medicine [29]. While integration of AI topics into formal curricula may be sufficient for most medical students, research opportunities and mentorships should be provided specifically for future physician–scientists and innovators. Figures 3 and 4 show potential areas of AI topic concentration that differ between the 2 groups. Topics such as “fundamental concepts of AI,” “strengths and weaknesses of AI,” and “ethics of AI” were deemed interesting by both groups of respondents (ie, ≤2 hours vs ≥3 hours per month preferred for studying AI). Topics such as “translational science,” “global health and AI,”
“AI in medical research” can be directed specifically toward medical students who wish to go beyond the minimum required knowledge for future physicians regarding AI topics. To deliver these 2 sets of AI competencies, different resources should be employed for each group of learners. For example, our study showed that programming workshops, research teams, and conferences should be created that are tailored to the “advanced” learners, while short lectures would be highly valuable for both sets of learners.

Limitations

Our study is admittedly not without limitations. First, although it agrees with results from other nations, our results do not fully represent the entire US medical student body due to a small sample size. Moreover, there might have been selection bias, because the respondents might have been students who were particularly interested in AI, especially considering that there was no financial incentive for survey completion. Finally, we did not analyze the importance of adding AI topics to the medical school curriculum in the context of other, already existing, medical school curricula, and we thus did not gauge the relative importance of AI topics.

Conclusion

A large majority of current surveyed US medical students recognized the important role of AI in medicine and expressed excitement to learn more about AI fundamentals and applications in medicine. Nonetheless, only a minority of the students had knowledge of AI and medicine. The surveyed students were excited to learn about this topic and preferred formal, yet flexible, ways to approach AI in medical schools. However, currently available resources to learn about AI-related topics are limited in most US medical schools. Based on our work and prior surveys in other nations, we highlight the acute need to incorporate AI-related topics in the medical school curriculum.

Acknowledgments

BJ acknowledges grant support from the National Institutes of Health (HL143082).

Disclaimer

The authors declare that they had full access to all of the data in this study and the authors take complete responsibility for the integrity of the data and the accuracy of the data analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Qualtrics survey assessing perceptions of US medical students on AI education. [PDF File (Adobe PDF File), 234 KB - mededu_v8i4e38325_app1.pdf]

Multimedia Appendix 2

Detailed answers to survey questions. [DOCX File, 36 KB - mededu_v8i4e38325_app2.docx]

References


Abbreviations

AI: artificial intelligence
Google Images Search Results as a Resource in the Anatomy Laboratory: Rating of Educational Value

Alexandra Elisabeth Wink1, MS, PhD; Amanda N Telfer2, MS; Michael A Pascoe2, PhD

1Division of Translational Anatomy, Department of Radiology, UMass Chan Medical School, Worcester, MA, United States
2Department of Cell and Developmental Biology, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, United States

Corresponding Author:
Alexandra Elisabeth Wink, MS, PhD
Division of Translational Anatomy
Department of Radiology
UMass Chan Medical School
55 Lake Avenue North
Worcester, MA, 01655
United States
Phone: 1 508 856 6920
Email: alexandra.wink@umassmed.edu

Abstract

Background: Preclinical medical learners are embedded in technology-rich environments, allowing them rapid access to a large volume of information. The anatomy laboratory is an environment in which faculty can assess the development of professional skills such as information literacy in preclinical medical learners. In the anatomy laboratory, many students use Google Images searches in addition to or in place of other course materials as a resource to locate and identify anatomical structures. However, the most frequent sources as well as the educational quality of these images are unknown.

Objective: This study was designed to assess the sources and educational value of Google Images search results for commonly searched anatomical structures.

Methods: The top 10 Google Images search results were collected for 39 anatomical structures. Image source websites were recorded and categorized based on the purpose and target audience of the site publishing the image. Educational value was determined through assessment of relevance (is the searched structure depicted in the image?), accuracy (does the image contain errors?), and usefulness (will the image assist a learner in locating the structure on an anatomical donor?). A reliable scoring rubric was developed to assess an image’s usefulness.

Results: A total of 390 images were analyzed. Most often, images were sourced from websites targeting health care professionals and health care professions students (38% of images), while Wikipedia was the most frequent single source of image results (62/390 results). Of the 390 total images, 363 (93.1%) depicted the searched structure and were therefore considered relevant. However, only 43.0% (156/363) of relevant images met the threshold to be deemed useful in identifying the searched structure in an anatomical donor. The usefulness of images did not significantly differ across source categories.

Conclusions: Anatomy faculty may use these results to develop interventions for gaps in information literacy in preclinical medical learners in the context of image searches in the anatomy laboratory.

(JMIR Med Educ 2022;8(4):e37730) doi:10.2196/37730

KEYWORDS
anatomy laboratory; information literacy; internet search; anatomical images; scoring rubric; Google; images; educational value; literacy information; medical education; medical students; anatomy
Medical students in their clerkship years and medical residents rely heavily on online resources such as UpToDate and Google for general study as well as when preparing for patient encounters [3-5]; thus, these learners need to develop strong information literacy skills [6].

Gross anatomy is often one of the first experiences of a learner’s medical school career, and therefore the anatomy laboratory is seen as an opportunity to teach and assess professional skills such as accountability and teamwork [7]. Because the anatomy laboratory is also an environment for (digital) information-seeking [8], this environment is one in which faculty can promote information literacy as a professional skill.

Many anatomy laboratories are equipped with computers or other internet-enabled devices that allow medical students to access dissection instructions and other course materials. Laboratory faculty often observe students performing web searches for images of anatomical structures [8], presumably as an alternative to using a hard-copy atlas or another course-sanctioned resource. This is consistent with students’ general preference for online resources such as Google and Wikipedia [3,9,10]. Potential reasons for this preference are the ease of access and interactivity associated with a search engine compared to flipping through a hard-copy textbook [11,12] or a perceived insufficiency of their other course materials [13].

Criticisms of Google Images include that the results are not specialized, detailed textual information is missing, image quality is variable, irrelevant results are time-consuming, and, importantly, images are not reliable or from valid sources [11]. The extent to which online resources are suitable for medical students has been debated by previous investigators [14-16], and this is perhaps dependent on the complexity and objectiveness of the subject matter. Few analyses of online images have been conducted, and those that were performed focused on certain medical specialties and conditions [17].

Medical students’ predilection for Google Images searches in the anatomy laboratory raises concerns about information literacy. Kingsley et al [18] found that students who preferred Google to other sources lacked the ability to retrieve and evaluate evidence-based information. Further, Google’s accessibility and ease of use may outweigh any concerns about the accuracy and trustworthiness of information [3]. In a study of online resource use by medical residents, Duran-Nelson et al [4] suggested that when using online resources, residents may value speed over quality of information. Assuming learners hold the same preferences when seeking online information in their preclinical years, a learner may select a top Google Images search result regardless of its educational quality. To address gaps in information literacy in preclinical learners, an examination of the content of these online search results is justified.

Given the gaps in knowledge presented above, the objectives of this study were to (1) report the sources of top Google Images search results for anatomical structures and concepts and (2) evaluate these images for their educational quality.

### Methods

#### Ethical Considerations

No ethical approval was required for this study as there were no human subjects; thus, this did not meet the criterion of "human subjects research" as defined by federal regulations and the UMass Chan Medical School Institutional Review Board.

#### Image Search Retrieval

To gather top Google Images search results for anatomical structures and concepts, one author (AEW) searched for 5-10 “high-yield” anatomical structures, groups of structures, or relations representing each of the regional content areas taught in a typical medical gross anatomy course: back and limbs, thorax, abdomen, pelvis, and head and neck. These structures included emphasized (eg, bolded) terms in laboratory manuals/dissection instructions, and structures that were frequently emphasized in didactic sessions or tested on practical examinations.

The Google Images searches were performed in January and May of 2020. Google places images closer to the top of the search results if the image is located centrally or at the top of a webpage, or if the webpage or image has been updated recently. Authority of the website is also an important factor in signaling where an image is ranked on a search results page [19]. Screenshots were taken of the top 10 image results for each term and organized in a slideshow file shared among the authors.

#### Source of Images

For each image, the name of the website that published the image was recorded and the website was visited to ascertain the following information: (1) author/creator, (2) target audience, and (3) mission/purpose of the website. Two authors (AEW and MAP) created and defined the categories of websites posthoc based on one or more of these three criteria. After creating and defining the categories, the two authors (AEW and MAP) sorted the websites into these categories independently and then compared their categorizations to calculate initial percent agreement. Every disagreement in categorization was then resolved through discussion to arrive at the final categorization.

#### Educational Quality

##### Relevance

An image was defined as “relevant” if it depicted the searched structure [20], and more particularly, if the image was of human anatomy (eg, an illustration of the broad ligament of the uterus of a horse was deemed not relevant). Images classified as not relevant were excluded from further analysis.

##### Accuracy of Images

All three authors independently assessed the relevant images for errors; if no errors were detected, the image was classified as “accurate.” Errors, as defined in this study, included mislabeled structures, and misrepresentations of anatomical structures, locations, and relationships. Anatomical variants, pathological presentations, omissions (eg, a structure not depicted for the sake of simplicity), minor misspellings (eg, supraspinatous vs supraspinatus), and outdated terminology no
Usefulness

Usefulness of an image was defined broadly by whether an image would allow a learner to successfully locate or identify the structure in a human anatomical donor during dissection. The lack of an existing, validated rubric to assess the usefulness of anatomical images according to this definition necessitated its development in this investigation. All authors constructed this rubric following the procedures outlined by Moskal and Leydens [21] and Mertler [22]. The authors validated the initial iteration of the rubric using a small sample of images, and then modified the rubric domains and definitions to rate all of the images. The domains present on the rubric were (1) completeness, (2) cognitive load, (3) realism, (4) accuracy, (5) representation, (6) labeling of intended structure, and (7) accessibility. Definitions of these criteria and a description of each level of the rubric are found in Multimedia Appendix 1. The maximum score possible on the rubric was 28 points. To create a binary classification of useful versus not useful, we established a threshold score of 25 points. This score precludes an image receiving the lowest score of 1 point in any criterion on the rubric without receiving a score of 4 in every other criterion.

Each author rated the usefulness of the relevant images independently. Following the individual ratings, the reliability of the scoring rubric (ie, interrater reliability) was assessed using the Cronbach $\alpha$ calculation (SPSS version 24, IBM). The median of the three individual scores was established as the final usefulness score for each image. To determine if there was any difference in usefulness based on the source of the image, final usefulness scores (dependent variable) were compared across website categories (independent variable) using a Kruskal-Wallis test. An $\alpha$ level of .05 was used to determine statistical significance.

Results

Search Result Overview

Thirty-nine anatomical structures and concepts were identified and the top 10 image results were collected, yielding a total of 390 image results. The 390 results were sourced from 130 distinct websites. The sites that appeared in the results with the highest frequency are shown in Table 1.

Table 1. Most frequent sources of Google Images search results for anatomic structures.

<table>
<thead>
<tr>
<th>Website title and URL</th>
<th>Website description/tagline</th>
<th>Total number of results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kenhub (kenhub.com)</td>
<td>“Learn Anatomy Faster”</td>
<td>21</td>
</tr>
<tr>
<td>Teach Me Anatomy (teachmeanatomy.info)</td>
<td>“The Ultimate Resource for Healthcare Professionals &amp; Medical Students”</td>
<td>21</td>
</tr>
<tr>
<td>Pinterest (pinterest.com/ch)</td>
<td>“A visual discovery engine for finding ideas like recipes, home and style inspiration, and more”</td>
<td>18</td>
</tr>
<tr>
<td>Science Direct (sciencedirect.com)</td>
<td>“Elsevier’s premier platform of peer-reviewed literature”</td>
<td>17</td>
</tr>
<tr>
<td>Earth’s Lab (earthslab.com)</td>
<td>“Setting up a new place where learning becomes habit”</td>
<td>16</td>
</tr>
<tr>
<td>YouTube (youtube.com)</td>
<td>Online video-sharing and social media platform</td>
<td>14</td>
</tr>
<tr>
<td>Quizlet (quizlet.com)</td>
<td>“A free website providing learning tools for students including flashcards, study and game modes”</td>
<td>13</td>
</tr>
<tr>
<td>Springer Link (link.springer.com)</td>
<td>“Providing researchers with access to millions of scientific documents from journals, books, series, protocols, reference works and proceedings”</td>
<td>12</td>
</tr>
<tr>
<td>Get Body Smart (getbodysmart.com)</td>
<td>“A fully animated and interactive eBook about human anatomy and physiology”</td>
<td>10</td>
</tr>
</tbody>
</table>

Source of Images

Evaluation of the websites and discussion between authors resulted in the creation of six distinct categories: (1) Health Professions Education, (2) Patient/Public Education, (3) General Reference, (4) Academic Reference/Research Articles, (5) Social Media, and (6) Other. Definitions and descriptions of each category are provided in Table 2.

Agreement between authors on categorization was strong, with 82 of the 130 distinct sites (63.1%) placed in the same category by both authors during the first independent categorization. The 48 conflicts were resolved through discussion and reexamination of the websites to arrive at the final categorization.

Of the 390 image search results, 147 (37.7%) were found on Health Professions Education websites, 73 (18.7%) were found on General Reference websites, 54 (13.8%) were found on Patient/Public Education websites, 52 (13.3%) were found on Academic Reference/Research Articles websites, 50 (12.8%) were found on Social Media websites, and 14 (3.6%) were found on websites categorized as Other. The distribution of image source categories for each structure is shown in Figure 1.

Health Professions Education websites included commercial anatomy tutoring sites, as well as medical school exam study sites and specialty-specific physician resources. All but two structures (Bile Duct and Coronary Arteries) had Images search results from Health Professions Education websites.
All but one structure (Rotator Cuff Muscles) had search results from General Reference websites, although these types of websites did not comprise a majority of the search results for any of the searched structures. Wikipedia had the highest frequency of appearance of the search results (62/390 images, 15.9%). The images published on the Wikipedia entries were either public domain images (with or without modifications) from sources such as Gray’s Anatomy of the Human Body or images published under creative commons licenses.

Patient/Public Education websites included public-facing provider and clinic websites as well as general health information sites. Of the 39 search terms, 16 yielded results from these websites. This category also yielded the highest number (44) of distinct sites, with no site being repeated more than four times. Structures with a high frequency of results from these sites were found on pages relating to injury (eg, rotator cuff tears, back pain) or disease (eg, bile duct cancer, coronary artery disease).

Twenty-three of the 39 structures had Images search results from Academic Reference/Research Articles websites. Inferior Epigastric Vessels had the highest number of results from these sites (6/10), with five of these results coming from one book chapter. Images on these sites included depictions of variations, pathologic presentations, and surgical approaches (in which the searched structure may have been altered or removed).

Twenty-six of the 39 structures had Images search results on Social Media websites, although Social Media sites did not comprise a majority of the search results for any of the searched structures. Several images (n=14) were stills from YouTube videos; therefore, the site publishing the image was recorded as YouTube and categorized as Social Media. In these cases, the name, and where possible, a description of the account publishing the video were identified. Occasionally, images from Social Media sources were identical to images from other sources yielded by the search. This was likely due to image sharing to social media sites (eg, Pinterest) from original sources.

Only 9 of the 36 structures yielded Images search results from sites categorized as Other. These were primarily images available for purchase from stock image repositories as well as images found on sites whose primary purpose was to generate advertising revenue.

Table 2. Descriptions, definitions, and examples of website categories for Google Images search results.

<table>
<thead>
<tr>
<th>Category name</th>
<th>Definition/description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Health Professions Education</td>
<td>Reference material for people working in or studying the medical sciences; assumes the audience has a baseline level of specialized technical knowledge (or is studying to acquire such knowledge) about the medical sciences</td>
<td>Kenhub (kenhub.com), Radiopaedia (radiopaedia.org), Statpearls (statpearls.com)</td>
</tr>
<tr>
<td>2. Patient/Public Education</td>
<td>Accessible health-related information for patients and the lay public, typically (but not necessarily) authored by an expert, clinician, or institution</td>
<td>Mayo Clinic (mayoclinic.org), WebMD (webmd.com), American Cancer Society (cancer.org)</td>
</tr>
<tr>
<td>3. General Reference</td>
<td>Material presented as a synthesis of several sources of information tailored to a general audience</td>
<td>Wikipedia (en.wikipedia.org), Exploring Nature Science Education Resources (exploringnature.org)</td>
</tr>
<tr>
<td>4. Academic Reference/Research Articles</td>
<td>Database of peer-reviewed articles or texts, tailored to an academic audience (includes academic publisher websites)</td>
<td>Science Direct (sciencedirect.com), McGraw Hill Medical (mhmedical.com), Journal of Neurosurgery (thejns.org)</td>
</tr>
<tr>
<td>5. Social Media</td>
<td>Platform for sharing user-generated content within a community (includes blogs)</td>
<td>Pinterest (pinterest.com), YouTube (youtube.com), Karmic Seeds Body Mind &amp; Spirit (karmicseedsyogaandfitness.blogspot.com)</td>
</tr>
<tr>
<td>6. Other</td>
<td>Media that do not fit into one of the previous categories with no meaningful content (includes commercial stock image repositories)</td>
<td>Redbubble (redbubble.com), Shutterstock (shutterstock.com)</td>
</tr>
</tbody>
</table>
**Figure 1.** Distribution of source website categories for the top 10 Google Images search results for 39 anatomical structures (390 images).

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Distribution of Result Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adductor Magnus Muscle</td>
<td>5 4 3 1 1 6</td>
</tr>
<tr>
<td>Ampulla of Ductus Deferens</td>
<td>1 3 5 6 6 3</td>
</tr>
<tr>
<td>Axillary Artery</td>
<td>1 3 5 5 6 6</td>
</tr>
<tr>
<td>Bile Duct</td>
<td>5 4 3 1 1 2</td>
</tr>
<tr>
<td>Broad Ligament of the Uterus</td>
<td>3 5 6 6 6 1</td>
</tr>
<tr>
<td>Coronary Arteries</td>
<td>3 2 4 5 6 1</td>
</tr>
<tr>
<td>Deep Inguinal Ring</td>
<td>2 3 4 1 1 1</td>
</tr>
<tr>
<td>Muscles of Facial Expression</td>
<td>4 1 2 3 5 1</td>
</tr>
<tr>
<td>Femoral Triangle</td>
<td>3 5 6 6 6 1</td>
</tr>
<tr>
<td>Flexor Pollicis Brevis Muscle</td>
<td>1 2 3 5 6 6</td>
</tr>
<tr>
<td>Genitofemoral Nerve</td>
<td>3 4 5 6 6 5</td>
</tr>
<tr>
<td>Greater Splanchnic Nerve</td>
<td>1 3 4 5 6 5</td>
</tr>
<tr>
<td>Homiazygos Vein</td>
<td>1 3 4 5 6 6</td>
</tr>
<tr>
<td>Hilum of the Lungs</td>
<td>1 3 5 6 6 6</td>
</tr>
<tr>
<td>Inferior Epigastric Vessels</td>
<td>1 3 4 5 6 5</td>
</tr>
<tr>
<td>Intercostal Nerve</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Ischiocavernosus Muscle</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Left Recurrent Laryngeal Nerve</td>
<td>1 3 4 5 6 5</td>
</tr>
<tr>
<td>Lumbosacral Trunk</td>
<td>1 3 4 5 6 5</td>
</tr>
<tr>
<td>Maxillary Artery</td>
<td>1 3 4 5 6 5</td>
</tr>
<tr>
<td>Maxillary Nerve</td>
<td>1 3 4 5 6 5</td>
</tr>
<tr>
<td>Middle Colic Artery</td>
<td>1 3 4 5 6 5</td>
</tr>
<tr>
<td>Pelvic Diaphragm</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Popliteal Artery</td>
<td>1 2 3 4 6 6</td>
</tr>
<tr>
<td>Posterior Cruciate Ligament</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Pudendal Nerve</td>
<td>1 2 3 4 6 6</td>
</tr>
<tr>
<td>Quadrangular Space</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Quadratus Lumborum Muscle</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Retromandibular Vein</td>
<td>1 3 4 5 6 6</td>
</tr>
<tr>
<td>Rotator Cuff Muscles</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Stylohyoideus Muscle</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Submandibular Duct</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Superficial Layer of Posterior Forearm Muscles</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Superior Gemellus Muscle</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Superior Gluteal Artery</td>
<td>1 3 4 5 6 6</td>
</tr>
<tr>
<td>Tarsal Tunnel</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Teres Major Muscle</td>
<td>1 2 3 4 5 6</td>
</tr>
<tr>
<td>Thoracic Duct</td>
<td>1 3 4 5 6 6</td>
</tr>
<tr>
<td>Triceps Brachial Muscle</td>
<td>1 3 4 5 6 6</td>
</tr>
</tbody>
</table>

**Legend:**
1. Health Professions Education websites
2. Patient/Public Education websites
3. General Reference websites
4. Academic Reference/Research Articles websites
5. Social Media websites
6. Other websites

**Educational Value**

**Relevance**

Of the 390 images, 25 were classified as “not relevant” because they either did not depict the searched structure (24/25, 96%) or the searched structure was depicted in a nonhuman species (1/25, 4%). Two additional images were removed from the analysis because the associated text was not in English. In total, 27 images were omitted from further analysis. The structures with the most nonrelevant images were Broad Ligament of the Uterus (4/10 results excluded), Inferior Epigastric Vessels (3/10 results excluded), and Submandibular Duct (2/10 results excluded). Of the remaining searched structures, 18 structures had one nonrelevant image and 18 structures had no excluded images.
Accuracy of Images

Of the 363 relevant images, 339 were accurate (93.4%) and 24 (6.6%) contained one or more errors. These errors were classified as either misrepresentations of a structure’s morphology, location, or relations (15/24, 63%), or mislabeled structures (9/24, 38%; Figure 2). Examples of errors of misrepresentation included a retromandibular vein not dividing (morphology), an intercostal bundle between the external and internal intercostal muscles (location), and a popliteal vein lateral to the popliteal artery (relation). The structures with the most images containing errors were Pelvic Diaphragm (4 images), Retromandibular Vein (3 images), Popliteal Artery (2 images), Intercostal Nerve (2 images), Quadratus Lumborum Muscle (2 images), and Middle Colic Artery (2 images). The remaining 33 structures had either one image with errors (9/33, 27%) or zero images with errors (24/33, 73%).

Usefulness

A scoring rubric was developed to assess the usefulness of the relevant image results. Reliability among the three raters was strong when independently rating the usefulness of each image (Cronbach $\alpha=.902$, 95% CI .883-.918). When the final usefulness score for each image was derived, the median usefulness score across all 363 relevant images was 24 (range 16-28). Of the 363 relevant images, 156 images (43.0%) were deemed useful when using the binary definition of a usefulness score of 25 points or greater. The structures with the most (>60% of relevant images) useful image results were Ischiocavernosus Muscle, Axillary Artery, Muscles of Facial Expression, Maxillary Artery, Posterior Cruciate Ligament, Maxillary Nerve, Intercostal Nerve, Stylopharyngeus Muscle, Genitofemoral Nerve, Pubedal Nerve, Superior Gluteal Artery, and Submandibular Duct. The only structure with zero useful images in the top 10 results was the Lumbosacral Trunk. The distribution of the number of useful images across structures is shown in Figure 3.

There was no statistically significant difference in median usefulness score across the six image source categories ($P=.17$; Figure 4). The percentage of useful images (score of 25 points or greater) varied across each category: 46.7% (64/137) of Health Professions Education images, 32% (17/53) of Patient/Public Education images, 45% (30/67) of General Reference images, 54% (25/46) of Academic Reference/Research Article images, 36% (18/50) of Social Media images, and 20% (2/10) of images found on Other sites were useful.
Figure 3. Usefulness of the top 10 Google Images search results for 39 anatomical structures (390 images).

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Distribution of Image Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adductor Magnus Muscle</td>
<td></td>
</tr>
<tr>
<td>Ampulla of Ductus Deferens</td>
<td></td>
</tr>
<tr>
<td>Axillary Artery</td>
<td></td>
</tr>
<tr>
<td>Bile Duct</td>
<td></td>
</tr>
<tr>
<td>Broad Ligament of the Uterus</td>
<td></td>
</tr>
<tr>
<td>Coronary Arteries</td>
<td></td>
</tr>
<tr>
<td>Deep Inguinal Ring</td>
<td></td>
</tr>
<tr>
<td>Muscles of Facial Expression</td>
<td></td>
</tr>
<tr>
<td>Femoral Triangle</td>
<td></td>
</tr>
<tr>
<td>Flexor Pollicis Brevis Muscle</td>
<td></td>
</tr>
<tr>
<td>Genitofemoral Nerve</td>
<td></td>
</tr>
<tr>
<td>Greater Splanchnic Nerve</td>
<td></td>
</tr>
<tr>
<td>Hemiazygos Vein</td>
<td></td>
</tr>
<tr>
<td>Hilum of the Lungs</td>
<td></td>
</tr>
<tr>
<td>Inferior Epigastric Vessels</td>
<td></td>
</tr>
<tr>
<td>Intercostal Nerve</td>
<td></td>
</tr>
<tr>
<td>Ischiocavernosus Muscle</td>
<td></td>
</tr>
<tr>
<td>Left Recurrent Laryngeal Nerve</td>
<td></td>
</tr>
<tr>
<td>Lumbosacral Trunk</td>
<td></td>
</tr>
<tr>
<td>Maxillary Artery</td>
<td></td>
</tr>
<tr>
<td>Maxillary Nerve</td>
<td></td>
</tr>
<tr>
<td>Middle Colic Artery</td>
<td></td>
</tr>
<tr>
<td>Pelvic Diaphragm</td>
<td></td>
</tr>
<tr>
<td>Popliteal Artery</td>
<td></td>
</tr>
<tr>
<td>Posterior Cruciate Ligament</td>
<td></td>
</tr>
<tr>
<td>Pudendal Nerve</td>
<td></td>
</tr>
<tr>
<td>Quadrangular Space</td>
<td></td>
</tr>
<tr>
<td>Quadratus Lumborum Muscle</td>
<td></td>
</tr>
<tr>
<td>Retromandibular Vein</td>
<td></td>
</tr>
<tr>
<td>Rotator Cuff Muscles</td>
<td></td>
</tr>
<tr>
<td>Stylopharyngeus Muscle</td>
<td></td>
</tr>
<tr>
<td>Submandibular Duct</td>
<td></td>
</tr>
<tr>
<td>Superior Gemellus Muscle</td>
<td></td>
</tr>
<tr>
<td>Superior Gluteal Artery</td>
<td></td>
</tr>
<tr>
<td>Superficial Posterior Forearm Muscles</td>
<td></td>
</tr>
<tr>
<td>Tarsal Tunnel</td>
<td></td>
</tr>
<tr>
<td>Teres Major Muscle</td>
<td></td>
</tr>
<tr>
<td>Thoracic Duct</td>
<td></td>
</tr>
<tr>
<td>Triceps Brachii Muscle</td>
<td></td>
</tr>
</tbody>
</table>

Legend: **USEFUL**, NOT USEFUL, NOT RELEVANT.
Discussion

Source of Images

Google Images searches were performed for 39 “high-yield” anatomical structures and concepts and the top 10 results were analyzed to determine their source and educational quality.

The largest proportion of Google Images search results were published on sites whose target audience is students and health professionals (147/390, 37.7%). These sites assume that their audience has a baseline level of knowledge of human anatomy (or is in the process of acquiring such knowledge). These sites included commercially available anatomy tutoring sites. Johnson et al [23] reported that while some learners found these sites helpful in learning anatomy, students tended to prefer materials that were specifically tailored to their courses. Not all websites in this category are necessarily held to the same standard of peer review, nor was the intention of each image to aid in locating a structure in the anatomy laboratory; therefore, images from these sites cannot be generalized as useful, as will be discussed below.

Patient and Public Education sites comprised 13.9% of the Images search results (54/390). These included general information sites (eg, WebMD, MedlinePlus), public-facing sites for major medical centers (eg, Mayo Clinic), and private provider (physician or allied health professional) websites. This category also included wellness sites that did not appear to be affiliated with any provider or practice. Although images from some of these sites can be used with confidence [17], using caution with unfamiliar sources is advised. A concern with these images is that to be accessible to the lay public, they may not provide adequate detail when applied to the study of anatomy [24]. Additionally, images published on clinical websites may represent an injured or pathological structure (eg, a torn posterior cruciate ligament) or a structure that has been surgically reconstructed and no longer resembles the typical anatomical presentation.

General Reference sites comprised 18.7% of the Images search results. Wikipedia comprised the majority of sites categorized as General Reference (62/390) and was the most frequent source of Images search results overall. Wikipedia is a popular resource among medical students [9,10,25], but its reliability is debated [26]. Arguments against Wikipedia as a resource cite poor quality of images [16] and insufficient detail [15,16,27]. London et al [14] found that Wikipedia was generally accurate and complete for basic anatomical information, despite asserting that textbooks should still be considered the gold standard. Images published on Wikipedia articles provided accurate depictions of anatomical structures; however, their educational value in terms of helping students identify and locate a structure in the laboratory varied. For instance, older public domain images published on Wikipedia (such as those from Gray’s Anatomy) included outdated terminology no longer accepted by Terminologia Anatomica. Because of the open-source nature of Wikipedia, it is an educational opportunity for anatomists to edit content as well as create and publish their own content [14,28].

Academic Reference sites comprised 13.3% (52/390) of the search results. These included pages that link to journal articles and other academic literature. These sites may seem attractive as reliable sources as they are peer-reviewed; however, students browsing images from these sites should be aware that this literature may include case reports of anatomical variants (such as atypical branching of the middle colic artery or congenital absence of the retromandibular vein) and may not reflect a typical anatomical presentation.

Social Media images comprised 12.8% (50/390) of the search results. The role of social media in anatomy education has been well-examined [29]. Pinterest was the fourth-most frequent
source of Images search results overall, with 18 results coming from that platform. Pinterest allows users to collect and organize images and has been considered a potential source for clinical specialists to curate educational images [30,31]. Fourteen of the search results were stills from YouTube videos. YouTube is a popular educational resource, especially for learners who are considered “digital natives.” Although the usefulness of YouTube as an anatomy resource has been questioned [32], this platform presents an opportunity for anatomy educators to create and promote educationally useful content [33]. Finding an anatomy image on a social media site suggests that a learner may have found the image to be educationally beneficial and therefore worthy of sharing with others. However, for many of the anatomy images from social media platforms seen in this study, it was not immediately clear who shared the image or what the original source of the image was, which may (rightfully) lead learners to question the trustworthiness of these images.

Educational Value

Relevance

An image was deemed “relevant” if it depicted the search term, although appropriate labeling was assessed separately using the usefulness rubric. This parallels the concept of “visual relevancy” as described by Sedghi et al [34], which is dependent on the learner’s ability to recognize what they are looking for in an image. Of the 390 Images search results, 363 were deemed relevant. A high proportion of relevant images in the top 10 search results for an anatomical structure reduces the effort required by learners to scroll through multiple search results to find the structure they are seeking.

Accuracy of Images

The accuracy of the Images search results exceeded the authors’ expectations, with 93.4% of images containing no errors. This finding may alleviate some concerns that students are being misled or taught incorrectly from online image search results. Other studies of online medical images found similar levels of accuracy [17,35], although the authors of these studies questioned the ability of nonexperts to determine accuracy [35]. This warrants further study of the ability of learners to appraise the accuracy of online anatomical images. This accuracy assessment was brought into the usefulness rubric (described below) and was one of the seven criteria used to assess usefulness.

Usefulness

Overview

The rubric for usefulness determined that 156 of the 363 relevant images (43.0%) would be useful in helping a learner locate or identify the searched structure in an anatomy laboratory setting. There was no statistically significant association between the source of the image and the usefulness of the image.

The usefulness rubric consisted of criteria supported by evidence to be of value when selecting an image that would assist a learner in locating an anatomical structure. It is important to note that what makes an image useful is highly subjective [13] and varies from structure to structure.

Completeness

The ability to locate an anatomical structure is dependent on an understanding of its location in the body and the key relations between surrounding structures. The criterion “completeness” was included to reflect these combined needs. An image would receive a low score if there was no sense of where in the body the structure is located and if it did not show relations to neighboring structures. The ubiquity of anatomic variation prevents the identification of anatomic structures based solely on absolute terms (eg, the occipital artery is always the third branch of the external carotid artery [36]). An example of a low-rated image (1 point) for completeness was a search result for Lumbosacral Trunk: the image was a freestanding sketch of the lumbar plexus that gave no indication of where in the body the nerve was located and showed no relations to surrounding structures.

Cognitive Load

According to cognitive load theory, labeling on images beyond what is relevant and a lack of focus can result in a diversion of the learner’s attention or mental activity (ie, increased extraneous load). This is particularly problematic in the anatomy laboratory setting, which is associated with a high complexity of content and skill (ie, intrinsic load). Cognitive load theory states that learning is negatively impacted if the combination of extraneous and intrinsic load exceeds the limited working memory of the learner [37]. Additionally, it has been strongly shown that multimedia full of irrelevant information distract the viewer from the main focus [38] and may impede a learner from identifying a structure in the anatomy laboratory. By rating the amount of extraneous material and the clarity of focus, the rubric captured how well an image result would assist a learner without overburdening them. For example, an image result showing the Left Recurrent Laryngeal Nerve had an excessive number of labels (45 counted) and received a low rating (1 point). In contrast, an image showing the Broad Ligament of the Uterus was rated high (4 points) because it contained all relevant labels (uterus, broad ligament, uterine tube, ovary) and the focus of the image was extremely clear.

Realism

The extent to which an image resembled the cadaveric presentation of the searched structure was reflected in the realism criterion. The majority of the published images were illustrations and schematic diagrams, and only two images were cadaveric photographs. Schematic diagrams are typically designed to highlight key structures through simplified or abstract representation [39]. Their usefulness in the anatomy laboratory is questionable because learners may not be able to translate these simplified images to their real-life presentations on a donor [39]. For example, a search result for Deep Inguinal Ring that depicted the ring as one end of a line-drawn cylinder representing the inguinal canal may be useful as an explanation of the structure of the inguinal canal, but would not help a learner locate the deep inguinal ring on the deep surface of the abdominal wall. In contrast, a high-quality illustration or photograph would help a learner translate the image to a cadaveric specimen more easily. The images rated in this analysis were two-dimensional images; whether images that
depicted structures in three dimensions (eg, GIFs or animations that allow rotation) would improve the usefulness of an image warrants further investigation [40].

**Representation**

Learners benefit the most from viewing images of typical anatomy when they are attempting to locate structures in the body. A challenge arises for learners when viewing an atypical image as reference (eg, showing pathology, variation, or surgical reconstruction). For example, an image result showing the popliteal artery with impingement at the gastrocnemius was an atypical variation and would not assist a learner in identifying that structure in its usual location.

**Labeling**

Presenting an image of an anatomic structure to a learner with little context necessitates appropriate labeling to indicate the target structure. Despite the intuitive nature of this principle, it was recognized that some image results did not provide labeling (or other indications such as leader lines) of the searched structure; therefore, this criterion needed to be part of the rubric.

**Accessibility**

Learning materials should be inclusive and accessible to learners with a wide variety of abilities. The use of color without any other labeling to indicate structures on an anatomical image is problematic for learners with color vision deficiency [41]. Low-resolution images and images obscured by watermarks may also be visually inaccessible, or at the very least unappealing, to learners who prefer a high-quality, unobscured image [11]. An example of a low-rated (1 point) image for accessibility was a result for Maxillary Nerve that indicated the divisions of the trigeminal nerve using red and green (colors unable to be distinguished by those with protanopia, deuteranopia, and achromatopsia).

**Recommendations Based on Findings**

The following recommendations are offered to educators who work with students in an anatomy laboratory setting based on the findings of our analysis. These recommendations are not necessarily universal but can be tailored to individual curricula or educational approaches.

Students have no trouble finding and accessing online resources, but they cannot necessarily discern a good resource from a bad one [23]. The students surveyed by Johnson et al [23] expressed that they need direction from educators to find reputable online sources. Nevertheless, O’Carroll et al [3] found that medical students accessed Google as a resource with high frequency despite being instructed to choose more reputable sources such as bibliographic databases. Translating this to the anatomy laboratory environment, learners will be likely to use Google Images searches despite any attestation that atlases or other course materials are the gold standard. Thus, anatomy faculty should be prepared to advise learners on best practices for Google Images searches in the anatomy laboratory.

The number of results produced by a Google Images search requires students to be aware of how to filter them effectively [6,42]. The results of the current analysis of Google Images search results for anatomical terms could be instrumental in developing guides for students on how to select reliable images for their study. These guides could include a summary of the types of websites publishing these images with guidance on how to interpret media on these sites, as well as a list of “faculty-recommended” sources.

Students should consider that websites publishing anatomical images may have agendas beyond anatomy education. These include websites promoting controversial scientific stances (eg, the Institute for Creation Research, whose mission is to promote research within the context of biblical creation) or websites advertising commercial products (eg, Whole Life Challenge, a subscription-based wellness and lifestyle brand). While the images published on these sites may be accurate and useful, there remains opportunity to assist learners in becoming aware of these agendas when selecting images from these sites.

When advising students on internet resource use in the anatomy lab, the opportunity arises to remind students of ethical behavior in the context of choosing resources. When selecting an image, students should be aware of whether the information they are using is plagiarized (eg, lecture slides shared without permission, blogs that copy text from other published material) or is published on a site that exists primarily for the purpose of helping students cheat on exams (including social media pages that circulate an institution’s previous exam questions).

**Limitations of the Study**

We acknowledge that the selection of search terms is a subjective process, based on one’s own experience with anatomy curricula. Some websites were no longer active at the time of secondary analysis; either the domains had expired or the company publishing the site had ceased operations. In these cases, however, sufficient descriptive information about the website was available to properly categorize the site. It is also worth noting that the usefulness of an image may vary based on the dissection approach utilized in the course. The current analysis assumed a regional anatomical dissection approach, and images useful for this approach may not prove to be useful in more surgically based dissection protocols.

**Future Work**

In the future, we intend to survey anatomy students to gauge their perceptions of anatomical Images search results and to determine whether images deemed “useful” by students meet the criteria for usefulness as defined in this rubric. These results would validate this rubric and guide the development of images in educational materials. It would also be of interest to assess learners’ ability to determine the relevance and accuracy of online anatomy images (ie, will learners be able to detect inaccuracies or a lack of relevance in an image that appears to be useful at first glance?). It would also be insightful to survey students on their overall perception of the types of websites that publish images yielded in online searches. These data would further assist educators in developing best-practice guides for anatomy image searches, such as a one-page informational handout.
Conclusions
Medical and health professions students must develop information literacy skills for selecting appropriate resources early in their training. This skill development may take place in the anatomy laboratory as students search for online images to assist them in locating and identifying structures. A large number of Google Images search results were acquired for highly relevant anatomical structures and concepts. These images were reliably categorized, with a plurality sourced from Health Professions Education websites. Wikipedia articles appeared the most frequently among the images collected, which falls in line with the high traffic and public domain status of its images. A high percentage of images were determined to be accurate, with errors in representation of morphology, location, or relations being the most common. A scoring rubric was successfully developed and used to reveal that only 43.0% of images were useful for identifying a structure in a human anatomic donor. Usefulness scores did not differ significantly across image source categories. Taken together, these results illuminate the need for students to consider the source and quality of anatomic images that they access frequently. This presents an opportunity for the development and distribution of guidelines to assist students of anatomy.

Acknowledgments
The authors wish to thank Dr Donna Perry and Dr Eustathia Lela Giannaris for their assistance with the study concept. Some of these data were previously presented in abstract form at the Annual Virtual Meeting of the American Association of Clinical Anatomists on June 15-19, 2020.

Conflicts of Interest
None declared.

Multimedia Appendix 1
Scoring rubric developed to determine usefulness of each relevant Google Images search result.
[DOCX File, 21 KB - mededu_v8i4e37730_app1.docx ]

References


©Alexandra Elisabeth Wink, Amanda N Telfer, Michael A Pascoe. Originally published in JMIR Medical Education (https://mededu.jmir.org), 21.10.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Education, is properly cited. The complete bibliographic information, a link to the original publication on https://mededu.jmir.org/, as well as this copyright and license information must be included.
Abstract

Background: Trainees rely on clinical experience to learn clinical reasoning in pediatric emergency medicine (PEM). Outside of clinical experience, graduate medical education provides a handful of explicit activities focused on developing skills in clinical reasoning.

Objective: In this paper, we describe the development, use, and changing perceptions of a web-based asynchronous tool to facilitate clinical reasoning discussion for PEM providers.

Methods: We created a case-based web-based discussion tool for PEM clinicians and fellows to post and discuss cases. We examined website analytics for site use and collected user survey data over a 3-year period to assess the use and acceptability of the tool.

Results: The learning tool had more than 30,000 site visits and 172 case comments for the 55 published cases over 3 years. Self-reported engagement with the learning tool varied inversely with clinical experience in PEM. The tool was relevant to clinical practice and useful for learning PEM for most respondents. The most experienced clinicians were more likely than fellows to report posting commentary, although absolute rate of commentary was low.

Conclusions: An asynchronous method of case presentation and web-based commentary may present an acceptable way to supplement clinical experience and traditional education methods for sharing clinical reasoning.

(JMIR Med Educ 2022;8(4):e38427) doi:10.2196/38427

KEYWORDS
asynchronous learning; clinical reasoning; emergency medicine; pediatrics; web-based learning tool

Introduction

Clinical reasoning—how clinicians process and apply medical knowledge—is one way by which expert clinicians distinguish themselves from novices [1]. For novice medical trainees, in-person case-based experience, clinical context, and learning through observation are critical to developing clinical reasoning skills [2]. The pediatric emergency department (ED) can be an exceptional place to learn clinical reasoning skills. Patient volume and relative acuity in the pediatric ED provides real-world learning opportunities that complement traditional textbooks or didactics. However, the breadth of cases an individual trainee encounters in the pediatric ED can vary, resulting in inconsistent opportunities to hone clinical reasoning strategies. In addition, barriers of shift schedules and a busy ED
can limit the sharing of clinical reasoning between providers. Finally, trainees in the ED may be only briefly observed directly by faculty, suggesting the existing apprenticeship model of learning clinical reasoning may have room for improvement [3].

Asynchronous learning—in which individuals direct their own learning at their own pace, often using web-based resources—may offer advantages uniquely suited for adult learning and emergency medicine [4,5]. A web-based asynchronous learning tool was a potentially effective way to improve knowledge in pediatric emergency medicine (PEM) [6]. An asynchronous e-learning module was associated with improved knowledge in PEM among residents, and it was similar to traditional lectures in knowledge acquisition and superior to no lectures at all [7,8]. However, most asynchronous learning interventions focus on acquiring knowledge, not sharing clinical reasoning strategies. When educational interventions do address clinical reasoning, they often focus on diagnostic reasoning while neglecting therapeutic reasoning [9].

The COVID-19 pandemic abruptly changed graduate medical education. Early studies of medical training programs across several specialties report decreased in-person clinical care experiences, missed work for COVID-19 infection or exposure, and increased remote learning [10-12]. Since the development of clinical reasoning skill is traditionally tied to in-person case-based experience, asynchronous learning approaches that focus on clinical reasoning may provide unique educational value.

To develop supplemental opportunities for clinical reasoning education that incorporates both learning through interactions with others and the unique advantages of asynchronous learning for emergency medicine, we created a web-based environment for clinicians to share case-based clinical reasoning challenges [13]. In the 8 years since its inception, we have shared over 190 user-selected cases and discussions. This paper describes the development and evaluation of this tool as well as the lessons learned from this still growing asynchronous web-based PEM case series over its initial 3-year period.

**Methods**

**Procedure**

We created our learning tool—called The Hot Seat—for 3 PEM fellowship programs in Virginia, Washington DC, and Maryland. The Hot Seat presents clinical cases that focus on one or multiple diagnostic or management dilemmas requiring participants to use available information to guide decision-making during various points of a patient encounter. Cases were selected and written by PEM fellows at one of the participating programs based on a predetermined schedule. PEM fellows were advised to select cases that “raised an important diagnostic or management dilemma” and discouraged from selecting cases only because they were rare diagnoses. A brief description of the chief complaint was listed at the top of the case followed by a history and physical examination. Case presentations were modified to deidentify patients and focus on clinical challenges. Each case included several associated multiple-choice questions that intentionally had no clear right or wrong answers. The choice of case, case presentations, and associated questions aimed to frame relevant clinical reasoning dilemmas, not test specific knowledge recall. A PEM faculty advisor reviewed and edited each case and published it on a website created and customized through WordPress—a popular commercial website development tool.

Cases were published about twice per month. For each case, one PEM faculty—who was blinded to the outcomes of the case—was on the Hot Seat and tasked with explaining their clinical reasoning related to the case’s challenges, pitfalls, diagnostic pearls, disposition, or immediate management. In addition to the PEM faculty on the Hot Seat, anyone who visited the site could read the case, answer multiple-choice questions, or share their clinical reasoning strategies by posting commentary.

After a 2-week period, we published the denouement—a summary of the case discussion, including responses to multiple choice responses, clinical reasoning pearls related to the case, and the case outcome.

We developed a survey to address the specific goals of our project, specifically the acceptability and perceived utility of our novel web-based tool. We distributed a web-based survey for self-reported use and clinical relevance of the learning tool with responses on a 5-point Likert scale. We surveyed PEM practitioners at participating institutions at the end of 3 consecutive academic years (2016 to 2018) using an anonymous REDcap survey [14]. Respondents were eligible if they were members of specific institutional email lists for active PEM practitioners. Respondents self-identified their experience in PEM (fellow or PEM faculty with experience <3 years; PEM faculty with experience ≥3 years). Participants confirmed consent to participation and publication of feedback data prior to completing the survey. We counted posted comments and recorded site usage data using Google Analytics. Repeated views of a single page or multiple page views by the same user were counted as separate views.

**Ethical Considerations**

This study was approved by the Children’s National Institutional Review Board (Pro00004269).

**Results**

During the first 3 years of the Hot Seat, we created 55 unique cases that generated 172 comments from readers. The site had 31,417 page views. Page views varied by month, with a low of 317 page views (July 2017) to a high of 1664 page views (January 2016) and were highest around the time of publication of each new case. We sent survey invitations to about 70 providers each year and received a total of 65 completed surveys over the 3-year study period (Table 1).

The survey asked how often respondents used specific features of Hot Seat cases. The frequency of “always or usually” reading the Hot Seat cases was inversely associated with experience in PEM, with all PEM fellows “always or usually” reading the Hot Seat. For all levels of clinical experience, the proportion

---

https://mededu.jmir.org/2022/4/e38427

JMIR Med Educ 2022 | vol. 8 | iss. 4 | e38427 | p.144

(page number not for citation purposes)
of respondents who “always or usually” read the denouement was smaller than the proportion who reported “always or usually” reading the case or responding to the poll. Few respondents in all groups reported commenting on cases (Figure 1).

All PEM fellows “agreed or strongly agreed” that the Hot Seat is relevant to their clinical practice or provides useful insight into clinical reasoning. All but one fellow also “agreed or strongly agreed” that the Hot Seat is an effective learning tool for PEM. Faculty of all levels of clinical experience generally “agreed or strongly agreed” that the Hot Seat is relevant to clinical practice, provides useful insight into clinical reasoning, and is an effective learning tool for PEM (Figure 2).

The survey included follow-up questions to better understand perceived usefulness of the Hot Seat. Of the 62 respondents who “agreed or strongly agreed” that “the Hot Seat is relevant to clinical practice,” 45 (73%) said the Hot Seat cases were similar to cases they have encountered, 58 (94%) said the Hot Seat prompted them to think about management of similar cases, and 31 (50%) said the Hot Seat cases and commentary reflected their thought processes.

Of the 64 respondents who “agreed or strongly agreed” that the Hot Seat provides “useful insight into clinical reasoning,” 55 (86%) said it was helpful to see how others approach the cases by reading the comments; 48 (75%) said the multiple-choice questions were helpful; and 46 (72%) said the cases helped them think about how they would manage similar cases in the future.

Table 1. Summary of survey respondents as well as Hot Seat cases and use by academic year.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>2015-2016 (n=15)</th>
<th>2016-2017 (n=22)</th>
<th>2017-2018 (n=28)</th>
<th>Total (n=65)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical experience of respondents, n (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fellow</td>
<td>5 (33)</td>
<td>3 (17)</td>
<td>5 (18)</td>
<td>13 (20)</td>
</tr>
<tr>
<td>Faculty &lt;3 years</td>
<td>3 (20)</td>
<td>6 (27)</td>
<td>9 (32)</td>
<td>18 (28)</td>
</tr>
<tr>
<td>Faculty ≥3 years</td>
<td>7 (47)</td>
<td>13 (59)</td>
<td>14 (50)</td>
<td>34 (52)</td>
</tr>
<tr>
<td>Hot seat cases, n (%)</td>
<td>19 (34)</td>
<td>18 (33)</td>
<td>18 (33)</td>
<td>55 (100)</td>
</tr>
<tr>
<td>Case comments, n (%)</td>
<td>57 (33)</td>
<td>55 (32)</td>
<td>60 (35)</td>
<td>172 (100)</td>
</tr>
<tr>
<td>Total page views, n (%)</td>
<td>11,632 (37)</td>
<td>9436 (30)</td>
<td>10,349 (32)</td>
<td>31,417 (100)</td>
</tr>
<tr>
<td>Time on page (seconds), mean (SD)</td>
<td>171</td>
<td>176</td>
<td>148</td>
<td>N/A*</td>
</tr>
</tbody>
</table>

N/A: not applicable.

Figure 1. Frequency of engagement with Hot Seat by clinical experience.
Discussion

Principal Findings

Interest and engagement in web-based medical educational resources have grown in recent years [15,16]. In comparison to many available web-based tools, the Hot Seat uniquely focuses on case discussion and clinical reasoning—important areas of medical education for which designing teaching initiatives can be challenging. In our experience, we found an engaged target audience with a range of clinical experience who reported useful insights into clinical reasoning and value in reading commentary, yet a reluctance to personally write commentary.

The proportion of respondents who engaged in the Hot Seat was high in all groups—with the PEM fellows reporting the highest share of engagement. Engaged PEM fellows were expected, since they create the cases and are presumably invested in their learning as trainees. Weekly PEM fellows’ conferences included designated time to discuss the current Hot Seat case, adding more impetus for fellows to engage. However, even among the senior clinicians, few respondents reported “rarely or never” engaging, suggesting that a web-based tool may be acceptable to a broader range of experience levels rather than just trainees.

Most respondents found value in reading the case comments and seeing how others would approach the cases. Despite the reported value of reading comments, most respondents rarely posted comments themselves. Reading, but not posting is a common social media behavior and consistent across learning platforms, where most people consume content, and a small proportion creates the majority of the content [17].

Most respondents found that comments are useful for learning, yet only about half of respondents said that the comments reflected their own thought processes. Sharing disagreements in reasoning is a feature of the Hot Seat that is distinct from many asynchronous educational approaches. Strategies to increase the sharing of reasoning and promote discussions among users may be an area of focus for future projects.

Although a small number of respondents posted comments to cases, the most experienced PEM faculty represented the group with the largest proportion of “usually or always” commenting on the cases. Experienced faculty contributing a high proportion of content is consistent with prior data on social media use in medical education [18]. Since experienced clinicians play an important role in sharing experience and clinical reasoning, future work should find ways to amplify engagement of these experienced clinical voices.

Limitations

Accessing a case discussion is not equivalent to learning clinical reasoning. Therefore, although our site usage data provide a broad picture of readership, it does not necessarily reflect educational engagement. Our survey questions sought to address this limitation, but relatively low response rates among faculty and the self-reported nature of survey results limited our conclusions on how individuals use the Hot Seat. We created our survey to explore the goals of our study and used common questionnaire development practices. However, we did not validate the survey, which may limit the interpretation of survey results. Response bias may skew survey results toward the positive and may not be applicable to a broader audience. Finally, commentary analysis was quantitative rather than qualitative. A more complete understanding of comment quality and relevance may be a useful next step.

Practical Lessons Learned

The Hot Seat’s blog format has advantages while also presenting challenges. The biggest advantage is the relatively low barrier to entry. WordPress has an accessible drag and drop interface, requires minimal prior coding experience, has a robust user community and associated support forums, and generally creates an affordable website format familiar to most people. On the other hand, a blog is not an ideal format to simulate a multilayered case discussion. Formatting and posting cases as well as creating the denouement require frequent, active inputs by a centralized group of people. Case commentary is typically individual statements rather than active discussions. Analytics are basic and cannot associate learning behavior to individuals.
making educational assessments challenging. A future platform for clinical reasoning education should be customizable and modern. Features might include automations for creating and posting the cases, an interface conducive to discussions, and reliable learner analytics.

**Conclusions**

Since we reviewed the initial 55 cases and learner data for this study, we have published over 140 additional cases. Case authors, learners, and cases have changed over time, making traditional pre- and postintervention comparisons challenging. Yet the longevity, variety, and evolving nature of our project demonstrates that clinical reasoning scenarios continually present themselves, and sharing them via an asynchronous web-based site may be an acceptable and useful approach to facilitating a clinical reasoning discussion.

**Conflicts of Interest**

None declared.

**References**


Abbreviations

**ED:** emergency department  
**PEM:** pediatric emergency medicine
The Potential and Concerns of Using AI in Scientific Research: ChatGPT Performance Evaluation

Zuheir N Khlaif1*, PhD; Allam Mousa2*, PhD; Muayad Kamal Hattab3*, PhD; Jamil Itmazi4*, PhD; Amjad A Hassan3*, PhD; Mageswaran Sanmugam5*, PhD; Abedalkarim Ayyoub1*, PhD

1Faculty of Humanities and Educational Sciences, An-Najah National University, Nablus, Occupied Palestinian Territory
2Artificial Intelligence and Virtual Reality Research Center, Department of Electrical and Computer Engineering, An Najah National University, Nablus, Occupied Palestinian Territory
3Faculty of Law and Political Sciences, An-Najah National University, Nablus, Occupied Palestinian Territory
4Department of Information Technology, College of Engineering and Information Technology, Palestine Ahliya University, Bethlahem, Occupied Palestinian Territory
5Centre for Instructional Technology and Multimedia, Universiti Sains Malaysia, Penang, Malaysia
*all authors contributed equally

Abstract

Background: Artificial intelligence (AI) has many applications in various aspects of our daily life, including health, criminal, education, civil, business, and liability law. One aspect of AI that has gained significant attention is natural language processing (NLP), which refers to the ability of computers to understand and generate human language.

Objective: This study aims to examine the potential for, and concerns of, using AI in scientific research. For this purpose, high-impact research articles were generated by analyzing the quality of reports generated by ChatGPT and assessing the application’s impact on the research framework, data analysis, and the literature review. The study also explored concerns around ownership and the integrity of research when using AI-generated text.

Methods: A total of 4 articles were generated using ChatGPT, and thereafter evaluated by 23 reviewers. The researchers developed an evaluation form to assess the quality of the articles generated. Additionally, 50 abstracts were generated using ChatGPT and their quality was evaluated. The data were subjected to ANOVA and thematic analysis to analyze the qualitative data provided by the reviewers.

Results: When using detailed prompts and providing the context of the study, ChatGPT would generate high-quality research that could be published in high-impact journals. However, ChatGPT had a minor impact on developing the research framework and data analysis. The primary area needing improvement was the development of the literature review. Moreover, reviewers expressed concerns around ownership and the integrity of the research when using AI-generated text. Nonetheless, ChatGPT has a strong potential to increase human productivity in research and can be used in academic writing.

Conclusions: AI-generated text has the potential to improve the quality of high-impact research articles. The findings of this study suggest that decision makers and researchers should focus more on the methodology part of the research, which includes research design, developing research tools, and analyzing data in depth, to draw strong theoretical and practical implications, thereby establishing a revolution in scientific research in the era of AI. The practical implications of this study can be used in different fields such as medical education to deliver materials to develop the basic competencies for both medicine students and faculty members.
Introduction

Background

Artificial intelligence (AI) has many applications in various aspects of our daily life, including health, criminal, education, civil, business, and liability law [1,2]. One aspect of AI that has gained significant attention is natural language processing; this refers to the ability of computers to understand and generate human language [3]. As a result, AI has the potential to revolutionize academic research in different aspects of research development by enabling the analysis and interpretation of vast amounts of data, creating simulations and scenarios, clearly delivering findings, assisting in academic writing, and undertaking peer review during the publication stage [4,5]. ChatGPT [6], one of the applications of AI, is a variant of the GPT language model developed by OpenAI and is a tool designed to generate human-like text in a conversational style that can engage in conversations on various topics. Trained on human-human conversation data, ChatGPT can generate appropriate responses to questions and complete discussions on its own, making it a valuable tool for natural language processing research. As a language model developed by OpenAI, ChatGPT has been widely used in various fields, such as language translation, chatbots, and natural language processing [7]. ChatGPT has many applications in multiple domains, including psychology, sociology, and education; additionally, it helps to automate some of the manual and time-consuming processes involved in research [8]. Furthermore, its language-generation capabilities make it a valuable tool for natural language processing tasks, such as summarizing complex scientific concepts and generating scientific reports [9]. The features of ChatGPT make it an attractive tool for researchers whose aim is to streamline their workflow, increase efficiency, and achieve more accurate results.

Research Gap

Many studies, preprints, blogs, and YouTube (YouTube, LLC/Google LLC) videos have reported multiple benefits of using ChatGPT in higher education, academic writing, technical writing, and medical reports [10]. However, many blogs have raised concerns about using ChatGPT in academic writing and research. Moreover, some articles consider ChatGPT to be an author and have listed it as a coauthor; this has raised many questions regarding research integrity, authorship, and the identity of the owners of a particular article [11]. What has been written regarding the issue of using ChatGPT as a research generator has been limited to opinions and discussions among researchers, editors, and reviewers. Various publishers have organized these discussions to explore possible agreements and the development of ethical policies regarding the use of ChatGPT in academic writing and research. However, there are limited practical studies on using ChatGPT in scientific research. There is a gap in understanding the potential, limitations, and concerns of using ChatGPT in scientific research as well as the ethical and social implications of incorporating AI in scientific work.

Additionally, there is a lack of standardized methods and best practices for using ChatGPT in scientific research. These gaps highlight the need for further research to investigate the effectiveness, accuracy, and trustworthiness of ChatGPT when used for scientific research, as well as to identify the ethical and societal implications of using AI in scientific inquiry. It is important to investigate the concerns of using ChatGPT in academic research to ensure safe practices and consider the required ethics of scientific research.

Purpose of the Study

This paper aims to examine the role of ChatGPT in enhancing academic performance in scientific research in social sciences and educational technology research. It also seeks to provide insights and guidance for researchers in different fields such as medical sciences, specifically research in medical education.

Contribution of the Study

Conducting scientific research regarding the potential for ChatGPT to be used in scientific research could provide researchers with valuable insights into the capabilities, benefits, and limitations of using AI in research. The findings of this research are expected to help identify the potential bias and ethical considerations associated with using AI to inform future developments and assess the implications of using AI in scientific research on different topics and multidisciplinary research fields such as medical education. Moreover, by exploring the use of ChatGPT in specific fields, such as social sciences and educational technology, researchers can gain a deeper understanding of how AI can enhance academic performance and support advancing knowledge in various education fields such as engineering and medical education. In the context of this study, the researchers consider ChatGPT to be an e-research assistant, or a helpful tool, for researchers to accelerate the productivity of research in their specific areas. Therefore, this study attempts to answer the following research questions:

- What are the best practices for using AI-generated text, specifically ChatGPT, in scientific research?
- What are the concerns about using AI-generated text, specifically ChatGPT, in academic research?

AI-Generated Text Revolutionizing Medical Education and Research

AI-generated text, powered by AI technologies such as ChatGPT, has revolutionized medical education and research. It offers unique opportunities to enhance learning experiences and provide medical professionals with up-to-date knowledge [12]. Its integration into medical education brings several advantages, such as real-time access to vast medical information, continuous learning, and evidence-based decision-making [13]. AI-generated text bridges the knowledge gap by providing...
accurate and current information from reputable sources, facilitating access to relevant medical literature, research studies, and clinical guidelines [14]. This personalized learning tool fosters critical thinking and self-directed exploration of medical concepts, while also offering instant feedback and adaptive learning experiences [15].

In medical research, ChatGPT plays a crucial role by assisting researchers in gathering and analyzing extensive medical literature, saving valuable time and effort [12,13]. It fosters collaboration among researchers and aids in data analysis, uncovering patterns and relationships within data sets [14]. Moreover, ChatGPT facilitates the dissemination of research findings by creating accessible summaries and explanations of complex work [15]. This promotes effective communication between research and clinical practice, ensuring evidence-based health care practices.

Despite its benefits, AI-generated text presents challenges in medical education and research. Ensuring the reliability and accuracy of information is essential, considering the potential for incorrect or misleading data [12]. Validating sources and aligning content with medical standards and guidelines are crucial steps. Additionally, AI-generated text may lack human interaction, which is vital for developing communication and empathy skills in medical practice [13]. To strike a balance, medical education should combine AI-generated text with traditional teaching methods, emphasizing direct patient interaction and mentorship [14].

Researchers must exercise caution when relying on AI-generated text, critically evaluating the information provided [15]. Human expertise and judgment remain indispensable to ensure the validity and ethical considerations of research findings [16]. Thoughtful integration of AI-generated text in medical education and research is essential to harness its potential while preserving the essential human touch required in medical practice [17].

In conclusion, AI-generated text has transformed medical education and research by offering accessible and up-to-date knowledge to medical professionals [12]. It empowers learners to engage in self-directed exploration and critical thinking, while also providing personalized feedback for improvement [15]. In research, ChatGPT aids in data analysis, communication, and dissemination of findings, bridging the gap between research and practice [15]. However, careful consideration of its reliability and integration with human expertise is crucial in both medical education and research settings [12]. By embracing AI-generated text thoughtfully, the medical field can leverage its potential to drive innovation and advance evidence-based health care practices [16].

**Exploring the Application of ChatGPT in Scientific Research**

Researchers can use ChatGPT in various ways to advance research in many fields. One of the applications of ChatGPT in scientific research is to provide researchers with instructions about how to conduct research and scientific research ethics [18].

Researchers can ask the application to provide a literature review in a sequenced way [19]. ChatGPT can organize information into tables based on the prompts used by researchers and the flow of the research stages [20]. Moreover, researchers can utilize ChatGPT’s ability to summarize data and write reports based on detailed data; the application also makes it easier for researchers and analysts to understand and communicate their findings. Practitioners and researchers have already been using language models such as ChatGPT to write, summarize published articles, talk, improve manuscripts, identify research gaps, and write suggested research questions [4]. Moreover, practitioners use AI-generated text tools to generate examination questions in various fields, while students use them to write computer code in competitions [18]. AI will soon be able to design experiments, write complete articles, conduct peer reviews, and support editorial offices accepting or rejecting manuscripts [21]. However, researchers have raised concerns about using ChatGPT in research because it could compromise the research integrity and cause significant consequences for the research community and individual researchers [22].

Although the development of ChatGPT has raised concerns, it has provided researchers with opportunities to write and publish research in various fields. Researchers can learn how to begin academic research; this is especially relevant to novice researchers and graduate students in higher education institutions. Various reports have been released regarding using ChatGPT when writing student essays, assignments, and medical information for patients [7]. However, many tools have also been released to look for and identify writing undertaken by ChatGPT [23,24].

ChatGPT has limitations when writing different stages of academic research; for example, the limitation of integrating real data in the generated writing, its tendency to fabricate full citations, and the fabrication of knowledge and information relating to the topic under investigation [7].

**Understanding the Legal Landscape of ChatGPT**

It was stated above that ChatGPT is an AI language model developed and owned by OpenAI, which can be used in higher education, academic writing, and research [10,18]. However, it is noticeable that the ChatGPT model does not currently include any terms and conditions, or a fair use policy per se, published on its system or website. Meanwhile, it is important to note that this situation may change in the future, as the model’s owner may, at any time, apply or enforce their policies, terms, and conditions, or membership and charges as they see fit. Nevertheless, the absence of any regulatory terms or policies on using ChatGPT should not subsequently mean that there are no other legal or ethical regulations that researchers must consider and follow when using ChatGPT services. Rules concerning data protection and intellectual property rights are equally relevant to protecting both the rights of the owner of the ChatGPT system and the intellectual property rights of authors that ChatGPT has sought and from whose work it has generated its information [25].

As to the rules concerning the property rights of the owner of the ChatGPT system, it is a criminal offense for anyone to engage in any harmful activity, misuse, damage, or cyberattack on the system and its operation [26,27]. Cybercrimes and copyright infringements may refer to any activity that is
considered illegal under domestic or international criminal law [28,29].

For example, using malicious tactics to cause damage to ChatGPT systems and user-mode applications; engaging in data theft, file removal or deletion, and digital surveillance; or attempting to gain external remote control over the ChatGPT system can be considered illegal acts that may result in criminal charges under national or international law [28,30].

Furthermore, it should be noted that cybercrimes and copyright infringements carry potential criminal consequences and civil liabilities [31]. This means that the owner of the ChatGPT system and any other third party affected by such acts have the right to seek remedies, including compensation, under the Law of Tort [32]. These rights are implied under national and international law and do not need to be explicitly stated on the ChatGPT website or within its system [33,34].

As to the second point relating to respecting intellectual property rights, while ChatGPT and its owners are generally not responsible for how a person uses the information provided by the system, it is the user’s liability if any information obtained from the system in any way constitutes a breach of national law, or could lead to a criminal conviction. This could include, for example, the commission of fraud, cyberbullying, harassment, or any other activity that violates an individual country’s applicable laws or regulations [27,30]. Suppose a person or group of people have published misleading or deceptive information. In that case, they may be at risk of being charged with a criminal offense [35], even though such information is gathered from the ChatGPT model. Accordingly, it remains the sole responsibility of every individual using the operation of ChatGPT to ensure that the information gathered or provided by the system is accurate, complete, and reliable. This means that it remains the sole responsibility of every individual to ensure that any information published or provided to others is done so in accordance with the applicable national or international law, including that related to intellectual property rights, data protection, and privacy of information [35,36].

Concerning academic writing and research, it is also important to note that researchers are responsible for and expected to follow ethical and professional standards when conducting, reporting, producing academic writing, or publishing their research [25,37]. Accordingly, if a researcher relies on ChatGPT in whole or in part for scientific research, the attribution of the information gathered would depend on the specific context and circumstances of the research. For example, research cannot provide scientific information discovered by others without appropriate reference to the original research [38]. The researcher cannot claim intellectual property rights or provide misleading information that may infringe on the proprietary rights of others [36,39].

Intellectual property rights are governed by national and international law and practice. Researchers using ChatGPT as a source in their manuscript writing are familiar with all ethical and legal policies and internal and international laws regulating their work, including copyright protection, privacy, confidentiality protection, and personal property protection. Therefore, users of ChatGPT should not rely on the service to engage in any activity that infringes upon the intellectual property rights of others, including but not limited to copyright, trademark, or privacy infringement [30]. Nevertheless, where data or published research is concerned, protecting intellectual property rights is not limited to one individual copyright but usually involves protecting the rights and interests of all members connected to the data and published research [40]. This includes but is not limited to educational institutions, government institutions or authorities, private sectors, and funding institutions [41]. This means that all parties involved in data collection, research publication, and other published information from which ChatGPT has gathered its information must be cited, attributed, and acknowledged in the manuscript submitted for publication. This is an ethical requirement and involves copyright and intellectual property rights [39].

In short, when using ChatGPT, researchers must be mindful of the legal and regulatory requirements related to their use of the service, including those relating to intellectual property rights, data protection, and privacy [42]. When using ChatGPT for research purposes, researchers should pay serious attention to the potential ethical implications of their research and take steps to ensure that their use of the service is responsible and in compliance with relevant standards and guidelines [43]. Ultimately, the responsibility for conducting research lies with researchers; they should follow best practices and scientific ethics guidelines in all research phases.

Based on a conversation between researchers participating in the study and ChatGPT, the application claims that it has revolutionized academic research writing and publishing within a short time compared with traditional writing. Gao et al. [44] examined the differences between the writing generated by human and AI-generated text, such as ChatGPT, and the style of writing. The findings of their study revealed that the AI detector used accurately identified the abstracts that ChatGPT wrote. The researchers then checked for plagiarism which was found to be 0%.

**Methods**

**Overview**

This study aimed to examine the potential for, and concerns of, using ChatGPT to generate original manuscripts that would be accepted for publication by journals indexed in Web of Science and Scopus. We used ChatGPT to generate 4 versions of a full manuscript in the field of educational technology, specifically technostress and continuance intention regarding the use of a new technology. The faked research aimed to identify the relationship between the factors influencing teachers’ technostress and continuance intention to use a new technology continually. Moreover, using ChatGPT, we generated more than 50 abstracts for articles in the fields of social sciences and educational technology. These articles were previously published in journals (with an impact factor exceeding 2) indexed in Scopus Q1 and Q2. The researchers developed the prompts through long conversations with the model. For example, the first prompt was simple and asked for general information about the research topics. Then, the researchers developed the initial
prompt based on the responses of the model by requesting more information, models, adding connector words, citations, etc.

**Description of the Full Articles**

The generated article was composed of the main sections typically found in published research in journals indexed in Scopus (Q1 and Q2) and Web of Science (having an impact factor or Emerging Sources Citation Index). These sections included an Introduction (including the background to the study, the research problem, the purpose and contribution of the study, and research questions); a Literature Review (including the framework of the study); the Methodology (including the research design, tools, and data collection); Data Analysis (including suggested tables to be included in the study); and a Citation and References List. To improve the outcomes, we iterated the initial writing of these sections on 4 occasions. To obtain more detailed responses from ChatGPT, we began by providing simple prompts that gradually became more detailed. In all of the prompts, we asked ChatGPT to add citations within the text and to include a references list at the end of each section. These articles were named version 1, version 2, version 3, and version 4, respectively.

**Generated Abstracts**

To generate the abstracts from published articles, we provided ChatGPT with a reference and asked it to generate an abstract from the article; this abstract was composed of no more than 200 words. In the prompt, we requested it to include the purpose of the study, the methodology, the participants, data analysis, the main findings, any limitations, future research, and contributions. We chose these terms based on the criteria for writing an abstract that would be suitable for publication in an academic journal. For the generated abstracts, we asked ChatGPT to provide us with abstracts from various fields of social sciences and educational technology. The criteria for writing such an abstract were that it should be suitable for publishing in an academic journal that is specified for the topic of the abstract and that it was composed of 200 words.

**Development of the Research Tool**

We developed an evaluation form based on the review forms used in some journals indexed in Scopus and Web of Science. The purpose of the form was to guide the reviewers to review the full articles generated by ChatGPT. We submitted the form to potential reviewers who worked on behalf of some of the journals, to validate and ensure the content of the items in the form was good enough. We asked them to provide feedback by editing, adding, and writing their comments on the form. Some reviewers requested to add a new column to write notes, while others asked to separate the introduction into research ideas and research problems and to include the overall quality of the research questions. After developing the form, a pilot review was conducted with 5 reviewers regarding actual studies written by humans. The final version of the evaluation form is available in Multimedia Appendix 1.

**Focus Group Session**

An online discussion group lasting 1 hour was conducted to discuss the quality of the abstracts and articles that were reviewed. An invitation was sent to 23 reviewers to attend the discussion session. Out of these 23 reviewers, 20 attended the online session. The discussion focused on how the reviewers judged the quality of the abstracts and articles, the content and sequence of ideas, and the writing style; 2 researchers moderated the discussion in the focus group session.

**Potential Reviewers**

We intended to recruit 50 reviewers from different fields of social sciences and educational technology to review the abstracts and the 4 generated articles. We used the snowball technique to find potential reviewers to revise the abstracts and the 4 generated papers. We recruited 23 reviewers, all of whom held a PhD in different fields. They were from different countries and had published research in international journals. All of the reviewers had a similar level of experience in reviewing material for high-ranking journals.

All the reviewers were anonymous, and the review process followed a single-blind peer-review model. A total of 3 reviewers assessed each abstract individually, while 7 reviewers evaluated each of the 4 articles generated by ChatGPT using a blind peer-review process. The researchers requested that the reviewers give verbal feedback on the overall quality of the written articles, and all the reviewers submitted their reports to the first author.

In the beginning, the researchers did not inform the reviewers that ChatGPT had generated the content they would be reviewing. However, at the beginning of the focus group session, the moderator informed the researchers that the content they reviewed had been generated using the ChatGPT platform. All the reviewers were informed that their identities would remain anonymous.

**Data Analysis**

The researchers analyzed the reviewers’ responses on the form using statistical analysis to identify the mean score of each item of the 4 versions of the research. They thereafter compared the results of the 4 articles to find the one-way ANOVA [45]. Moreover, the researchers analyzed the qualitative data comprising notes written in the note section on the evaluation form and the data obtained from the focus group session using thematic analysis [46]. The purpose of the qualitative data was to gain insights and a deeper understanding of the quality of the articles and abstracts from the reviewers’ perspective. The researchers used thematic analysis to analyze the qualitative data from the focus group session.

**Qualitative Data Analysis Procedures**

The researchers recorded the focus group discussion session and one of the researchers took notes during the discussion. The audio file of the recorded session was transcribed. The researchers sent the text file to the participants to change, edit, or add new information. After 1 week, the researchers received the file without any changes. The unit analysis was a concept or idea related to the research questions; 2 researchers independently analyzed the qualitative data. After completing the analysis, the raters exchanged the data analysis files and found an intrarater reliability of 89%. Any discrepancy between
the coders, and the researchers, was resolved by negotiation to achieve agreement.

**Ethical Consideration**

The researchers received approval to conduct this study from the Deanship of Scientific Research at the An Najah National University in Palestine (approval number ANNU-T010-2023). A consent form was obtained from participants in the focus group session and from the reviewers to use their records for academic research. Therefore, the statement was “Do you agree or not? If you agree please sign at the end of the form.” All the participants were informed that participation in the reviewing process and discussion in the focus group were voluntary and free without any compensation. Moreover, we informed them that their identity will be anonymous. At the end of the paragraph, the following sentence was added: “If you agree, we consider you signed the form, if not you can stop your participation in the study.”

**Results**

**Reviewers’ Assessment of Research Quality and References in the 4 Study Versions**

Based on the statistical data analysis performed by estimating the mean and SDs of the reviewers’ responses, as well as ANOVA [45], both Tables 1 and 2 represent the analysis results of the reviewers’ reports on the 4 study versions. All research stages were evaluated by the reviewers based on criteria developed by the researchers through the study of reviewing processes in high-impact journals such as Nature, Science, and Elsevier. The scale used to evaluate each stage of the research was as follows: 1=strongly disagree and 10=strongly agree. The midpoint of the scale represents the level at which a paper would be accepted for publication and was 5.5 in this study. Based on the findings presented in Tables 1 and 2, we found that the overall average quality of the research ranged from 5.13 to 7.08 for version 1 to version 4, respectively. Therefore, based on the midpoint criteria (5.5), not all of the papers would have been successful in being selected for publication. For example, version 2 scored less than the midpoint. Based on these findings, the weakest part of the developing studies, with less improvement in the 4 versions, was the cited references list. The value ranged from 4.74 to 5.61, leading to only 1 version (ie, version 4) of the generated studies being eligible for acceptance based on the references. However, upon checking whether the references listed were available, we found that only 8% of the references were available on Google Scholar (Google LLC/Alphabet Inc.)/Mendeley (Mendeley Ltd., Elsevier).

In addition, we found that the development of the prompts did not improve the quality of the research idea; for example, in version 3 (Table 1), the quality of the research idea was less than the quality of the idea in version 2. The major improvement could be seen in the writing of the literature review. We noticed an improvement between versions 1 and 4. The range was from 5.88 (version 1) to 7.32 (version 4).

According to data displayed in Table 2, there were differences in reviewing the 4 versions of the generated study (P<.02). The results differ because, according to the posttest (least significant difference), versions 3 and 4 exhibited greater significance (P<.001) compared with version 1. Therefore, there was no significant difference (P<.001) between versions 3 and 4, and both were superior to version 1. However, version 2 did not differ significantly (P<.001) from the other 3 versions. This result was due to the quality of the prompts used in the first version. The researchers used a simple prompt without any directions to write the abstract. Therefore, the reasonable difference between versions 3 and 4 was due to the difference in the stated prompts.

Moreover, there were significant differences (P<.001) in the research phases between the 4 versions, which were related to the development of the prompts used by the researchers to request responses from ChatGPT. An interesting finding in the citation and references list was that both versions 3 and 4 showed no differences in the development of in-text citations and the references list. Although the researchers used detailed prompts to train ChatGPT to improve in-text citations and references, there was no significant (P<.001) development. Hence, the response was simple—containing neither references nor connecting words as shown in Figure 1. We developed the prompt accordingly and managed to obtain a more professional response as illustrated in Figures 2 and 3. The whole idea is illustrated in Multimedia Appendix 2.

ChatGPT’s performance in version 4 is notable, not only in the overall quality but also in most research stages. However, in the focus group discussion, despite the improvement in the quality of research based on the enhancement of the prompts and the use of more context and detail, the reviewers mentioned that the quality of writing, especially the consequences and the use of conjunctions between ideas, was lacking and needed improvement. They mentioned that it was easy for the reviewers who had experience in reviewing articles to identify that the writing was accomplished using a machine rather than a human.

The reviewers in this study concluded that if journal reviewers have experience, they will realize that an AI tool, such as ChatGPT, has written the manuscript they are reviewing.
<table>
<thead>
<tr>
<th>Description</th>
<th>Mean (SD)</th>
<th>Version 1</th>
<th>Version 2</th>
<th>Version 3</th>
<th>Version 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.04 (0.71)</td>
<td>7.26 (0.62)</td>
<td>7.48 (0.51)</td>
<td>7.57 (0.51)</td>
<td>7.34 (0.62)</td>
</tr>
<tr>
<td>Research idea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.30 (0.35)</td>
<td>6.51 (0.40)</td>
<td>6.47 (0.38)</td>
<td>6.85 (0.33)</td>
<td>6.53 (0.41)</td>
</tr>
<tr>
<td>Literature review</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.88 (0.36)</td>
<td>6.43 (0.61)</td>
<td>6.80 (0.57)</td>
<td>7.32 (0.37)</td>
<td>6.61 (0.71)</td>
</tr>
<tr>
<td>Methodology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.59 (0.49)</td>
<td>6.02 (0.51)</td>
<td>6.61 (0.54)</td>
<td>6.93 (0.53)</td>
<td>6.29 (0.73)</td>
</tr>
<tr>
<td>Citation and references</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.74 (0.54)</td>
<td>5.00 (0.50)</td>
<td>5.41 (0.54)</td>
<td>5.61 (0.69)</td>
<td>5.19 (0.66)</td>
</tr>
<tr>
<td>Plagiarism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.67 (0.36)</td>
<td>6.78 (0.36)</td>
<td>7.46 (0.30)</td>
<td>7.83 (0.42)</td>
<td>7.18 (0.60)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.01 (0.12)</td>
<td>6.32 (0.28)</td>
<td>6.61 (0.23)</td>
<td>6.97 (0.23)</td>
<td>6.48 (0.42)</td>
</tr>
<tr>
<td>Research components and source</td>
<td>Sum of squares</td>
<td>Mean of squares</td>
<td>df</td>
<td>F, 3,88</td>
<td>P value</td>
<td>Least significant difference</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>----</td>
<td>---------</td>
<td>---------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>Abstract</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BGa</td>
<td>3.77</td>
<td>1.26</td>
<td>3</td>
<td>3.59</td>
<td>.02</td>
<td>Version 3 and version 4=version 1</td>
</tr>
<tr>
<td>WGb</td>
<td>30.78</td>
<td>0.35</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>34.55</td>
<td>N/A</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research idea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG</td>
<td>3.65</td>
<td>1.22</td>
<td>3</td>
<td>9.06</td>
<td>&lt;.001</td>
<td>Version 4&gt;version 1, version 2, and version 3</td>
</tr>
<tr>
<td>WG</td>
<td>11.80</td>
<td>0.13</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15.45</td>
<td>N/A</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Literature review</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG</td>
<td>25.19</td>
<td>8.40</td>
<td>3</td>
<td>35.27</td>
<td>&lt;.001</td>
<td>Version 1&lt;version 2&lt;version 3&lt;version 4</td>
</tr>
<tr>
<td>WG</td>
<td>20.95</td>
<td>0.24</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>46.14</td>
<td>N/A</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Methodology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG</td>
<td>24.92</td>
<td>8.31</td>
<td>3</td>
<td>30.85</td>
<td>&lt;.001</td>
<td>Version 1&lt;version 2&lt;version 3&lt;version 4</td>
</tr>
<tr>
<td>WG</td>
<td>23.70</td>
<td>0.27</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>48.62</td>
<td>N/A</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citation and references</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG</td>
<td>10.68</td>
<td>3.56</td>
<td>3</td>
<td>10.90</td>
<td>&lt;.001</td>
<td>Version 1&lt;version 2&lt;version 3=version 4</td>
</tr>
<tr>
<td>WG</td>
<td>28.74</td>
<td>0.33</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>39.42</td>
<td>N/A</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plagiarism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BG</td>
<td>20.88</td>
<td>6.96</td>
<td>3</td>
<td>53.36</td>
<td>&lt;.001</td>
<td>(Version 1 and version 2)&lt;version 3=version 4</td>
</tr>
<tr>
<td>WG</td>
<td>11.48</td>
<td>0.13</td>
<td>88</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>32.36</td>
<td>N/A</td>
<td>91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>76.90</td>
<td>N/A</td>
<td>91</td>
<td></td>
<td>&lt;.001</td>
<td>Version 1&lt;version 2&lt;version 3&lt;version 4</td>
</tr>
</tbody>
</table>

*aBG: between groups.

*bWG: within the group.

*cN/A: not applicable.
Figure 1. The simple question and ChatGPT’s response.

Write 100 words about teachers technostress

Teachers technostress is a term used to describe the stress and anxiety that teachers experience as a result of integrating technology into their teaching practice. As schools and classrooms become more technologically advanced, teachers are expected to keep up with new technologies and incorporate them into their lesson plans. This can be a daunting task for many teachers, especially those who are not comfortable with technology. They may feel overwhelmed, frustrated, and anxious about their ability to use technology effectively in the classroom. Teachers technostress can have negative impacts on job satisfaction, teacher burnout, and student learning outcomes. To mitigate technostress, teachers need support, training, and resources to help them effectively integrate technology into their teaching practice.

Figure 2. Developing the prompt to include connector words and add citation from specific years.

Can you try again but using “and, but, therefore, however, Moreover” in writing and adding citation inside text and adding references list in the end of the paragraph. Please use published papers during 2017-2023

Figure 3. ChatGPT’s response with more citations and references.

Teachers technostress is a growing concern, particularly in the era of remote learning, with teachers facing an increasing number of demands related to technology integration. Moreover, the pace of technological change means that teachers need to continually learn new skills and tools to be effective in their role (Akçayır & Akçayır, 2017). On the one hand, technology can provide many benefits to teachers, including new opportunities for engagement and student-centered learning (Farr, 2019). However, on the other hand, the pressure to learn and use new technologies can lead to feelings of anxiety, stress, and frustration among teachers.

Studies have shown that teachers technostress can have negative impacts on their job satisfaction and overall well-being (Khan et al., 2020). Therefore, it is important for schools to provide teachers with training and resources to help them effectively integrate technology into their teaching practice. Furthermore, teachers need to be given the time and support necessary to learn new technologies and develop their skills, rather than being expected to learn on their own time (Koseoğlu & Eyyam, 2020).

However, it is also important to recognize that technostress is not solely the responsibility of schools and educational institutions. Teachers also need to take an active role in managing their own stress and anxiety related to technology use. For example, some research has suggested that mindfulness training can help to reduce technostress among teachers (Du et al., 2021). By taking steps to manage their own stress, teachers can maintain their well-being and continue to provide high-quality education to their students.

References:

Qualitative Findings

Findings of Best Practice When Using ChatGPT in Academic Writing

Based on the findings and analysis of the reviewers’ reports from the focus group sessions, the optimal ways to use ChatGPT can be categorized into the following themes: Using Descriptive Prompts, Providing Context to the Prompts, Using Clear Language, and Checking The Outputs Obtained From ChatGPT.

Using Descriptive Prompts

During the development of the 4 articles, the researchers used a wide range of prompts, from simple to detailed, which significantly ($P<.001$) impacted the quality of each version of the research. For example, the researchers started the conversation with ChatGPT by using a simple prompt such as the one illustrated in Multimedia Appendix 2 (the prompts used in the study to generate the 4 versions). Using more detailed prompts improved the quality of the generated text (see data for version 4 in Table 1). By contrast, vague or simple prompts resulted in unrelated responses. For example, when we simply asked about the factors influencing technostress, ChatGPT responded that “TAM was the major framework used to understand technostress”; this is, however, untrue because technostress is related to technology acceptance and adoption.

Providing Context to the Prompts

It is important to provide context when requesting a platform (ChatGPT in this case) to summarize literature or generate a good research idea. Therefore, stating the context you are looking for will maximize your chances of receiving a strong, relevant response to your research. In the second version of the generated article, we noticed an improvement in the clarity of the research idea when we added the context to the prompt. In the first version of the article, we did not ask about teachers’ technostress, whereas in the second version, we solely focused on technostress. Adding context to the prompts was essential to provide us with an accurate response. Therefore, when researchers are experts in their field of study, technology such as ChatGPT can be a helpful tool for them. However, despite an improvement after adding context, the generated text continued to lack the quality and depth of academic writing.

Based on the experience of the researchers, their practices, and the responses of the reviewers in the focus group session, the quality of generating text using ChatGPT depended on the quality of the prompts used by the researchers. ChatGPT produced simple and basic content on a specific topic; in the context of this study this was about teachers’ technostress and continuance intention to use a new technology. The response was more developed with in-text citations, a references list, and the use of connectors between the sentences.

Using Clear Language

To obtain high-quality ideas for their research, researchers need to use simple, clear language containing comprehensive details about the subject matter of the research. We advise researchers to ensure that their prompts contain correct grammar and that they can make any corrections using ChatGPT before writing their prompts; this can be done by asking it to “correct the sentence or the paragraph.”

Checking the Outputs Obtained From ChatGPT

Humans are the experts and ChatGPT is a help tool. After receiving ChatGPT’s writing, it is therefore necessary to check it in terms of the quality of the content, the consequences of the ideas, in-text citations, and the list of references. On examination, we found that the in-text citations were incorrect and the references in the list were fabricated by the application; this therefore influenced the integrity of the research and brought us to the conclusion that ChatGPT can generate fake ideas and unauthentic references. Moreover, the predicted plagiarism in the generated text ranged from 5% to 15% depending on the general concepts used in the articles.

Furthermore, ChatGPT failed to cite in-text references, a common feature in academic writing. The references cited regarding all the queries were also grossly inappropriate or inaccurate. Citing inappropriate or inaccurate references can also be observed in biological intelligence, reflecting a lack of passion, and it is therefore not surprising to see a similar response from an AI tool like ChatGPT.

In addition to the quality of the references cited, this number of references cited was a concern. This problem can often be seen in the context of academic writing from an audience that is not committed to the topic being researched. Another major flaw in the response from ChatGPT was the misleading information regarding the framework of the generated studies, as illustrated in Figure 4.

We used Google Scholar to search for the titles of the cited articles but were unable to find them; this confirmed that ChatGPT fabricated the titles. Moreover, while asking ChatGPT to provide us with a summary of the findings of published articles from 2021 to 2023 in the field of technology integration in education, its response was “training only goes until 2021” as illustrated in Figure 5.
Concerns About Using ChatGPT in Scientific Research

Overview

The potential reviewers in the focus group sessions raised various concerns in terms of using ChatGPT in scientific research. The researchers categorized these concerns into 4 themes, namely, Research Ethics, Research Integrity, Research Quality, and Trustworthiness of Citations and References.

Research Ethics

Most participants in the focus groups raised concerns about the use of ChatGPT due to the risk that it might lead to bias in the information provided in the responses. The reason for the biased responses is due to the use of words and concepts in the wrong context and, as reported by many participants, the application trying to convince researchers that it knows what it is doing. One of the reviews mentioned, “when I reviewed the fourth version, I noticed that all the information about using technology was a bright side and its effects were positive and bright when this was not the case. Based on my experience and research, technology also has a negative effect on users”.

Another ethical concern raised by the reviewers in the focus group discussion regarding the use of ChatGPT was false and misleading information while writing the literature review; this was reported by a majority of the reviewers. One reviewer confirmed that the inaccurate information provided by ChatGPT in scientific research could influence the integrity of the research itself.

Another issue relating to research ethics is authorship; this relates to the authors who have contributed to the research. The reviewers agreed that ChatGPT could not be regarded as a coauthor because it is not human and the generated responses it gives are from data that it has been trained to use. In addition, only a few reviewers connected research ethics to the whole process of conducting research; most viewed it as pertinent solely to the publication phase. They reported that ChatGPT could not be regarded as a corresponding author and could not work on the revision of the research, as can be seen from their reasons given below.
Copyright and ownership are additional ethical concerns when using ChatGPT in scientific research. This is because ChatGPT cannot sign an agreement to publish an article in a journal after it has been accepted for publication. This concern was identified by all the reviewers in the focus group discussion. One reviewer raised the question of ownership regarding the information and ideas generated by ChatGPT: Does the ownership lie with the researcher or with ChatGPT? At the end of the discussion, the reviewers and researchers involved in this study collectively determined that the responsibility for ensuring research accuracy and adherence to all research ethics primarily rested with human researchers. It was therefore the responsibility of researchers to ensure the integrity of their research so that it could be accepted and published in a scientific journal.

**Research Integrity**

Based on the findings of the reviewers’ reports, as well as the discussion among the researchers and the reviewers in the focus group, there was agreement among the participants that they could not have confidence in and trust ChatGPT when it came to scientific research. Some reviewers mentioned that the transparency of providing researchers with information is unknown: How does ChatGPT foster originality in ideas and idea-generation methods? Moreover, according to data in Tables 1 and 2, specifically with regard to citation and references list, ChatGPT fabricated references as well as providing inaccurate information about the theoretical framework of the developed studies, which was also noticed and reported by the reviewers.

**Research Quality**

All the reviewers confirmed that ChatGPT cannot generate original ideas; instead, it merely creates text based on the outlines it is trained to use. Moreover, some reviewers insisted that the information provided by ChatGPT is inconsistent and inaccurate. Therefore, it can mislead researchers, especially a novice who does not have much experience in their field. One reviewer mentioned that ChatGPT can generate reasonable information and provide a researcher with a series of ideas albeit without suitable citations or correct references.

The quality of abstracts generated from published articles was both poor and misleading. The quality was less than the midpoint for acceptance to be published in peer-reviewed journals. An example of the abstract is provided in Multimedia Appendix 3.

**Trustworthiness of Citations and References**

The majority of the reviewers reported that the research generated by ChatGPT lacked in-text citations, which can be considered a type of plagiarism. A few reviewers also expressed that ChatGPT fabricated the references listed. Some related examples are provided in Multimedia Appendices 4 and 5.

**Discussion**

**Principal Findings**

Using AI as an assistive tool in medical education validates its benefits in both medical education and clinical decision-making [47]. The findings of this study underscore the importance of training users to effectively utilize AI-generated text in various fields, particularly in alignment with recent studies advocating for the use of AI tools in medical education [48]. Therefore, AI-generated text tools can be integrated into the medical curriculum and can be used in medical research [49]. The findings of other studies have revealed that ChatGPT, as an example of an AI-generated text tool, can be used as a tool for performing data analysis and making data-driven recommendations/decisions, as demonstrated by the generation of the 4 articles [48,50]. However, the generated knowledge or decision needs approval from humans as mentioned previously [49,50]. AI-generated text assists researchers and medical educators in formulating their decisions by developing the prompts they use in their conversations with the AI tool.

One of the challenges associated with ChatGPT and other AI-generated text tools that emerged during this study is the potential for incorrect information and fake references and in-text citations, which could influence the quality of medical education negatively as reported by [51,52]. In addition, the credibility of scientific research deeply depends on the accuracy of references and resources; however, these are not currently available in AI tools.

**Conclusions**

Based on the analysis of the reviewers’ reports, as well as the focus group discussions, we found that the quality of text generated by ChatGPT depends on the quality of the prompts provided by researchers. Using more detailed and descriptive prompts, as well as appropriate context, improves the quality of the generated text, albeit the quality of the writing and the use of conjunctions between ideas still need improvement. The study identified weaknesses in the list of references cited (with only 8% of the references available when searched for on Google Scholar/Mendeley). We also identified a lack of citations within the text. The study’s findings can inform the use of AI-generated text tools in various fields, including medical education, as an option to assist both practitioners in the field of medicine and researchers in making informed decisions.

In various countries, the issue of journal copyright laws and publishing policies regarding AI-generated text in scientific research requires an ethical code and guidelines; this is in place to address concerns regarding plagiarism, attribution, authorship, and copyright. In a scientific research collaboration with ChatGPT, the generated text should be iterated with human insight, allowing researchers to add their input and thus take ownership of the resulting work. This can lead to higher-level research studies using private data and systematic iteration of the research, making ChatGPT an e-research assistant when used appropriately. Previous studies on using AI-generated text have primarily focused on creating research abstracts and literature syntheses, while some have used AI in different aspects of conducting research. However, the functionality of AI text generators for scientific research highlights the need for the development of an ethical code and guidelines for the use of advanced technology in academic publishing, specifically in relation to concerns regarding plagiarism, attribution, authorship, and copyright. Although ChatGPT is highly efficient in generating answers, it draws information from various sources on the internet, raising concerns about the accuracy and originality of academic papers.
The practical implications of this study are the importance of using descriptive prompts with clear language, and the provision of a relevant context, to improve the accuracy and relevance of the generated text in different fields such as medical education. It is also important to check the outcomes obtained from ChatGPT and to be aware that AI-generated text may be recognized by experienced journal reviewers. The theoretical implications of the study highlight not only the potential of AI-generated text in academic writing but also the need for further research to address the limitations and challenges of this technology. Overall, this study provides insights for researchers and practitioners on how to effectively use ChatGPT in academic writing. Moreover, the tool can be used in the medical research field to analyze data; however, the researchers need to double-check the output to ensure accuracy and validity.

Acknowledgments
All authors declared that they had insufficient or no funding to support open access publication of this manuscript, including from affiliated organizations or institutions, funding agencies, or other organizations. JMIR Publications provided article processing fee (APF) support for the publication of this article. We used the generative AI tool ChatGPT by OpenAI to draft questions for the research survey, which were further reviewed and revised by the study group. The original ChatGPT transcripts are made available as Multimedia Appendices 1-5.

Data Availability Statement
Data will be available upon request.

Authors' Contributions
ZNK contributed to conceptualization, project administration, supervision, data curation, methodology, and writing (both review and editing). AM and JI revised the text, performed the literature review, wrote the original draft, and proofread the paper. AA performed data analysis. MKH and AAH wrote the legal part of the literature review and proofread the paper. MS wrote the original draft.

Conflicts of Interest
None declared.

References

[DOCX File, 46 KB - mededu_v9i1e47049_app1.docx]
[DOCX File, 153 KB - mededu_v9i1e47049_app2.docx]
[DOCX File, 82 KB - mededu_v9i1e47049_app3.docx]
[DOCX File, 50 KB - mededu_v9i1e47049_app4.docx]
[DOCX File, 61 KB - mededu_v9i1e47049_app5.docx]


Abbreviations

AI: artificial intelligence