Medical Time-Series Data Generation using Generative Adversarial Networks

Saloni Dash, Andrew Yale, Isabelle Guyon, Kristin Bennett

Medical Time-Series Data Generation using Generative Adversarial Networks

Saloni Dash[1], Andrew Yale[2], Isabelle Guyon[3], and Kristin P. Bennett[2]

[1] BITS Pilani, Department of CSIS, Goa Campus, India
[2] Rensselaer Polytechnic Inst. Troy, New York
[3] UPSud/INRIA U. Paris-Saclay, France

Abstract. Medical data is rarely made publicly available due to high de-identification costs and risks. Access to such data is highly regulated due to it's sensitive nature. These factors impede the development of data-driven advancements in the healthcare domain. Synthetic medical data which can maintain the utility of the real data while simultaneously preserving privacy can be an ideal substitute for advancing research. Medical data is longitudinal in nature, with a single patient having multiple temporal events, influenced by static covariates like age, gender, comorbidities, etc. Extending existing time-series generative models to generate medical data can be challenging due to this influence of patient covariates. We propose a workflow wherein we leverage existing generative models to generate such data. We demonstrate this approach by generating synthetic versions of several time-series datasets where static covariates influence the temporal values. We use a state-of-the-art benchmark as a comparative baseline. Our methodology for empirically evaluating synthetic time-series data shows that the synthetic data generated with our workflow has higher resemblance and utility. We also demonstrate how stratification by covariates is required to gain a deeper understanding of synthetic data quality and underscore the importance of including this analysis in evaluation of synthetic medical data quality.

Keywords: Synthetic Data · Generative Adversarial Networks · Time-Series

1 Introduction

Medical data in the form of Electronic Medical Records (EMR) has been widely used by hospitals in the United States for aiding hospital processes like quality improvement, monitoring patient safety etc. [14]. Furthermore, EMR data has also been used for advancing healthcare research for decades [13]. However, high de-identification costs and risks severely limit public access to such data. This imposes huge restrictions on data-driven clinical research and makes studies that use this data difficult to reproduce.

A generative model that samples from the distribution of the health data, while simultaneously preserving it's privacy is an ideal solution to the problem.

Generative models like Generative Adversarial Networks (GANs) [15,4] (HealthGAN, medGAN) explicitly generate snapshots of EMR type data. However, real EMR data is longitudinal in nature and falls in the domain of time-series generative modelling. A key aspect of medical data is static covariates that heavily influence temporal variables. For instance, a patient record not only contains details of hospital visits over a period of time but also static demographic details like gender, ethnicity, comorbidities etc. We characterize a good medical time-series generative model as one that jointly models the distribution of the static as well as the temporal variables.

We address this problem in the paper and provide a simple baseline that can be used for comparison against future medical time-series generative models. The primary contributions of this paper[†] are:

1. Illustration of an efficient, flexible workflow to facilitate joint modelling and synthesis of static and temporal variables.
2. Explicit qualitative evaluation of influence of static covariates on time-series variables.
3. Reproducing clinical time-series benchmarks on synthetic versions of a publicly available and widely used medical dataset.

2 Related Work

An open source synthetic patient generator called Synthea [5] uses hand-crafted modules aided by health care practitioners and statistics derived from real data to generate patients from their birth day to the present day. It does not violate any privacy restrictions because it does not use real patients to generate the data. It also claims to maintain utility as the generator uses underlying rules manually derived from the real data. However, the time-series generated for a record are not necessarily representative of real patient trajectories. Additionally, the custom designed rules severely limit the type of data that can be generated and are not easily extendable to other distributions of data.

Recurrent (Conditional) GAN (RCGAN) [8] uses recurrent neural networks (RNNs) in the GAN framework, to generate real valued time-series medical data like respiratory rate, heart rate etc. In the conditional setting, both the generator and discriminator are conditioned on labels sampled from the real data during training, and generated from independent distributions during the generation process. The labels guide the generative process but are not modelled jointly with the time-series variables.

Time-Series GAN [16] explicitly models time-series distributions as a joint distribution of static and temporal variables. It produces realistic time-series by jointly optimizing adversarial and supervised losses. We found TimeGAN to be the only time-series generative model that addresses the problem of jointly modelling static and temporal variables, and use it as a comparative baseline for our methodology.

[†]This paper is an extension of [6] submitted to the ML4H workshop at NeurIPS 2019.

3 Method

We illustrate our approach by generating time-series datasets for three time-series datasets where the covariates have a strong influence on the time-series variables. The datasets are (1) PJM Hourly Energy Consumption Dataset (Kaggle) [12],(2) Sleeping Patterns from American Time Use Survey (ATUS) [1], (3) Medical Information Mart for Intensive Care (MIMIC-III) [10].

The workflow is as follows:

1. Identify appropriate summary statistic(s) for time-series variables (e.g. mean, median, skew, count etc.)
2. Compute summary statistic(s) for fixed time-intervals over the whole time period.
3. Append summary statistic(s) to static variables. This maps the time-series data frame to a cross-sectional data frame.
4. Use a generative model of your choice to generate this transformed data.

For (1), we choose summary statistics inspired by downstream applications of the synthetic data. For (4), we use HealthGAN, a Wasserstein GAN [15], to generate the transformed data. The HealthGAN includes encoding mappings for categorical, numerical and ordinal variables of which the ordinal mappings particularly boosted our results for the ATUS dataset. We evaluate resemblance of the synthetic data to real data by assessing summary statistics conditioned on covariates and the utility by reproducing published research results.

This workflow is best suited for data where the time-series can be dissected into meaningful intervals to compute summary statistics relevant to the downstream application of synthetic data. An appropriate transformation of computing mean, variance, skew, kurtosis etc. for the whole time-series is always feasible.

We use TimeGAN[†] as a comparison for this workflow for two of the above datasets. It should be noted that we use the default parameters for TimeGAN and do not fine-tune the model while generating the data.

4 Results

4.1 PJM Hourly Energy Consumption

Our first dataset is not a medical dataset, but it provide an illustration of the challenge of synthesizing time series datasets wiith covariates. PJM Interconnection LLC (PJM) is a transmission organization (RTO) which is part of the Eastern Interconnection grid operating an electric transmission system serving specific regions of the United States. The dataset[†] primarily comprises of a datetime stamp and the average energy consumed in Mega Watts (MW) in that hour.

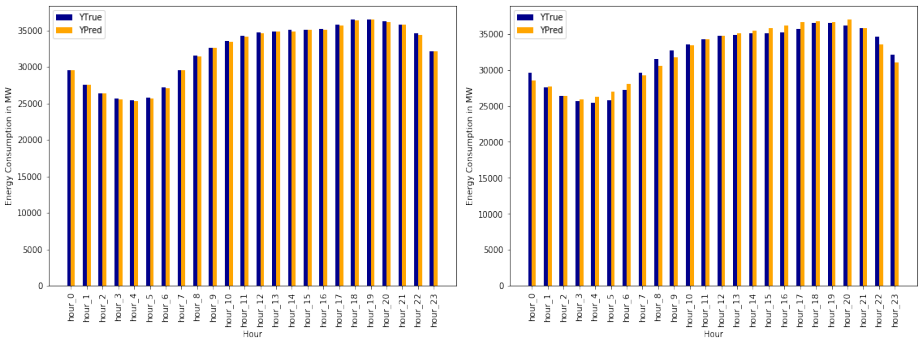
[†]We use the source code available at `https://bitbucket.org/mvdschaar/mlforhealthlabpub/src/master/alg/timegan/`

[†]`https://www.kaggle.com/robikscube/hourly-energy-consumption`

A natural summary statistic for the dataset is the average energy consumed per hour. We set the time period to be one day. We hence get twenty-four time-series statistics, one for each hour of the day which we append to the static variables of day of week and month derived from the datetime stamp. The transformed data now has twenty-six variables which are generated by HealthGAN. We also separately generate the original data using TimeGAN.

We then qualitatively evaluate the synthetic datasets by comparing trends in the real data. Figure 1 shows close resemblance of the summary statistic of average hourly energy consumption across twenty-four hours for the real and synthetic datasets (derived from HealthGAN and TimeGAN). A Welsch t-test of the samples binned by hour shows that the hourly means from HealthGAN (p-value = 0.51) as well as the hourly means from TimeGAN (p-value = 0.18) do not significantly differ from the hourly means of the real data. Both generations methods are seemingly performing well.

Figure 2 analyses the influence of the static covariates of day of week and month on the average energy consumption in the real data, which reports highest energy consumption during the weekdays in the evening hours and the summer months. These trends are mimicked in the synthetic data generated by HealthGAN. In the synthetic data generated by TimeGAN, the hourly and weekly trends are captured reasonably well but when examining by the covariate months, the peak at months 7 and 8 is missed.

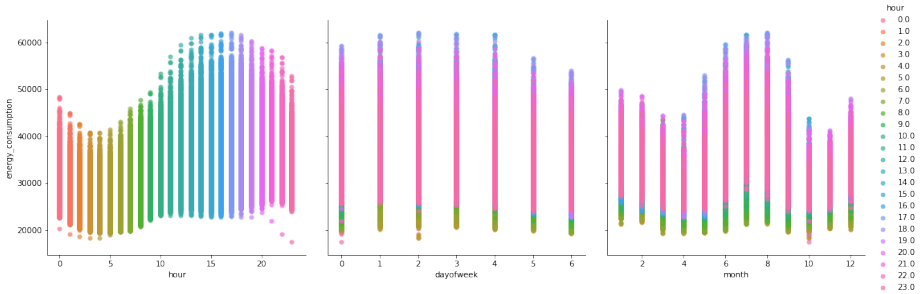


(a) HealthGAN (b) TimeGAN

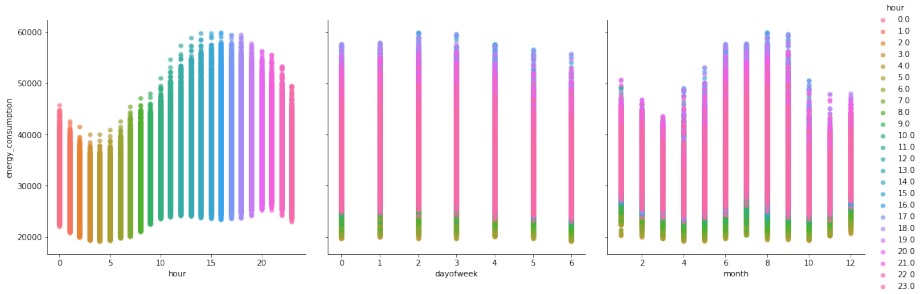
Fig. 1: **Hourly Energy Consumption - YTrue is the hourly energy consumption in the real data and YPred is the hourly energy consumption in the generated data. The values in both the real and synthetic datasets match closely.**

4.2 American Time Use Survey (ATUS)

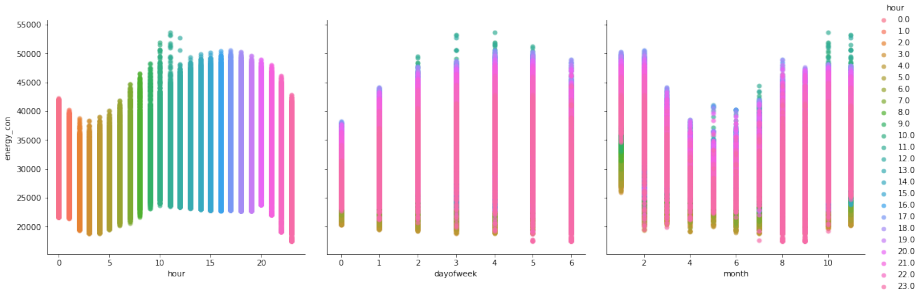
We generate sleeping patterns from American Time Use Survey (ATUS), a federally administered, annual survey on time use in the United States [1]. The



(a) Real Data



(b) Synthetic Data from HealthGAN



(c) Synthetic Data from TimeGAN

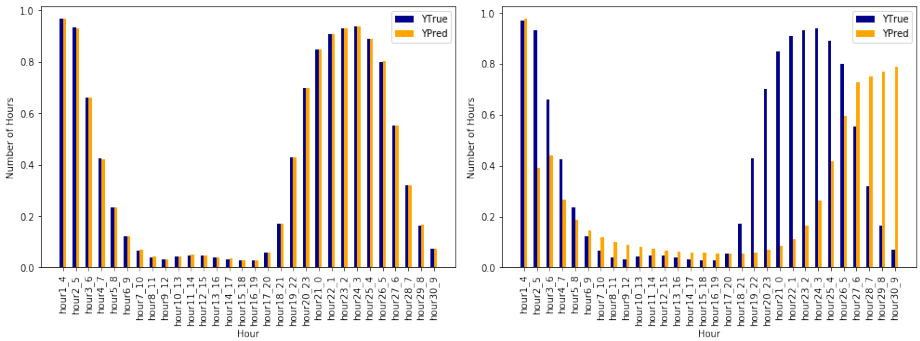
Fig. 2: **Average Energy Consumed vs Hour, Day of Week and Month for real and synthetic data sets shows HealthGAN synthetic data highly resembles real data even when covariates are considered.**

survey records how Americans divide their time among life's activities in a nationally representative sample. There are many different types of events per person. However, we choose to restrict our data to only the sleep activities of the people over a period of thirty hours. (Please refer to [6] for more details). We divide the thirty hours into thirty events of one hour each, and compute average sleep in that hour. We then append them to the static variables of age, sex, day

of the week and month of the year. The data is now in matrix form, consisting of 34 (including covariates) features per patient, ready for the HealthGAN to generate this data. Sleep data is synthesized from TimeGAN as well to use as a comparative baseline.

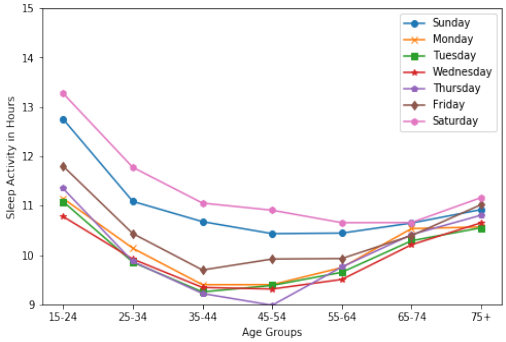
Figure 3 shows the average sleep trends in the real data and synthetic datasets. The average sleep per hour in the HealthGAN synthetic data closely resembles the real data. Most people are awake by 10:00 am and asleep by 12:00 am. The synthetic data from TimeGAN does not follow this distribution. A Welsch t-test of the sleep times binned by hour shows that the mean sleep time per hour in the data from HealthGAN (p-value = 0.58) is not statistically different from that in the real data. However, for the data from TimeGAN (p-value = 0.012), the means appear to be significantly different.

To analyse the relationship between the static covariates and time-series variables, we reproduce a sleep study [2] which analyses the average sleeping time stratified by age and day of week. Figure 4 shows the variations in sleep depending on age group and day of the week. In the real data, the average sleeping time on weekends is significantly different from that on weekdays. Overall, teenagers and young adults sleep (age group 15-24) significantly more than other age groups, whereas adults between 35-54 years in general require less sleep than other age groups. These trends are captured well in the synthetic data from HealthGAN. The variations in average sleep times binned by group and day of the week are so high in the real data that although the synthetic plot appears to be shifted by an hour, it still falls within the 95% confidence intervals of the means in the real data. For the synthetic data from TimeGAN, the distributions for days Tuesday through Saturday are missing completely, suggesting mode collapse, and the trends captured are significantly different from those in the real data.

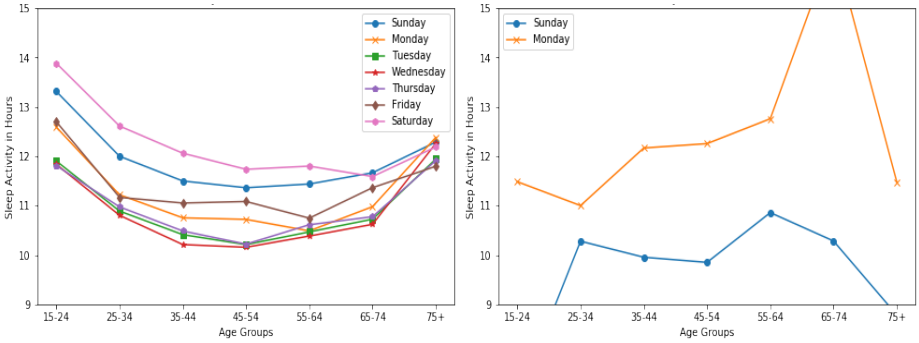


(a) HealthGAN (b) TimeGAN

Fig. 3: **Average Hourly Sleep Trends of real (blue) and synthetic data (yellow) generated by HealthGann and TimeGAN**



(a) Real Data



(b) HealthGAN

(c) TimeGAN

Fig. 4: **Average hours of sleep grouped by age and day of the week for real and synthetic HealthGAN and TimeGAN data**

4.3 MIMIC - III

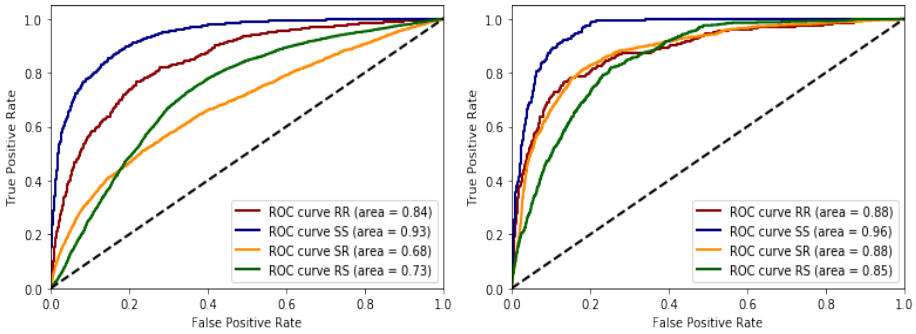
The MIMIC - III dataset [10] is a publicly available critical care database which is widely used in research studies [3,11,7]. Specifically we use three clinical prediction time-series benchmarks derived from MIMIC - III [9]. These tasks consist of:

1. **In-Hospital Mortality Prediction** - Predicting In-Hospital mortality based on 48 hours of ICU data.
2. **Decompensation Prediction** - Predicting whether a patient's health will worsen over the next 24 hours.
3. **Phenotype Classification** - Predicting which of the 25 acute care conditions are present in a patient record

For each of the above tasks, the logistic regression baseline specifies which summary statistics to extract from the time-series. We attempt to reproduce these baseline results in the generated data as well. The paper identifies 17 clinical variables as primary temporal variables. For each variable, six different sample statistic features (mean, std dev, skew etc.) are computed on seven

different subsequences of a given time series (full, first 10%, first 25% etc.). Please refer to [9] for more details. In total there are 714 temporal variables. In (1) the static covariates are age, gender and the mortality label. This results in a total of 717 variables for each patient. The task is a binary classification task with the primary metric being AUC-ROC. In (2) the static variables are age, gender and decompensation label, resulting in 717 variables. This is also a binary classification task with the primary metric being AUC-ROC. The results for (1) and (2) are summarized in Figure 5. In (3) the static variables are age, gender and the 25 acute care conditions, resulting in 741 variables. This is a multi-label classification problem to predict phenotype with the primary metric being AUC - ROC for each variable treated independently. The results for the 25 prediction prediction tasks are illustrated in Figure 6.

In figures 5 and 6, RR refers to train on real test on real, RS to train on real test on synthetic, etc. RS scores indicate whether the synthetic data can be substituted for the real data for a downstream application, while SR score indicates whether the synthetic data has realistic features. Overall, across all 27 MIMIC-III tasks, we report RS and SR scores to be reasonably close to RR scores. Note, however, that the SS scores tend to usually overshoot the RR scores indicating that the generated distribution is more regular than the real data.



(a) Mortality AUC ROC (b) Decompensation AUC ROC

Fig. 5: **AUC ROC for MIMIC-III Mortality and Decompensation tasks. First letter indicates training set (Real or Synthetic). Second letter indicates testing set**

5 Conclusion and Future Work

Medical time-series data sets are characterized by both static as well as temporal variables which must be modelled jointly to generate realistic medical time-series data. We provide a simple, flexible and effective workflow to generate this kind of data. We test our methodology by synthesizing three different time-series datasets,

PHENOTYPE	AUC-ROC RR	AUC-ROC SS	AUC-ROC SR	AUC-ROC RS
Hypertension with complications	0.964	0.994	0.721	0.912
Diabetes mellitus with complications	0.929	0.976	0.749	0.765
Chronic kidney disease	0.927	0.992	0.725	0.960
Shock	0.917	0.992	0.689	0.794
Respiratory failure	0.870	0.977	0.769	0.832
Acute cerebrovascular disease	0.870	0.956	0.754	0.690
Coronary atherosclerosis and related	0.866	0.977	0.749	0.896
Septicemia (except in labor)	0.848	0.988	0.720	0.915
Acute and unspecified renal failure	0.845	0.962	0.716	0.870
Diabetes mellitus without complication	0.840	0.978	0.616	0.727
Acute myocardial infarction	0.828	0.977	0.577	0.622
Essential hypertension	0.822	0.969	0.655	0.719
Congestive heart failure	0.816	0.981	0.649	0.441
Pneumonia	0.815	0.942	0.578	0.757
Disorders of lipid metabolism	0.789	0.976	0.682	0.895
Cardiac dysrhythmias	0.780	0.975	0.622	0.631
Conduction disorders	0.765	0.966	0.591	0.644
Fluid and electrolyte disorders	0.764	0.963	0.656	0.881
Other liver diseases	0.756	0.993	0.582	0.412
Gastrointestinal hemorrhage	0.748	0.988	0.565	0.782
Chronic obstructive pulmonary disease	0.701	0.969	0.541	0.672
Complications of surgical/medical care	0.690	0.982	0.597	0.492
Other upper respiratory disease	0.690	0.956	0.521	0.551
Pleurisy; pneumothorax; pulmonary collapse	0.665	0.968	0.542	0.625
Other lower respiratory disease	0.628	0.976	0.526	0.398

Fig. 6: **Phenotype Classification Heatmap - In general, RR, RS, SR scores fall in the same range. SS scores tend to overshoot RR scores significantly, suggesting increased regularity in synthetic data as compared to the real data.**

two of which are health datasets. We empirically show that the data generated by HealthGan not only shows close univariate resemblance with the real data but also captures trends influenced by static covariates. We use a state-of-the-art benchmark[†] as a comparative baseline. We highlight the importance of evaluating synthetic medical data with respect to critical covariates and the importance of including such analysis in time-series generative models for medical data. We plan to use super-resolution multivariate GANs trained on varying length interval summaries to capture more complex EMR data in the future.

[†]We would like to thank the authors of TimeGAN for their help with implementation and evaluation details.

References

1. American time use survey. `https://www.bls.gov/tus/home.htm`. Accessed: 2019-09-10.
2. Mathias Basner, Kenneth M Fomberstein, Farid M Razavi, Siobhan Banks, Jeffrey H William, Roger R Rosa, and David F Dinges. American time use survey: sleep time and its relationship to waking activities. *Sleep*, 30(9):1085–1095, 2007.
3. Somnath Bose, Alistair E. W. Johnson, Ari Moskowitz, Leo Anthony Celi, and Jesse D. Raffa. Impact of intensive care unit discharge delays on patient outcomes: A retrospective cohort study. *Journal of Intensive Care Medicine*, 34(11-12):924–929, 2019. PMID: 30270722.
4. Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete electronic health records using generative adversarial networks. *CoRR*, abs/1703.06490, 2017.
5. MITRE Corporation. Synthetic patient generation. `https://synthetichealth.github.io/synthea/`. Accessed: 2019-05-16.
6. Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, Andrew Yale, and Kristin P Bennett. Synthetic event time series health data generation. *arXiv preprint arXiv:1911.06411*, 2019.
7. Rodrigo Octávio Deliberato, Stephanie Ko, Matthieu Komorowski, MA Armengol de La Hoz, Maria P Frushicheva, Jesse D Raffa, Alistair EW Johnson, Leo Anthony Celi, and David J Stone. Severity of illness scores may misclassify critically ill obese patients. *Critical care medicine*, 46(3):394–400, 2018.
8. Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.
9. Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1), Jun 2019.
10. Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
11. Sharukh Lokhandwala, Ned McCague, Abdullah Chahin, Braiam Escobar, Mengling Feng, Mohammad M Ghassemi, David J Stone, and Leo Anthony Celi. One-year mortality after recovery from critical illness: A retrospective cohort study. *PloS one*, 13(5), 2018.
12. Rob Mulla. https://www.kaggle.com/robikscube/hourly-energy-consumption.
13. Amy Harris Nordo, Hugh P. Levaux, Lauren B. Becnel, Jose Galvez, Prasanna Rao, Komathi Stem, Era Prakash, and Rebecca Daniels Kush. Use of ehrs data for clinical research: Historical progress and current applications. *Learning Health Systems*, 3(1):e10076, 2019. e10076 LRH2-2018-04-0019.R3.
14. Sonal Parasrampuria and Jawanna Henry. Hospitals' use of electronic health records data, 2015-2017, 2019.
15. Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Privacy preserving synthetic health data. In *Proceedings of the 27. European Symposium on Artificial Neural Networks ESANN*, pages 465–470, 2019.
16. Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 5509–5519, 2019.