Viewpoint

Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project

Nicholas I-Hsien Kuo[1*], PhD; Oscar Perez-Concha[1*], PhD; Mark Hanly[1], PhD; Emmanuel Mnatzaganian[2], MSc; Brandon Hao[2], BA; Marcus Di Sipio[2], BHSc; Guolin Yu[2], BA; Jash Vanjara[2], MD; Ivy Cerelia Valerie[2], MD; Juliana de Oliveira Costa[3], PhD; Timothy Churches[4,5], MBBS; Sanja Lujic[1], PhD; Jo Hegarty[6], BIT; Louisa Jorm[1], PhD; Sebastiano Barbieri[1], PhD

[1]Centre for Big Data Research in Health, The University of New South Wales, Sydney, Australia
[2]The University of New South Wales, Sydney, Australia
[3]Medicines Intelligence Research Program, School of Population Health, The University of New South Wales, Sydney, Australia
[4]School of Clinical Medicine, University of New South Wales, Sydney, Australia
[5]Ingham Institute of Applied Medical Research, Liverpool, Sydney, Australia
[6]Sydney Local Health District, Sydney, Australia
[*]these authors contributed equally

Corresponding Author:
Nicholas I-Hsien Kuo, PhD
Centre for Big Data Research in Health
The University of New South Wales
Level 2, AGSM Building (G27), Botany St, Kensington NSW
Sydney, 2052
Australia
Phone: 61 0293850645
Email: n.kuo@unsw.edu.au

Abstract

Large-scale medical data sets are vital for hands-on education in health data science but are often inaccessible due to privacy concerns. Addressing this gap, we developed the Health Gym project, a free and open-source platform designed to generate synthetic health data sets applicable to various areas of data science education, including machine learning, data visualization, and traditional statistical models. Initially, we generated 3 synthetic data sets for sepsis, acute hypotension, and antiretroviral therapy for HIV infection. This paper discusses the educational applications of Health Gym's synthetic data sets. We illustrate this through their use in postgraduate health data science courses delivered by the University of New South Wales, Australia, and a Datathon event, involving academics, students, clinicians, and local health district professionals. We also include adaptable worked examples using our synthetic data sets, designed to enrich hands-on tutorial and workshop experiences. Although we highlight the potential of these data sets in advancing data science education and health care artificial intelligence, we also emphasize the need for continued research into the inherent limitations of synthetic data.

(JMIR Med Educ 2024;10:e51388) doi: 10.2196/51388

KEYWORDS

medical education; generative model; generative adversarial networks; privacy; antiretroviral therapy (ART); human immunodeficiency virus (HIV); data science; educational purposes; accessibility; data privacy; data sets; sepsis; hypotension; HIV; science education; health care AI

Introduction

Clinical data gathered from health care institutions are crucial for enhancing health care quality [1-3]. These data sets can feed into artificial intelligence (AI) and machine learning (ML) models to refine patient prognosis [4,5], diagnosis [6,7], and treatment optimization [8]. Furthermore, statistical models applied to these data sets can uncover association and causal paths [9]. However, stringent privacy regulations protecting patient confidentiality often hamper the prompt availability of these data sets for research and educational usage [10-14].

Gaining access to clinical and health care data sets is a critical aspect of health data science education. This exposure provides trainees with invaluable practical experience, offering profound insights into the complexities of real-world health care scenarios [15]. However, obtaining access to these sensitive data sets is a challenging endeavor—often involving a lengthy process of securing ethics approvals, institutional support, and data clearance [16]. Moreover, the approved users may be required to work on-site under the direct supervision of the data custodian to prevent data leakage [17]. These rigorous security measures, while essential for patient confidentiality, can hamper scalable training of future health data scientists.

During this era of big data, with a soaring demand for skilled health data scientists [18,19], synthetic data sets can bridge the gap between analytical skills and health context comprehension. As Kolaczyk et al [20] astutely asserted, "Theory informs principle, and principle informs practice; practice, in turn, informs theory."

A promising solution to the lack of clinical and health care data is the utilization of generative AI to generate synthetic data sets. These data sets provide controlled, context-specific learning experiences that parallel real-world situations while maintaining patient privacy. The Health Gym project exemplifies this approach [21]. Leveraging generative adversarial networks (GANs) [22-24], Health Gym creates synthetic medical data sets, establishing a secure yet realistic platform for trainees to hone their health data analytical skills. The data sets, covering key health conditions such as sepsis, acute hypotension, and antiretroviral therapy (ART) for HIV infection, can be accessed at [25]. The project's open-source code is also available on GitHub at [26] under the MIT License [27].

As an integral part of the Master of Science in Health Data Science Program at the University of New South Wales (UNSW), Australia [28] and a Datathon event [29], the Health Gym synthetic data sets have proven their versatility and effectiveness in enriching health care education. They are freely accessible to the wider research and education community while complying with stringent security standards such as those specified by Health Canada [30] and the European Medicines Agency [31], thus minimizing patient data disclosure risks.

In this viewpoint paper, we discuss the application of Health Gym synthetic data sets, their role in health data science education, and their potential in nurturing proficient health data scientists. We provide adaptable worked examples (accessible through Section A in Multimedia Appendix 1) by using our synthetic data sets, crafted to enrich hands-on tutorial and workshop experiences. We underline the importance of acknowledging the limitations of synthetic data to ensure their valid use in the creation of statistical models and AI applications in health care and the enhancement of health care education. Although synthetic data sets cannot supersede real-world data, they are a vital tool for training future health data scientists and supporting data-driven innovative approaches in health care.

Ethics Approval

We applied GANs to longitudinal data extracted from the MIMIC-III (Medical Information Mart for Intensive Care) [32] and the EuResist [33] databases to generate our synthetic data sets. This study was approved by the UNSW's human research ethics committee (application HC210661). For patients in MIMIC-III, requirement for individual consent was waived because the project did not impact clinical care and all protected health information was deidentified [32]. For people in the EuResist integrated database, all data providers obtained informed consent for the execution of retrospective studies and inclusion in merged cohorts [34].

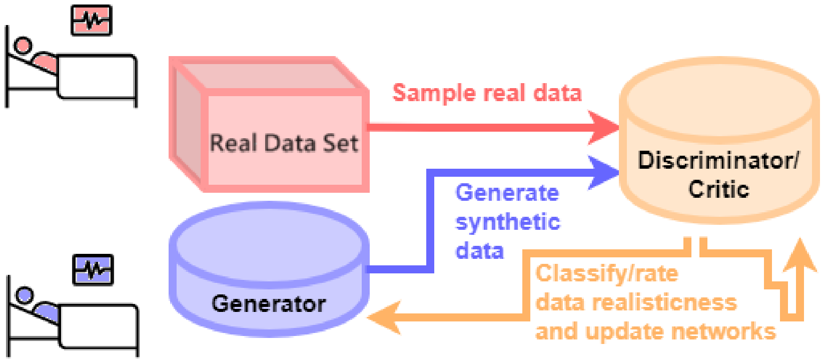
Health Gym

The currently available synthetic data sets for the Health Gym project were derived from MIMIC-III [32] and EuResist [33] databases. MIMIC-III is a comprehensive database of anonymized health data associated with patients admitted to the critical care units of the Beth Israel Deaconess Medical Center, including data on laboratory tests, procedures, and medications. The EuResist network aims to develop a decision support system to optimize ART for individuals living with HIV, leveraging extensive clinical and virological data.

After applying published selection or exclusion criteria, we extracted relevant data from databases that could facilitate the development of patient care algorithms. These data sets, focusing on sepsis, acute hypotension, and ART for HIV, served as the basis for our synthetic data creation. The synthetic data generation employed in the Health Gym was accomplished using GANs. The GAN model, as shown in Figure 1, consists of 2 primary components: a generator and a discriminator. The process starts by sampling real patient records (depicted in pink) and employing the generator to create synthetic patient records (depicted in violet). Both the real and synthetic records are then forwarded to the discriminator network, which is tasked with differentiating the genuine data from the counterfeit. Both networks are trained in an adversarial process—the generator is updated to create more realistic records, while the discriminator is refined to identify generated records more accurately. As a result, the quality of the synthetic data is progressively enhanced, and the synthetic patient records become increasingly representative of the ground truth. The iterative training concludes when the discriminator can no longer reliably distinguish the synthetic records from the real records. Refer to more details in Kuo et al [21].

Leveraging generative AI, Health Gym provides highly authentic clinical data sets, enriching health care education. Each data set undergoes rigorous quality assessment and security verification (detailed in Section B of Multimedia Appendix 1). These synthetic data sets foster engaging learning experiences, aiding educators in developing tailored educational strategies. The following sections will illuminate the application of Health Gym in university-level courses, exemplified through ART for HIV data set.

Figure 1. Generative adversarial network setup.



Synthetic ART for HIV Data Set

The Health Gym data sets contain mixed-type longitudinal data, including numerical, binary, and categorical variables. They encompass patient demographics, vital signs measurements, and pathology results. The data sets hence reflect the complexities of real-life data, thereby making them suitable for training health data scientists in university courses. This paper will primarily delve into the application of synthetic data in health care education focusing on the ART for HIV data set. Readers interested in the sepsis and the acute hypotension data sets should refer to Section C in Multimedia Appendix 1.

Data Set Description

Our synthetic HIV data set, informed by the selection or exclusion criteria proposed by Parbhoo et al [35] and drawn from the EuResist database, targets individuals living with HIV who initiated therapy after 2015 per the World Health Organization's guidelines [36]. ART for HIV typically includes a mix of 3 or more antiretroviral agents from at least 2 distinct medication classes. The dynamism of ART lies in its frequent regimen modifications resulting from various circumstances such as treatment failure due to poor adherence or viral resistance, intolerance to ART, clinical events such as pregnancy or coinfections, or optimization of therapy to support better adherence, reduce drug-drug interactions, maximize ART response, or prevent the emergence of drug-resistant viral strains [36,37].

In addition to ART information, the data set encompasses vital indicators of ART success and disease progression, namely, viral load (VL) and CD4 cell count. Successful ART is often indicated by VL below 1000 copies/mL, while a CD4 cell count exceeding 500 cells/mm^3 signifies healthy immunological status [36]. The complex interactions of these elements in our data set create a rich learning platform for health data science education.

Table 1 encapsulates the data set's 3 numeric, 5 binary, and 5 categorical variables. Numeric variables include VL, CD4 cell count, and relative CD4 laboratory test results. Treatment regimens follow those of Tang et al [38], breaking down the ART regimen into several parts. The data set includes 50 combinations of 21 unique medications. The antiretroviral medication classes are nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs), nonnucleoside reverse transcriptase inhibitors (NNRTIs), integrase inhibitors (INIs), protease inhibitors (PIs), and pharmacokinetic enhancers (pk-En). We deconstructed the ART regimen into its constituent parts: base drug combination (base drug combo), complimentary INIs (comp INIs), comp NNRTIs, extra PIs, and extra pk-En. The base drug combo primarily consists of NRTIs, with inclusion of other antiretroviral classes as well.

Recognizing the notable amount of missing data in the original EuResist database, we added a suffix (M) to variables to denote whether measurements were recorded at specific time points. In the authentic data set, measurements were reported at 24.27% (129,835/534,960) for VL (measured), 22.21% (118,815/534,960) for CD4 (measured), and 85.13% (455,411/534,960) for drug (measured). The absence of some CD4 and VL records may be attributable to specific clinical practices and the frequency of test requests [39-42]. For instance, it is common for clinicians to discontinue requesting a CD4 cell count if the previous result exceeded 500 cells/mm^3 and the individual had an undetectable VL. Similarly, VL is typically measured in the first 3 months, at 6 months, 12 months, and then annually.

Constructed using the GAN model developed by Kuo et al [43], this data set comprises 8916 synthetic patients tracked over 60 months, resulting in 534,960 records (8916 × 60). Figure 2 showcases a sample generated by the code in Figure 3 [44,45]. Each record features 15 columns, including a patient identifier, a time point, and 13 ARTs for HIV variables highlighted in Table 1. The synthetic data sets can be freely accessed in [46] and [47] on Figshare, a digital platform for research output sharing.

Table 1. The variables of antiretroviral therapy in the HIV data set.

Variable name	Data type	Unit	Valid categorical options
Viral load (VL)	numeric	copies/mL	N/A[a]
Absolute count for CD4 (CD4)	numeric	cells/µL	N/A
Relative count for CD4 (Rel CD4)	numeric	cells/µL	N/A
Gender	binary	N/A	Male, Female
Ethnicity (Ethnic)	categorical	N/A	Asian, African, Caucasian, other
Base drug combination (Base drug combo)	categorical	N/A	FTC[b] + TDF[c], 3TC[d] + ABC[e], FTC + TAF[f], DRV[g] + FTC + TDF, FTC + RTVB[h] + TDF, other
Complementary integrase inhibitor (Comp INI)	categorical	N/A	DTG[i], RAL[j], EVG[k], not applied
Complementary nonnucleoside reverse transcriptase inhibitor (Comp NNRTI)	categorical	N/A	NVP[l], EFV[m], RPV[n], not applied
Extra protease inhibitor (Extra PI)	categorical	N/A	DRV, RTVB, LPV[o], RTV[p], ATV[q], not applied
Extra pharmacokinetic enhancer (Extra pk-En)	binary	N/A	False, True
Viral load measured (VL) (M)[r]	binary	N/A	False, True
CD4 (M)	binary	N/A	False, True
Drug recorded (M)	binary	N/A	False, True

[a] N/A: not applicable.

[b] FTC: emtricitabine.

[c] TDF: tenofovir disoproxil fumarate.

[d] 3TC: lamivudine.

[e] ABC: abacavir.

[f] TAF: tenofovir alafenamide.

[g] DRV: darunavir.

[h] RTVB: ritonavir.

[i] DTG: dolutegravir.

[j] RAL: raltegravir.

[k] EVG: elvitegravir.

[l] NVP: nevirapine.

[m] EFV: efavirenz.

[n] RPV: rilpivirine.

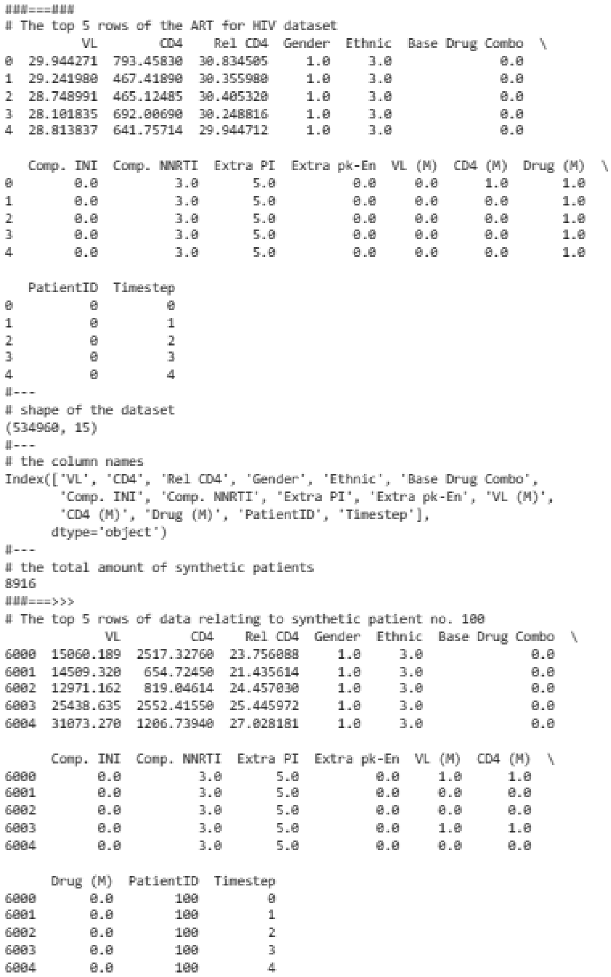
[o] LPV: lopinavir.

[p] RTV: ritonavir.

[q] ATV: atazanavir.

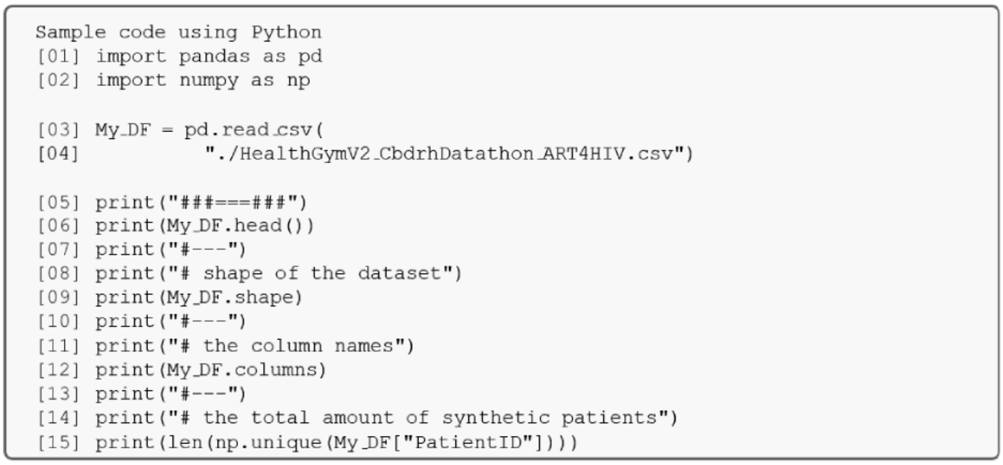
[r] (M): measured.

Figure 2. Inspecting the antiretroviral therapy for an HIV data set (output of the code in Figure 3).



```
###===###
# The top 5 rows of the ART for HIV dataset
          VL        CD4    Rel CD4  Gender  Ethnic  Base Drug Combo  \
0  29.944271  793.45830  30.834505     1.0     3.0              0.0   
1  29.241980  467.41890  30.355980     1.0     3.0              0.0   
2  28.748991  465.12485  30.405320     1.0     3.0              0.0   
3  28.101835  692.00690  30.248816     1.0     3.0              0.0   
4  28.813837  641.75714  29.944712     1.0     3.0              0.0   

   Comp. INI  Comp. NNRTI  Extra PI  Extra pk-En  VL (M)  CD4 (M)  Drug (M)  \
0        0.0          3.0       5.0          0.0     0.0      1.0       1.0   
1        0.0          3.0       5.0          0.0     0.0      0.0       1.0   
2        0.0          3.0       5.0          0.0     0.0      0.0       1.0   
3        0.0          3.0       5.0          0.0     0.0      0.0       1.0   
4        0.0          3.0       5.0          0.0     0.0      0.0       1.0   

   PatientID  Timestep
0          0         0
1          0         1
2          0         2
3          0         3
4          0         4
#---
# shape of the dataset
(534960, 15)
#---
# the column names
Index(['VL', 'CD4', 'Rel CD4', 'Gender', 'Ethnic', 'Base Drug Combo',
       'Comp. INI', 'Comp. NNRTI', 'Extra PI', 'Extra pk-En', 'VL (M)',
       'CD4 (M)', 'Drug (M)', 'PatientID', 'Timestep'],
      dtype='object')
#---
# the total amount of synthetic patients
8916
###===>>>
# The top 5 rows of data relating to synthetic patient no. 100
             VL         CD4    Rel CD4  Gender  Ethnic  Base Drug Combo  \
6000  15060.189  2517.32760  23.756088     1.0     3.0              0.0   
6001  14509.320   654.72450  21.435614     1.0     3.0              0.0   
6002  12971.162   819.04614  24.457030     1.0     3.0              0.0   
6003  25438.635  2552.41550  25.445972     1.0     3.0              0.0   
6004  31073.270  1206.73940  27.028181     1.0     3.0              0.0   

      Comp. INI  Comp. NNRTI  Extra PI  Extra pk-En  VL (M)  CD4 (M)  \
6000        0.0          3.0       5.0          0.0     1.0      1.0   
6001        0.0          3.0       5.0          0.0     0.0      0.0   
6002        0.0          3.0       5.0          0.0     0.0      0.0   
6003        0.0          3.0       5.0          0.0     1.0      1.0   
6004        0.0          3.0       5.0          0.0     0.0      0.0   

      Drug (M)  PatientID  Timestep
6000       0.0        100         0
6001       0.0        100         1
6002       0.0        100         2
6003       0.0        100         3
6004       0.0        100         4
```

Figure 3. Code in Python for generating the output shown in Figure 2. This code uses pandas [44] and NumPy [45]. Base drug combo: base drug combination; comp INI: complementary integrase inhibitor; comp NNRTI: complementary nonnucleoside reverse transcriptase inhibitor; PI: protease inhibitor; pk-En: pharmacokinetic enhancer; VL: viral load.



```
Sample code using Python
import pandas as pd
import numpy as np

My_DF = pd.read_csv(
           "./HealthGymV2_CbdrhDatathon_ART4HIV.csv")

print("###===###")
print(My_DF.head())
print("#---")
print("# shape of the dataset")
print(My_DF.shape)
print("#---")
print("# the column names")
print(My_DF.columns)
print("#---")
print("# the total amount of synthetic patients")
print(len(np.unique(My_DF["PatientID"])))
```

Applications and Case Studies

This section highlights the use of our synthetic ART for HIV data set in a collaborative Datathon event and as an effective teaching tool at UNSW for medical education.

Center for Big Data Research in Health Data Science Datathon

The synthetic data set for ART for HIV was a central component of the UNSW Center for Big Data Research in Health Datathon [48], an event merging theoretical learning with practical application. The Datathon was an enriching exercise in multidisciplinary collaboration. The event involved 6 teams, with a total of 24 participants, offering a tangible experience in data analysis. The student teams were supported by a group of mentors—a blend of data scientists, clinicians, health professionals, and government health informatics specialists from a local health district in Sydney, Australia [49]. The data scientists and the panel of authors of the Health Gym project (ie, Kuo et al [21]) elaborated on the technical aspects and navigated the participants through the intricacies of data analysis, including the assumptions we made to use the data (eg, time 0 corresponded to the date of ART initiation, the laboratory tests occurred before modifications in therapy). Meanwhile, clinicians and health professionals provided their expertise to guide students toward meaningful research questions (eg, discussing VL and CD4 count monitoring, drug-drug interactions, and metabolic toxicity [50]). Government health informaticians, experienced in electronic medical records and real-world population health application and impact, evaluated the usefulness of the students' findings.

This collaborative effort facilitated a comprehensive learning experience, encompassing the development of analytical models, data visualization, and effective communication of research outcomes. Using our synthetic data sets, participants gained valuable insights into working with data sets that emulate real-world health scenarios, thereby providing a bridge between theoretical academia and practical execution.

We summarize the findings of the 2 participating teams below. Detailed reports for Team 1 and Team 2 can be found in Section D and Section E of Multimedia Appendix 1, respectively. In addition, the associated codes for the 2 teams can be found in Section A of Multimedia Appendix 1.

Findings of Team 1

Team 1 investigated the effectiveness of medications, categorized by antiretroviral class, in achieving HIV suppression. Utilizing survival analysis, they assessed the time between the initiation of ART to the first occurrence of viral suppression, defined as VL below 1000 copies/mL [36]. They also assessed the time to CD4 cell count exceeding 500 cells/mm^3 [51], which indicates a healthy immunological status.

With Cox proportional hazards models [52] featuring time-varying covariates, the team identified particular antiretroviral agents associated with viral suppression. These findings were purely associative due to data set limitations, which did not account for factors such as age, socioeconomic status, comorbidities, and concurrent medications (of other illnesses).

Findings of Team 2

Team 2 focused on predicting the necessity of altering an individual's ART regimen over a 5-year time span, factoring in disease flare-ups, resistance, or side effects. They formulated a "sliding search" function that generated individual records for each 12-month period, with predictions for antiretroviral modification and adherence to therapy in the subsequent year by using neural networks. The team's methodology produced promising results, with an accuracy rate of 78% in predicting antiretroviral modification and 93% in predicting adherence to therapy. The algorithm detected trends in CD4 and VL results across the 12-month periods, which appeared to be the key predictive features. In addition, the team suggested that there could be potential benefits from exploring recurrent neural networks (eg, long short-term memory [53]).

Serving as UNSW Coursework Materials

Beyond their utility in the Datathon, our synthetic data sets contribute to UNSW courses in the Master of Science in Health Data Science Program [54], namely, HDAT9800 Visualization & Communication and HDAT9510 Machine Learning II.

HDAT9800 teaches future health data scientists the skills to visually communicate complex data effectively to diverse audiences. The course emphasizes the significance of clear data visualization and advocates for transparency and reproducibility in scientific work. It employs R [55] and Python [56] to demonstrate best practices in data analysis and visualization. Our synthetic data sets provide rich resources to enhance the learning in this setting. For instance, Marchesi et al [57] used our data sets to present patient states via t-distributed stochastic neighbor embedding visualization techniques [58].

Meanwhile, HDAT9510 explores advanced modern ML algorithms and methods such as convolutional neural networks [59], autoencoders [60], and reinforcement learning (RL) [61]. As the synthetic data sets consist of time-series variables, students can develop both feedforward and recurrent neural networks. See example models built using our data set in Marchesi et al [57] with recurrent neural networks and even decision trees [62] and hidden Markov models [63], as in a similar data set suggested by Wu et al [64]. Furthermore, with the presence of nonnumeric variables, students can learn about embedding [65]—transforming nonnumeric levels into real-valued vectors so that similar levels that are closer in the vector space carry more analogous meaning. The presence of missing data in the synthetic data sets also encourages students to formulate plausible assumptions about the structure of the clinical data set prior to data modelling.

We provide 3 adaptable worked examples using our ART for HIV data set, suitable for workshops and lectures. The associated codes for the worked examples can be found in Section A of Multimedia Appendix 1. Our synthetic data set supports a variety of student engagements, from understanding complex data structures to developing advanced RL algorithms for optimizing clinical interventions. Moreover, the low patient disclosure risk associated with our data sets (refer to Section B in Multimedia Appendix 1) eliminates the need for ethics approval [66]. This makes these data sets ideal for a range of settings—from small seminars to larger lecture groups.

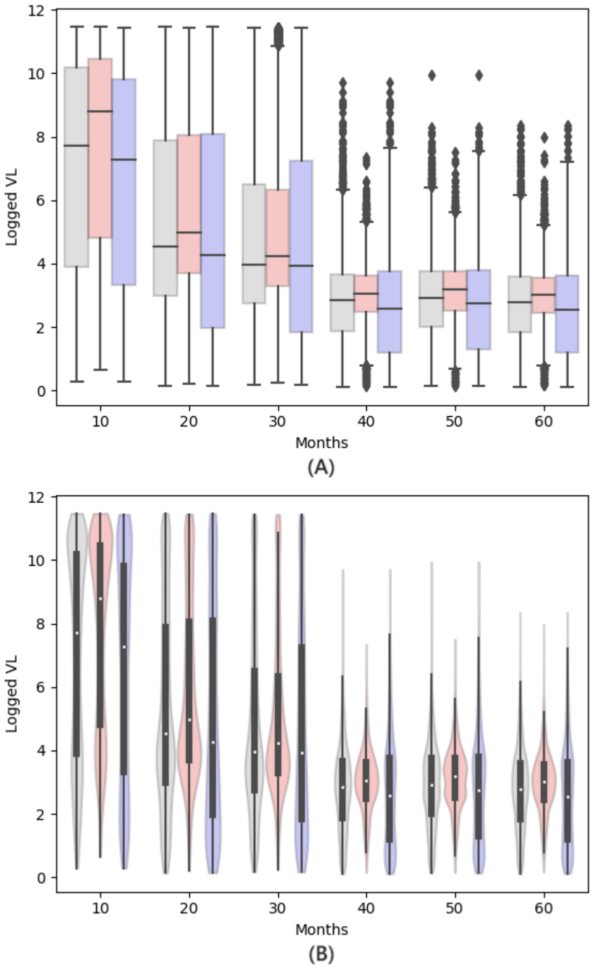
Worked Example 1

The first exercise, focused on data visualization using Python, compares VL trends over time among patients who commenced their ART with different base drug combos, against the general trend in all patients. The results of our worked example are depicted in Figure 4.

This multifaceted exercise requires students to create sub–data sets based on specific starting base drug combos (ie, FTC + TDF [emtricitabine + tenofovir disoproxil fumarate] and 3TC + ABC [lamivudine + abacavir]), extract data for defined periods, and familiarize themselves with box and violin plots [67]. They are also tasked with organizing the visual data as side-by-side plots.

Through this exercise, students will understand the limitations of box plots, which cannot visualize underlying data distributions. They will learn about the additional insights provided by advanced plotting techniques such as violin plots. In addition, students will note that people who start with FTC + TDF and those who start with 3TC + ABC display similar patterns as the overall ART for HIV cohort. The overlap of the interquartile ranges across all box plots indicates a consistent behavior.

Figure 4. Viral load distribution. Subplot (A) shows a box plot comparison of viral load across base drug combinations across time, and subplot (B) shows a violin plot comparison of viral load across base drug combinations across time. Grey indicates all patients, red indicates those initiating treatment with FTC + TDF (emtricitabine + tenofovir disoproxil fumarate), and blue indicates those initiating treatment with 3TC + ABC (lamivudine + abacavir). VL: viral load.



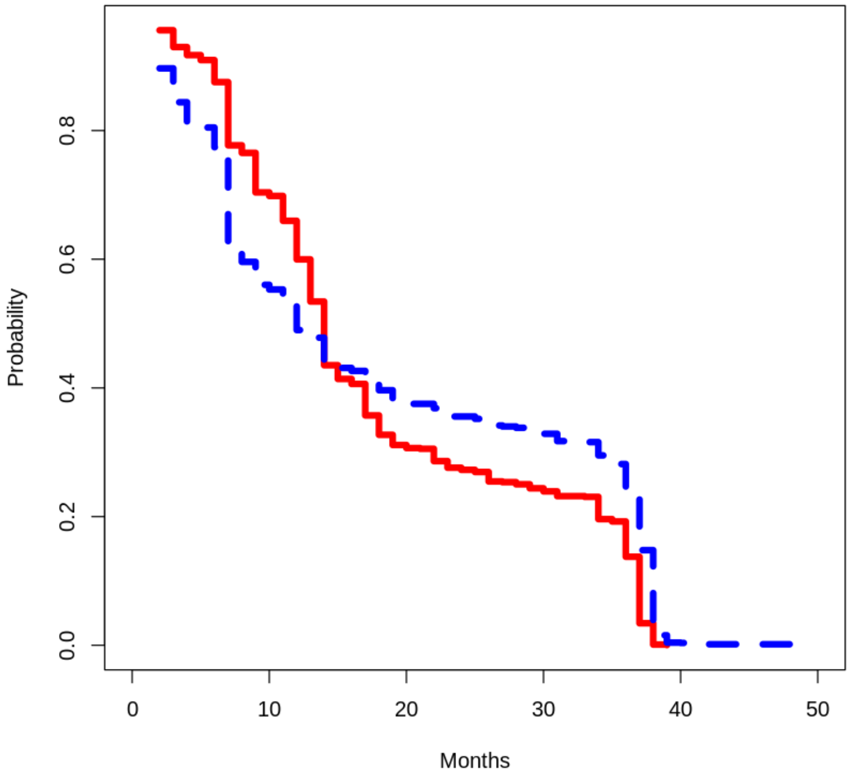
Worked Example 2

The second exercise delves into survival analysis using R [55], building on insights from the initial data visualization task. The exercise continues to compare results among people starting with the base drug combo of FTC + TDF and those initiating with the base drug combo of 3TC + ABC. The goal is to estimate the time necessary for a person on ART to successfully suppress their VL. The results of our worked example are depicted in Figure 5.

This task proves to be more complex than the first, requiring HIV domain knowledge, such as an understanding that a reasonable threshold for ART in HIV treatment is 1000 copies/mL [36]. This threshold indicates slowed viral replication and immune system damage. Thus, students should select patients who commence ART with VL above 1000 copies/mL (ie, not experiencing the outcome of interest at baseline).

Creating an appropriate data set for survival analysis is key, as is pinpointing when each patient’s VL first drops to or below 1000 copies/mL. In addition, students need to grasp the concept of right censoring [68] and utilize Kaplan-Meier curves [69] for time-to-event estimations. This offers an opportunity to engage with the influential survival package [70] in the R language. Upon examining the results in Figure 5, students will note no significant differences in the timing of VL suppression between people who started with the base drug combo of FTC + TDF and those who initiated with the base drug combo of 3TC + ABC.

Figure 5. Time-to-event estimation of viral load suppression for viral load lower than 1000 copies/mL. Red indicates those initiating treatment with FTC + TDF (emtricitabine + tenofovir disoproxil fumarate) and blue for those initiating treatment with 3TC + ABC (lamivudine + abacavir).



Worked Example 3

The third exercise immerses students in the process of developing an RL agent using Python. RL is a type of ML that learns an evidence-based policy to connect states (the current scenario) to actions (the potential responses to that scenario). In the context of our HIV treatment example, states refer to the representation of the patient's current health status and medication history, while action refers to the selection of medication to use in response to each state.

The RL agent selects an action based on a policy that optimizes for maximum cumulative rewards, even as environments evolve. This approach has particular relevance to health care. Clinicians often need to adapt treatment plans to each patient's unique circumstances, and RL can help them to individualize treatment durations, dosages, or types. For example, they may alter the regimen, class, or specific agents of medication to better serve the patient's needs. The outcomes of our example are visualized in Figure 6. This exercise highlights the potential of RL to enhance patient care through personalization—an aspect that is becoming increasingly important in today's medical landscape.

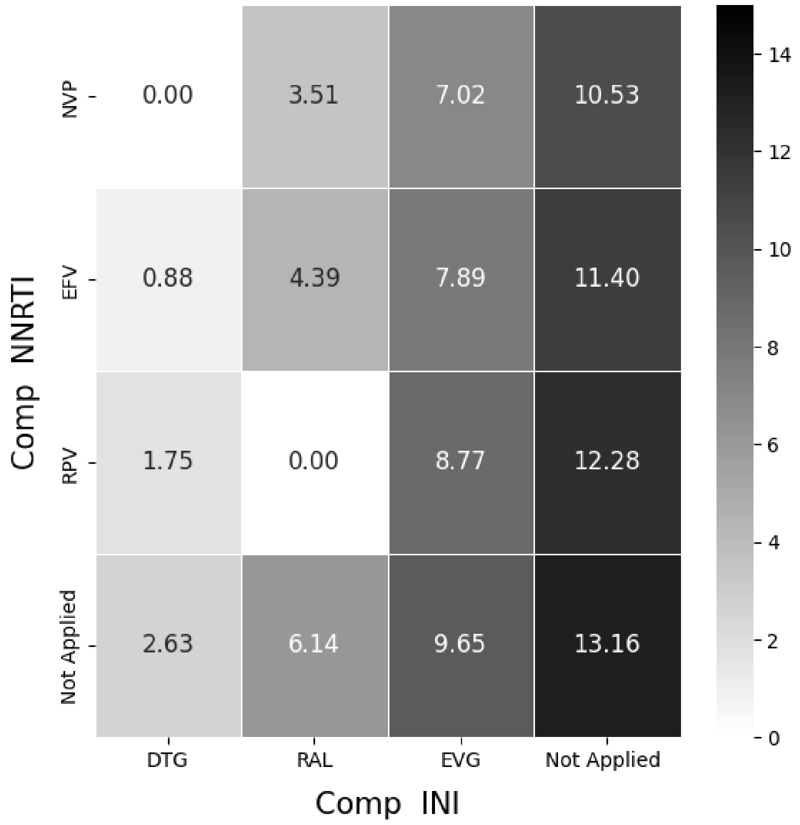
This complex exercise is designed for advanced students, posing challenges across multiple dimensions. It commences with data wrangling, where students scrutinize numeric variable distributions and evaluate the necessity for transformations such as rescaling, normalization [71], power transformation [72], or Box-Cox transformation [73].

In the next stage, students encounter categorical feature representation for medication regimens, practicing their skills in implementing embeddings. Advanced students can explore transfer learning for feature representation [74]. This exercise also presents real-world challenges, requiring students to handle mixed-type data progression. During the model fitting phase, students must employ suitable ML models, distinguishing between RL method archetypes [75] and considering their clinical implications.

Data visualization is the next task, encouraging students to articulate model-derived insights into digestible visuals for a diverse audience. The concluding phase involves refining assumptions and model performance, incorporating multiple tests to identify optimal hyperparameters [76]. Here, students peek into the "black box" nature of ML and gain an intuition for effective module combinations [77-79]. This step becomes critical for causal inference tasks that necessitate rigorous input data validation [80].

Figure 6 showcases the strategy employed by an RL agent in HIV therapy. Heatmaps visualize the relative frequencies of chosen actions (ie, the selected antiretroviral), where each tile represents a unique action and its frequency as a proportion of all actions. The example output shows that the RL agent consistently suggests the EFV + RAL (efavirenz + raltegravir)—a combination of comp NNRTIs and comp INIs—4.39% of the time, while never recommending the RPV + RAL (rilpivirine + raltegravir) combination. More information on the steps taken to create the output for this task can be found in Section F of Multimedia Appendix 1.

Figure 6. Visualizing the learned reinforcement learning policy. Comp INI: complementary integrase inhibitor; Comp NNRTI: complementary nonnucleoside reverse transcriptase inhibitor; DTG: dolutegravir; EFV: efavirenz; EVG: elvitegravir; NVP: nevirapine; RAL: raltegravir; RPV: rilpivirine.



Discussion

This paper demonstrates the transformative potential of synthetic health data sets in health care education, especially in the evolving context of generative AI integration. These data sets provide a realistic representation of real-world health data complexities while preserving patient confidentiality, facilitating experiential learning, skills enhancement, and interdisciplinary collaboration. However, this significant stride toward AI integration in education is not without challenges, and the creation of AI models trained on curated quality data sets emerges as a promising research area.

Despite our best efforts, the Health Gym synthetic data sets might not fully capture the complexity and diversity of real-world scenarios. For instance, some critical health determinants such as socioeconomic status [81] and comorbidities [82] are missing from the ART for HIV synthetic data sets. The absence of these factors mirrors the broader issues concerning data accessibility [83], particularly when it involves specialized or rare disease information. Furthermore, synthetic data might overlook uncontrolled variables or confounders inherent in real-world data [84,85], posing pedagogical challenges. However, this limitation is not solely attributable to our methodology. Since the socioeconomic status variable is not present in the EuResist database, our model lacked the necessary reference data from the outset.

In the field of health data science, proficient data set management and curation are essential due to the decentralized nature of health care data collection. Many entities contribute to health data, each using their own systems [86]. Privacy laws such as Australia's Privacy Act 1988 [87] and the United States' Health Insurance Portability and Accountability Act [88] complicate the sharing of data, resulting in a fragmented view of patient information.

Record linkage techniques [89] such as probabilistic matching [90] bridge this gap by linking disparate data records, offering a more comprehensive view of a patient's health. Nevertheless, our synthetic data sets, despite their potential, carry limitations such as the absence of a master linkage key [91], thereby reducing their applicability in university courses for data management and curation. Having such linked data sets are also great for health data science students to test hypotheses on the effects of comorbidities. Our experiences from the Datathon suggest that the Health Gym synthetic data sets are best used for creating algorithms to enhance patient care within specific disease management paradigms.

Our Health Gym initiative leverages a unique application of generative AI, differing from those used in emerging AI-assisted chatbots, which have also shown promise as potent educational tools. AI chatbots, with their personalized and interactive responses using large language models, can significantly incite interest and foster self-directed learning in medical students [92]. However, advanced AI tools such as OpenAI's ChatGPT [93] and Google's BARD [94] bring with them valid concerns around precision, reliability, potential misuse, and adherence to academic integrity [95,96]. In contrast, the synthetic clinical data sets, the generative product of our Health Gym project, offer controlled, scenario-specific learning environments that closely reflect real-world conditions while preserving patient privacy.

Access to clinical data sets is integral to health data science education, but the necessity of maintaining patient confidentiality can hinder the training of future health data scientists on a larger scale. This may exacerbate the digital divide [97,98], which is a prominent challenge in the broader AI integration into education. As we shift toward AI-driven educational resources, it is essential to prioritize equitable access across varied socioeconomic backgrounds. Future research should evaluate the long-term effects of AI on student learning, clinical judgment, patient outcomes, and the development of educational resources for effective AI integration. The secure, realistic synthetic data sets of Health Gym may provide a valuable solution, potentially facilitating equal access to educational materials.

Conclusion

Despite their limitations, the Health Gym synthetic health data sets have demonstrated their value in educating and training future health data scientists. Their integration into interdisciplinary platforms such as Datathon illustrates their potential in promoting collaborative learning, skills enhancement, and innovative research. In addition, synthetic data sets offer a learning platform that balances realistic health scenario representation with data privacy preservation.

Although we have primarily demonstrated the utility of Health Gym's synthetic data sets by using the ART for HIV data set, we emphasize the importance of the additional acute hypotension and sepsis data sets that we have developed (see Section C in Multimedia Appendix 1). These data sets broaden the scope of medical education by providing insight into managing illnesses in intensive care units, encompassing a unique set of measurements and pathology information. As such, these synthetic data sets offer students an enriched, realistic learning environment for health data science education, complementing the HIV data set and furthering the applicability and versatility of synthetic health data.

The majority of generative ML research is centered on computer vision [99,100] and, to a lesser extent, natural language processing [101], leaving clinical health care data relatively unexplored. This gap suggests a valuable opportunity for future research, particularly considering that clinical data being longitudinal, mixed-type time series variables have a fundamentally different nature. As demonstrated in our prior studies [21,43,102], we have ascertained that our synthetic data sets attain a robust level of validity and are readily available to support both clinical research and medical pedagogy; predictive models instantiated on our synthetic data sets parallel those of the original data sets in their characteristics. We will focus our future work on comparing synthetic data sets created using various generative ML architectures, for example, GANs, variational autoencoders [103], diffusion probabilistic models [102,104], and transformer-based models [105].

GANs, like other ML models, can only optimize according to predefined optimization functions. Given the current lack of research on the use of GANs in health care, more utility studies are necessary to fully comprehend the potential of our synthetic data sets. We are committed to continuing collaboration with clinicians and health professionals to better understand the practical strengths and weaknesses of synthetic data sets, including how to better evaluate and contain the risk of private information disclosure. Through these collective efforts, we aim to improve the quality of synthetic data sets, enhancing hands-on learning experiences for students in health data analytics.

Acknowledgments

This study benefited from data provided by the EuResist Network EIDB, and this project has been funded by a Wellcome Trust Open Research Fund (reference 219691/Z/19/Z). JdOC is supported by the Medicines Intelligence Center of Research Excellence (grant 1196900).

Authors' Contributions

Authors NI-HK and SB were responsible for the design, implementation, and validation of the deep learning models employed to generate the synthetic data sets for the Health Gym project. The inception of Datathon was conceived by OP-C and MH who liaised with various disciplinary personnel to realize this initiative. JdOC contributed specialist knowledge on antiretroviral therapy for HIV to Datathon, while JH offered expertise in the evaluation of Datathon projects. Furthermore, TC and SL, alongside OP-C and MH, leveraged their extensive teaching experience to guide Datathon participants and explore further applications of the Health Gym synthetic data sets. LJ provided key insights on the potential risk of sensitive information disclosure. Datathon participants EM, BH, MDS, GY, JV, and ICV gave critical feedback on the strengths and shortcomings of the synthetic data sets, in addition to providing valuable reflections on the event itself. This manuscript was prepared by NI-HK. All authors contributed to interpreting the findings and revising the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary data.
[DOCX File , 38 KB-Multimedia Appendix 1]

References

1. Alsuliman T, Humaidan D, Sliman L. Machine learning and artificial intelligence in the service of medicine: necessity or potentiality? Curr Res Transl Med. Nov 2020;68(4):245-251. [doi: 10.1016/j.retram.2020.01.002] [Medline: 32029403]
2. Naseem M, Akhund R, Arshad H, Ibrahim MT. Exploring the potential of artificial intelligence and machine learning to combat COVID-19 and existing opportunities for LMIC: a scoping review. J Prim Care Community Health. 2020;11:2150132720963634. [FREE Full text] [doi: 10.1177/2150132720963634] [Medline: 32996368]
3. Wood D. Wicked problems: using data for better public policy. The Australian Parliamentary Budget Office. URL: https://www.pbo.gov.au/sites/default/files/2023-03/PBO%20Conference_Danielle%20Wood_Data%20and%20wicked%20problems.pdf [accessed 2023-12-26]
4. Jin X, Gallego Luxan B, Hanly M, Pratt NL, Harris I, de Steiger R, et al. Estimating incidence rates of periprosthetic joint infection after hip and knee arthroplasty for osteoarthritis using linked registry and administrative health data. The Bone & Joint Journal. Sep 01, 2022;104-B(9):1060-1066. [doi: 10.1302/0301-620x.104b9.bjj-2022-0116.r1]
5. Barbieri S, Mehta S, Wu B, Bharat C, Poppe K, Jorm L, et al. Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach. Int J Epidemiol. Jun 13, 2022;51(3):931-944. [FREE Full text] [doi: 10.1093/ije/dyab258] [Medline: 34910160]
6. Feng YZ, Liu S, Cheng ZY, Quiroz JC, Rezazadegan D, Chen PK, et al. Severity assessment and progression prediction of COVID-19 patients based on the LesionEncoder framework and chest CT. J Med Internet Res. Preprint posted online on March 18, 2021. [doi: 10.2196/preprints.28903]
7. Bayer J, Spark J, Krcmar M, Formica M, Gwyther K, Srivastava A, et al. The SPEAK study rationale and design: A linguistic corpus-based approach to understanding thought disorder. Schizophr Res. Sep 2023;259:80-87. [doi: 10.1016/j.schres.2022.12.048] [Medline: 36732110]
8. Bachmann N, Von Siebenthal C, Vongrad V, Turk T, Neumann K, Beerenwinkel N, et al. Determinants of HIV-1 reservoir size and long-term dynamics during suppressive ART. Nature Communications. Jul 19, 2019:1. [doi: 10.1101/19013763]
9. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet. 2019;10:524. [FREE Full text] [doi: 10.3389/fgene.2019.00524] [Medline: 31214249]
10. Nosowsky R, Giordano TJ. The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. Annu Rev Med. 2006;57:575-590. [doi: 10.1146/annurev.med.57.121304.131257] [Medline: 16409167]
11. O'Keefe CM, Connolly CJ. Privacy and the use of health data for research. Med J Aust. Nov 01, 2010;193(9):537-541. [doi: 10.5694/j.1326-5377.2010.tb04041.x] [Medline: 21034389]
12. Bentzen HB, Castro R, Fears R, Griffin G, Ter Meulen V, Ursin G. Remove obstacles to sharing health data with researchers outside of the European Union. Nat Med. Aug 2021;27(8):1329-1333. [FREE Full text] [doi: 10.1038/s41591-021-01460-0] [Medline: 34345050]
13. de Oliveira Costa J, Bruno C, Schaffer AL, Raichand S, Karanges EA, Pearson S. The changing face of Australian data reforms: impact on pharmacoepidemiology research. Int J Popul Data Sci. Apr 15, 2021;6(1):1418. [FREE Full text] [doi: 10.23889/ijpds.v6i1.1418] [Medline: 34007904]
14. Pearson S, Pratt N, de Oliveira Costa J, Zoega H, Laba T, Etherton-Beer C, et al. Generating real-world evidence on the quality use, benefits and safety of medicines in Australia: history, challenges and a roadmap for the future. IJERPH. Dec 18, 2021;18(24):13345. [doi: 10.3390/ijerph182413345]
15. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data. Jun 19, 2019;6(1):1. [doi: 10.1186/s40537-019-0217-0]
16. Data availability and transparency bill 2022. Australian Parliament House. URL: https://www.aph.gov.au/Parliamentary_Business/Bills_LEGislation/Bills_Search_Results/Result?bId=r6649 [accessed 2023-12-26]
17. The Five Safes framework. Australian Bureau of Statistics. URL: http://tinyurl.com/4t3nnxpf [accessed 2023-12-26]
18. Miller S, Hughes D. The Quant Crunch: how the demand for data science skills is disrupting the job market. Business-Higher Education Forum. 2017. URL: https://www.bhef.com/publications/quant-crunch-how-demand-data-science-skills-disrupting-job-market [accessed 2023-12-26]
19. Columbus L. IBM predicts demand for data scientists will soar 28% by 2020. Forbes. May 13, 2017. URL: https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/?sh=7fe27cff7e3b [accessed 2023-12-26]
20. Kolaczyk ED, Wright H, Yajima M. Statistics practicum: placing 'practice' at the center of data science education. Harvard Data Science Review. Jan 29, 2021:1. [doi: 10.1162/99608f92.2d65fc70]
21. Kuo NIH, Polizzotto MN, Finfer S, Garcia F, Sönnerborg A, Zazzi M, et al. The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms. Sci Data. Nov 11, 2022;9(1):693. [FREE Full text] [doi: 10.1038/s41597-022-01784-7] [Medline: 36369205]
22. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. Oct 22, 2020;63(11):139-144. [doi: 10.1145/3422622]

23. Arjovsky M, Chintala S, Bottou L. Wasserstein Generative Adversarial Networks. Presented at: International Conference on Machine Learning; August 6, 2017; Sydney, Australia. URL: https://proceedings.mlr.press/v70/arjovsky17a.html
24. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. Presented at: Neural Information Processing Systems; 2017, 2017; Long Beach, California.
25. Kuo NIH. The Health Gym. HealthGym.ai. URL: https://healthgym.ai/ [accessed 2023-12-26]
26. Nic5472K / ScientificData2021_HealthGym. GitHub. URL: https://github.com/Nic5472K/ScientificData2021_HealthGym [accessed 2023-12-27]
27. Rosen L. Open Source Licensing: Software Freedom and Intellectual Property Law. Jul 01, 2004. URL: https://www.immagic.com/eLibrary/ARCHIVES/EBOOKS/R050225R.pdf [accessed 2023-12-27]
28. Graduate certificate in Health Data Science. The University of New South Wales. URL: https://www.unsw.edu.au/study/postgraduate/graduate-certificate-in-health-data-science?studentType=Domestic [accessed 2023-12-27]
29. CBDRH Health Data Science Datathon 2023. GitHub. URL: https://cbdrh-hds-datathon-2023.github.io/ [accessed 2023-12-27]
30. Public release of clinical information: guidance document. Government of Canada. URL: https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance/document.html, [accessed 2023-12-27]
31. Clinical data publication. European Medicines Agency. URL: https://www.ema.europa.eu/en/human-regulatory-overview/marketing-authorisation/clinical-data-publication#:~:text=The%20Agency%20intends%20to%20gradually,%3A%2014%2D15%20December%202022 [accessed 2023-12-27]
32. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. May 24, 2016;3:160035. [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
33. Zazzi M, Incardona F, Rosen-Zvi M, Prosperi M, Lengauer T, Altmann A, et al. Predicting response to antiretroviral treatment by machine learning: the EuResist project. Intervirology. 2012;55(2):123-127. [FREE Full text] [doi: 10.1159/000332008] [Medline: 22286881]
34. Prosperi MCF, Rosen-Zvi M, Altmann A, Zazzi M, Di Giambenedetto S, Kaiser R, et al. Correction: antiretroviral therapy optimisation without genotype resistance testing: a perspective on treatment history based models. PLoS ONE. Apr 26, 2011;6(4):1. [doi: 10.1371/annotation/d0254103-21b9-4078-836b-57ba5bd1c26a]
35. Parbhoo S, Bogojeska J, Zazzi M, Roth V, Doshi-Velez F. Combining kernel and model based learning for HIV therapy selection. AMIA Jt Summits Transl Sci Proc. 2017;2017:239-248. [FREE Full text] [Medline: 28815137]
36. Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach, 2nd ed. World Health Organization. URL: https://www.who.int/publications/i/item/9789241549684 [accessed 2023-12-27]
37. Bennett DE, Bertagnolio S, Sutherland D, Gilks CF. The World Health Organization's global strategy for prevention and assessment of HIV drug resistance. Antiviral Therapy. Feb 01, 2008;13(2_suppl):1-13. [doi: 10.1177/135965350801302s03]
38. Tang MW, Liu TF, Shafer RW. The HIVdb system for HIV-1 genotypic resistance interpretation. Intervirology. 2012;55(2):98-101. [FREE Full text] [doi: 10.1159/000331998] [Medline: 22286876]
39. Fox MP, Brennan AT, Nattey C, MacLeod WB, Harlow A, Mlisana K, et al. Delays in repeat HIV viral load testing for those with elevated viral loads: a national perspective from South Africa. J Int AIDS Soc. Jul 2020;23(7):e25542. [FREE Full text] [doi: 10.1002/jia2.25542] [Medline: 32640101]
40. Hill AL, Rosenbloom DIS, Goldstein E, Hanhauser E, Kuritzkes DR, Siliciano RF, et al. Real-time predictions of reservoir size and rebound time during antiretroviral therapy interruption trials for HIV. PLoS Pathog. Apr 2016;12(4):e1005535. [FREE Full text] [doi: 10.1371/journal.ppat.1005535] [Medline: 27119536]
41. What's new in treatment monitoring: viral load and CD4 testing. World Health Organisation. URL: https://www.who.int/publications/i/item/WHO-HIV-2017.22 [accessed 2023-12-27]
42. NSW HIV strategy 2021-2025. New South Wales Health. URL: https://www.health.nsw.gov.au/endinghiv/Pages/nsw-hiv-strategy-2021-2025.aspx [accessed 2023-12-27]
43. Kuo NIH, Garcia F, Sönnerborg A, Böhm M, Kaiser R, Zazzi M, EuResist Network study group; et al. Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV. J Biomed Inform. Aug 2023;144:104436. [FREE Full text] [doi: 10.1016/j.jbi.2023.104436] [Medline: 37451495]
44. Mckinney W. Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference (SciPy 2010). 2010:1. [doi: 10.25080/majora-92bf1922-00a]
45. van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Comput Sci Eng. Mar 2011;13(2):22-30. [doi: 10.1109/mcse.2011.37]
46. Kuo NIH. The Heath Gym synthetic HIV dataset. Figshare. URL: https://figshare.com/articles/dataset/The_Heath_Gym_Synthetic_HIV_Dataset/19838470 [accessed 2023-12-27]
47. Kuo NIH. The Health Gym v2.0 synthetic antiretroviral therapy (ART) for HIV dataset. Figshare. URL: https://figshare.com/articles/dataset/The_Health_Gym_v2_0_Synthetic_Antiretroviral_Therapy_ART_for_HIV_Dataset/22827878 [accessed 2023-12-27]

48. Datathon highlights. CBDRH Health Data Science Datathon 2023. URL: https://cbdrh-hds-datathon-2023.github.io/review.html [accessed 2023-12-27]
49. Sydney local health district. New South Wales Health. URL: https://slhd.health.nsw.gov.au/ [accessed 2023-12-27]
50. de Oliveira Costa J, Lau S, Medland N, Gibbons S, Schaffer AL, Pearson S. Potential drug-drug interactions due to concomitant medicine use among people living with HIV on antiretroviral therapy in Australia. Br J Clin Pharmacol. May 2023;89(5):1541-1553. [doi: 10.1111/bcp.15614] [Medline: 36434744]
51. Garcia SAB, Guzman N. Acquired immune deficiency syndrome CD4+ count. StatPearls. Aug 14, 2023:1. [Medline: 30020661]
52. Cox DR. Regression models and life tables. Journal of the Royal Statistical Society: Series B (Methodological). Dec 05, 2018;34(2):187-202. [doi: 10.1111/j.2517-6161.1972.tb00899.x]
53. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. Nov 15, 1997;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
54. Master of Science in Health Data Science. The University of New South Wales. URL: https://www.unsw.edu.au/study/postgraduate/master-of-science?cq_plac=&studentType=Domestic [accessed 2023-12-27]
55. R: a language and environment for statistical computing. R-Project. URL: https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing [accessed 2023-12-28]
56. Van Rossum G. Python tutorial. Centrum Wiskunde & Informatica Institutional Repository. 1995. URL: https://ir.cwi.nl/pub/5007 [accessed 2023-12-27]
57. Marchesi R, Micheletti N, Jurman G, Osmani V. Mitigating health data poverty: generative approaches versus resampling for time-series clinical data. ArXiv. Preprint posted online on October 26, 2022. [doi: 10.48550/arXiv.2210.13958]
58. Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research. 2008. URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf [accessed 2023-12-27]
59. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. May 28, 2015;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
60. Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal. Jun 17, 2004;37(2):233-243. [doi: 10.1002/aic.690370209]
61. Sutton R, Barto A. Reinforcement learning: an introduction. IEEE Trans Neural Netw. Sep 1998;9(5):1054-1064. [doi: 10.1109/tnn.1998.712192]
62. Winterfeldt D, Edwards W. Decision Analysis and Behavioral Research. Cambridge, Massachusetts, USA. Cambridge University Press; Aug 26, 1986.
63. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Statist. Dec 1966;37(6):1554-1563. [doi: 10.1214/aoms/1177699147]
64. Wu M, Hughes M, Parbhoo S, Zazzi M, Roth V, Doshi-Velez F. Beyond sparsity: tree regularization of deep models for interpretability. In: AAAI. Presented at: AAAI Conference on Artificial Intelligence; April 25, 2018; Chicago, Illinois, USA. [doi: 10.1609/aaai.v32i1.11501]
65. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. ArXiv. Preprint posted online on September 7, 2013 [FREE Full text] [doi: 10.48550/arXiv.1301.3781]
66. Barnett AG, Campbell MJ, Shield C, Farrington A, Hall L, Page K, et al. The high costs of getting ethical and site-specific approvals for multi-centre research. Res Integr Peer Rev. 2016;1:16. [FREE Full text] [doi: 10.1186/s41073-016-0023-6] [Medline: 29451546]
67. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. May 2007;9(3):90-95. [doi: 10.1109/mcse.2007.55]
68. Wu MC, Carroll RJ. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics. Mar 1988;44(1):175. [doi: 10.2307/2531905]
69. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. Jun 1958;53(282):457-481. [doi: 10.1080/01621459.1958.10501452]
70. Therneau TM, Lumley T, Elizabeth A, Cynthia C. survival: Survival Analysis. The Comprehensive R Archive Network. 2015. URL: https://cran.r-project.org/web/packages/survival/index.html [accessed 2023-12-27]
71. Patro SGK, Sahu KK. Normalization: a preprocessing stage. ArXiv. Preprint posted online on March 19, 2015 [FREE Full text] [doi: 10.17148/iarjset.2015.2305]
72. Caroll RJ, Ruppert D. On prediction and the power transformation family. Biometrika. 1981;68(3):609-615. [doi: 10.1093/biomet/68.3.609]
73. Box GEP, Cox DR. An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological). Dec 05, 2018;26(2):211-243. [doi: 10.1111/j.2517-6161.1964.tb00553.x]
74. Bengio Y. Deep learning of representations for unsupervised and transfer learning. Presented at: International Conference on Machine Learning Unsupervised and Transfer Learning Workshop; July 2, 2011; Bellevue, Washington, USA. URL: https://proceedings.mlr.press/v27/bengio12a/bengio12a.pdf
75. Levine S, Kumar A, Tucker G, Fu J. Offline reinforcement learning: tutorial, review, and perspectives on open problems. ArXiv. Preprint posted online on November 1, 2020 [FREE Full text] [doi: 10.48550/arXiv.2005.01643]

76. Bergstra J, Yamins D, Cox DD. Making a science of model search. Presented at: International Conference on Machine Learning; June 21, 2013; Atlanta, USA.
77. Kuo NIH. Understanding and modifying dynamical Hopfield neural networks for generating multiple coherent patterns [PhD thesis]. The University of Auckland. 2017. URL: https://researchspace.auckland.ac.nz/handle/2292/34849 [accessed 2023-12-27]
78. Kuo NIH, Harandi M, Fourrier N, Walder C, Ferraro G, Suominen H. An input residual connection for simplifying gated recurrent neural networks. Presented at: International Joint Conference on Neural Networks; July 19, 2020; Glasgow, United Kingdom. [doi: 10.1109/ijcnn48605.2020.9207238]
79. Kuo NIH, Harandi M, Fourrier N, Walder C, Ferraro G, Suominen H. Plastic and stable gated classifiers for continual learning. Presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops; June 19, 2021; Online. [doi: 10.1109/cvprw53098.2021.00394]
80. Walker AR, Luque D, Le Pelley ME, Beesley T. The role of uncertainty in attentional and choice exploration. Psychon Bull Rev. Dec 2019;26(6):1911-1916. [doi: 10.3758/s13423-019-01653-2] [Medline: 31429060]
81. Socioeconomic indexes for areas. Australian Bureau of Statistics. URL: https://www.abs.gov.au/websitedbs/censushome.nsf/home/seifa [accessed 2023-12-27]
82. Chronic conditions and multimorbidity. Australian Institute of Health and Welfare. URL: https://www.aihw.gov.au/reports/australias-health/chronic-conditions-and-multimorbidity [accessed 2023-12-27]
83. Filkins BL, Kim JY, Roberts B, Armstrong W, Miller MA, Hultner ML, et al. Privacy and security in the era of digital health: what should translational researchers know and do about it? Am J Transl Res. 2016;8(3):1560-1580. [FREE Full text] [Medline: 27186282]
84. Corley DA, Jensen CD, Marks AR, Zhao WK, de Boer J, Levin TR, et al. Variation of adenoma prevalence by age, sex, race, and colon location in a large population: implications for screening and quality programs. Clinical Gastroenterology and Hepatology. Feb 2013;11(2):172-180. [doi: 10.1016/j.cgh.2012.09.010]
85. Earnshaw VA, Bogart LM, Dovidio JF, Williams DR. Stigma and racial/ethnic HIV disparities: moving toward resilience. Am Psychol. 2013;68(4):225-236. [FREE Full text] [doi: 10.1037/a0032705] [Medline: 23688090]
86. Datasets - CHeReL. Centre for Health Record Linkage. URL: https://www.cherel.org.au/datasets [accessed 2023-12-27]
87. Privacy act 1988. The Australian Government Federal Register of Legislation. URL: https://www.legislation.gov.au/Details/C2014C00076 [accessed 2023-12-27]
88. Health information privacy. The US Department of Health & Human Services. URL: https://www.hhs.gov/hipaa/index.html [accessed 2023-12-27]
89. Fellegi IP, Sunter AB. A theory for record linkage. Journal of the American Statistical Association. Dec 1969;64(328):1183-1210. [doi: 10.1080/01621459.1969.10501049]
90. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int J Epidemiol. Dec 2002;31(6):1246-1252. [doi: 10.1093/ije/31.6.1246] [Medline: 12540730]
91. Lujic S, Randall DA, Simpson JM, Falster MO, Jorm LR. Interaction effects of multimorbidity and frailty on adverse health outcomes in elderly hospitalised patients. Sci Rep. Aug 19, 2022;12(1):14139. [FREE Full text] [doi: 10.1038/s41598-022-18346-x] [Medline: 35986045]
92. Han J, Park J, Lee H. Analysis of the effect of an artificial intelligence chatbot educational program on non-face-to-face classes: a quasi-experimental study. BMC Med Educ. Dec 01, 2022;22(1):830. [FREE Full text] [doi: 10.1186/s12909-022-03898-3] [Medline: 36457086]
93. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2023-12-27]
94. An important next step on our AI journey. Google. URL: https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/ [accessed 2023-12-27]
95. ‘We are a little bit scared’: OpenAI CEO warns of risks of artificial intelligence. The Guardian. URL: https://www.theguardian.com/technology/2023/mar/17/openai-sam-altman-artificial-intelligence-warning-gpt4 [accessed 2023-12-27]
96. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ. Jun 06, 2023;9:e48163. [FREE Full text] [doi: 10.2196/48163] [Medline: 37279048]
97. Lembani R, Gunter A, Breines M, Dalu MTB. The same course, different access: the digital divide between urban and rural distance education students in South Africa. Journal of Geography in Higher Education. Nov 22, 2019;44(1):70-84. [doi: 10.1080/03098265.2019.1694876]
98. van de Werfhorst HG, Kessenich E, Geven S. The digital divide in online education: Inequality in digital readiness of students and schools. Computers and Education Open. Dec 2022;3:100100. [doi: 10.1016/j.caeo.2022.100100]
99. Kazeminia S, Baur C, Kuijper A, van Ginneken B, Navab N, Albarqouni S, et al. GANs for medical image analysis. Artif Intell Med. Sep 2020;109:101938. [doi: 10.1016/j.artmed.2020.101938] [Medline: 34756215]
100. Armanious K, Jiang C, Fischer M, Küstner T, Hepp T, Nikolaou K, et al. MedGAN: Medical image translation using GANs. Comput Med Imaging Graph. Jan 2020;79:101684. [doi: 10.1016/j.compmedimag.2019.101684] [Medline: 31812132]
101. Bose P, Srinivasan S, Sleeman WC, Palta J, Kapoor R, Ghosh P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. Applied Sciences. Sep 08, 2021;11(18):8319. [doi: 10.3390/app11188319]

102. Kuo NIH, Garcia F, Sonnerborg A, Bohm M, Kaiser R, Zazzi M, et al. Synthetic health-related longitudinal data with mixed-type variables generated using diffusion models. ArXiv. Preprint posted online on March 22, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.12281]
103. Kingma DP, Welling M. Auto-encoding variational Bayes. ArXiv. Preprint posted online on December 10, 2022 [FREE Full text] [doi: 10.48550/arXiv.1312.6114]
104. Sohl-Dickstein J, Weiss EA, Maheswaranathan N, Ganguli S. Deep unsupervised learning using nonequilibrium thermodynamics. ArXiv. Preprint posted online on November 18, 2015. [doi: 10.48550/arXiv.1503.03585]
105. OpenAI. GPT-4 technical report. ArXiv. Preprint posted online on December 19, 2023. 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.08774]

Abbreviations

3TC: lamivudine
ABC: abacavir
AI: artificial intelligence
ART: antiretroviral therapy
Base drug combo: base drug combination
Comp INI: complementary integrase inhibitor
EFV: efavirenz
FTC: emtricitabine
GAN: generative adversarial network
INI: integrase inhibitor
MIMIC: Medical Information Mart for Intensive Care
ML: machine learning
NNRTI: nonnucleoside reverse transcriptase inhibitor
NRTI: nucleotide reverse transcriptase
PI: protease inhibitor
pk-En: pharmacokinetic enhancer
RAL: raltegravir
RL: reinforcement learning
RPV: rilpivirine
TDF: tenofovir disoproxil fumarate
UNSW: University of New South Wales
VL: viral load

Edited by G Eysenbach, K Venkatesh, MN Kamel Boulos; submitted 30.07.23; peer-reviewed by S Seevanayanagam, M Black; comments to author 14.10.23; revised version received 20.10.23; accepted 08.11.23; published 16.01.24

Please cite as:
Kuo NIH, Perez-Concha O, Hanly M, Mnatzaganian E, Hao B, Di Sipio M, Yu G, Vanjara J, Valerie IC, de Oliveira Costa J, Churches T, Lujic S, Hegarty J, Jorm L, Barbieri S
Enriching Data Science and Health Care Education: Application and Impact of Synthetic Data Sets Through the Health Gym Project
JMIR Med Educ 2024;10:e51388
URL: https://mededu.jmir.org/2024/1/e51388
doi: 10.2196/51388
PMID: 38227356