Synthetic Acute Hypotension and Sepsis Datasets Based on MIMIC-III and Published as Part of the Health Gym Project

Nicholas I-Hsien Kuo[1,†], **Simon Finfer**[2,3,4], **Louisa Jorm**[1], **Sebastiano Barbieri**[1]

[1]Centre for Big Data Research in Health, University of New South Wales, Sydney, Australia
[2]The George Institute for Global Health, Sydney, Australia
[3]University of New South Wales, Sydney, Australia
[4]Imperial College London, London, United Kingdom

[†]n.kuo@unsw.edu.au

Abstract

These two synthetic datasets comprise vital signs, laboratory test results, administered fluid boluses and vasopressors for $3,910$ patients with acute hypotension and for $2,164$ patients with sepsis in the Intensive Care Unit (ICU). The patient cohorts were built using previously published inclusion and exclusion criteria and the data were created using Generative Adversarial Networks (GANs) and the MIMIC-III Clinical Database. The risk of identity disclosure associated with the release of these data was estimated to be very low (0.045%). The datasets were generated and published as part of the *Health Gym*, a project aiming to publicly distribute synthetic longitudinal health data for developing machine learning algorithms (with a particular focus on offline reinforcement learning) and for educational purposes.

1 Background

Due to their highly confidential nature, clinical data can usually not be shared without establishing formal collaborations and executing extensive data use agreements. This hampers the development of robust machine learning algorithms for healthcare and the use of clinical data for educational purposes. One approach to overcome these barriers consists of generating synthetic data that closely resembles the original dataset but does not allow re-identification of individual patients and can therefore be freely distributed.

We publish two synthetic but realistic datasets related to patients with acute hypotension and with sepsis in the Intensive Care Unit (ICU). The datasets were created using *Generative Adversarial Networks* (GANs) (Goodfellow et al., 2014; Gulrajani et al., 2017) and the MIMIC-III Clinical Database (Johnson et al., 2016). Two patient cohorts were identified within MIMIC-III and used to generate the synthetic data: $3,910$ patients with acute hypotension (Gottesman et al., 2020) and $2,164$ patients with sepsis (Komorowski et al., 2018), with related timeseries of vital signs, laboratory test results, medications (*e.g.,* administered fluid boluses and vasopressors), and demographics.

The datasets were generated and published as part of the *Health Gym*, a project aiming to publicly distribute synthetic longitudinal health data for developing machine learning algorithms (with a particular focus on offline reinforcement learning) and for educational purposes. The datasets are highly realistic (a publication detailing the generation and quality assurance process is currently in preparation) and here we report on the risk of identity disclosure associated with the release of these data, using current best practices (Goncalves et al., 2020; El Emam et al., 2020).

2 Methods: Identity Disclosure Risk

The MIMIC-III Clinical Database contains only non-identifiable data; however, there is a small remaining risk of sensitive information being disclosed if an adversary is able to link the published synthetic data to specific records in MIMIC-III. A two step process is used to assess this risk. In the first step we verify that no data is simply copied by the GAN from the real training dataset to the generated synthetic dataset. This is done by ensuring that the Euclidean distance between any record (*i.e.,* all variables recorded at a specific point in time for an individual) in the real dataset and any record in the synthetic dataset is greater than zero.

In the second step we compute the probability of successfully gaining additional information about an individual by matching records in the synthetic dataset with individuals in the population used to sample the real dataset, following the approach by El Emam et al. (2020). An adversary may have access to partial information (*quasi-identifiers* such as age and gender) about individuals in the population and may attempt to determine whether additional information about an individual can be gained from the synthetic dataset (population-to-sample attack), or whether an individual in the synthetic dataset can be matched to an individual in the population (sample-to-population attack). Under the assumption that an adversary will only attempt one of these attacks, but without knowing which one, the overall probability of one of these attacks being successful is given by the maximum probability of either attack being successful (El Emam et al., 2020).

3 Content Description

This section describes the format, variables, and identity disclosure risk for the two published datasets. The patient cohorts used to generate the synthetic datasets were identified in MIMIC-III using previously published inclusion and exclusion criteria. Specifically, the cohort of patients with acute hypotension was built according to Gottesman et al. (2020) and the cohort of patients with sepsis was built following Komorowski et al. (2018).

3.1 Acute Hypotension Dataset

Variable Name	**Data Type**	**Unit**
Mean Arterial Pressure (MAP)	numeric	mmHg
Diastolic Blood Pressure (Diastolic BP)	numeric	mmHg
Systolic Blood Pressure (Systolic BP)	numeric	mmHg
Urine	numeric	mL
Alanine Aminotransferase (ALT)	numeric	IU/L
Aspartate Aminotransferase (AST)	numeric	IU/L
Partial Pressure of Oxygen (PaO2)	numeric	mmHg
Lactate	numeric	mmol/L
Serum Creatinine	numeric	mg/dL
Fluid Boluses	categorical	mL
Vasopressors	categorical	mcg/kg/min
Fraction of Inspired Oxygen (FiO2)	categorical	fraction
Glasgow Coma Scale Score (GCS)	categorical	-
Urine Data Measured (Urine (M))	binary	-
ALT or AST Data Measured (ALT/AST (M))	binary	-
FiO2 (M)	binary	-
GCS (M)	binary	-
PaO2 (M)	binary	-
Lactic Acid (M)	binary	-
Serum Creatinine (M)	binary	-

Table 1: Variables included in the acute hypotension dataset.

The acute hypotension dataset is stored as a *comma separated value* (CSV) file with a size of 23.0 MB. It includes $3,910$ synthetic patients and each patient is associated with measurements over 48 hours. There are hence $187,680$ (=$3,910 \times 48$) records (rows) in total.

The dataset contains 22 variables (columns). The first 20 variables (9 numeric, 4 categorical, and 7 binary) are listed in Table 1 and the remaining two variables contain the IDs of the synthetic patients and the timepoints. The 7 binary variables (with suffix *(M)*) indicate whether a variable was measured at a specific point in time, which in medical time series is usually highly informative.

In a reinforcement learning context, the fluid boluses and vasopressors variables can be used to define the discrete action space for managing acute hypotension, with the remaining variables defining the state space (Gottesman et al., 2020).

3.1.1 Identity Disclosure Risk

Since the acute hypotension dataset does not contain any *quasi-identifiers*, we only verified that the synthetic dataset does not contain any exact copies of records in the real dataset. Indeed the smallest Euclidean distance between any synthetic record and any real record was 49.06 (> 0).

3.2 Sepsis Dataset

The sepsis dataset is stored as a CSV file with a size of 16.2 MB. It includes $2,164$ synthetic patients with 20 time points per patient, representing 80 hours of data aggregated across 4-hour windows ($80 = 20 \times 4$). There are hence $43,280$ ($= 2,164 \times 20$) records (rows) in total.

The dataset contains 46 variables (columns). The first 44 variables (35 numeric, 3 binary, and 6 categorical) are listed in Tables 3 and 4 and the remaining two variables contain the IDs of the synthetic patients and the timepoints. Besides inherently categorical variables such as the Glasgow Coma Scale (GCS, a clinical scale between 3 and 15 used to measure a person's level of consciousness), this dataset also contains numeric variables which were categorised into deciles to simplify the data generation process (SpO2, Temp, PTT, PT, and INR).

In a reinforcement learning context, the fluid boluses and vasopressors variables can be used to define the discrete action space for managing sepsis, with the remaining variables defining the state space (Komorowski et al., 2018).

3.2.1 Identity Disclosure Risk

The synthetic dataset does not contain any exact copies of records in the real dataset (the smallest Euclidean distance between any pair of records was 328.78 (> 0)).

The sepsis dataset contains the quasi-identifiers age and gender which could be used to match records in the synthetic dataset with individuals in the population used to sample the real dataset. To compute the probability of a successful population-to-sample attack or sample-to-population attack, population statistics were determined using the entire MIMIC-III Clinical Database. The 'population' contained $248,930$ records whereas the sample of patients with sepsis (the real dataset) contained $23,882$ records.

Parameter	**Synthetic Data Risk**		**Real Data Risk**	
	Population-to-Sample	Sample-to-Population	Population-to-Sample	Sample-to-Population
Sepsis	0.044%	0.045%	0.057%	0.057%

Table 2: Indentity disclosure risk of the sepsis datasets.

The probabilities of successful attacks are listed in Table 2, for both synthetic and real datasets. These estimates are conservative since they were not adjusted for incorrect matches or for whether the adversary 'learned something new' from a match (El Emam et al., 2020).

Therefore, the publication of the synthetic sepsis dataset is associated with a maximum disclosure risk of 0.045% (*i.e.,* 0.045% probability that an individual in the synthetic dataset can be matched to an individual in the entire MIMIC-III Clinical Database). This is lower than the risk of such disclosure in the real sepsis dataset (0.057%), and far below the threshold of 9% proposed by the European Medicines Agency (2014) and Health Canada (2014) for the public release of clinical data. This is also lower than the risk threshold of 5% used in El Emam et al. (2020).

Variable Name	Data Type	Unit
Age	numeric	year
Heart Rate (HR)	numeric	bpm
Systolic BP	numeric	mmHg
Mean BP	numeric	mmHg
Diastolic BP	numeric	mmHg
Respiratory Rate (RR)	numeric	bpm
Potassium (K^+)	numeric	meq/L
Sodium (Na^+)	numeric	meq/L
Chloride (Cl^-)	numeric	meq/L
Calcium (Ca^{++})	numeric	mg/dL
Ionised Ca^{++}	numeric	mg/dL
Carbon Dioxide (CO2)	numeric	meq/L
Albumin	numeric	g/dL
Hemoglobin (Hb)	numeric	g/dL
Potential of Hydrogen (pH)	numeric	-
Arterial Base Excess (BE)	numeric	meq/L
Bicarbonate (HCO3)	numeric	meq/L
FiO2	numeric	fraction
Glucose	numeric	mg/dL
Blood Urea Nitrogen (BUN)	numeric	mg/dL
Creatinine	numeric	mg/dL
Magnesium (Mg^{++})	numeric	mg/dL
Serum Glutamic Oxaloacetic Transaminase (SGOT)	numeric	u/L
Serum Glutamic Pyruvic Transaminase (SGPT)	numeric	u/L
Total Bilirubin (Total Bili)	numeric	mg/dL
White Blood Cell Count (WBC)	numeric	E9/L
Platelets Count (Platelets)	numeric	E9/L
PaO2	numeric	mmHg
Partial Pressure of CO2 (PaCO2)	numeric	mmHg
Lactate	numeric	mmol/L
Total Volume of Intravenous Fluids (Input Total)	numeric	mL
Intravenous Fluids of Each 4-Hour Period (Input 4H)	numeric	mL
Maximum Dose of Vasopressors in 4H (Max Vaso)	numeric	mcg/kg/min
Total Volume of Urine Output (Output Total)	numeric	mL
Urine Output in 4H (Output 4H)	numeric	mL

Table 3: Numeric variables included in the sepsis dataset.

Variable Name	Data Type	Unit
Gender	binary	0=male, 1=female
Readmission of Patient (Readmission)	binary	-
Mechanical Ventilation (Mech)	binary	-
GCS	categorical	-
Pulse Oximetry Saturation (SpO2)	categorical	%
Temperature (Temp)	categorical	Celcius
Partial Thromboplastin Time (PTT)	categorical	sec
Prothrombin Time (PT)	categorical	sec
International Normalised Ratio (INR)	categorical	-

Table 4: Non-numeric variables included in the sepsis dataset.

Besides the standard security setup outlined in El Emam et al., we further consider the situation where an attacker has access to both the MIMIC-III database and the patient selection criteria[1]. In this situation, we treat the real sepsis dataset as the population and treat the synthetic sepsis dataset as the sample. The new sample-to-population (synthetic-dataset-to-real-dataset) risk is 0.796%. This value is notably larger than the synthetic-dataset-to-database risk of 0.045% in Table 2. The main reason is that the real dataset is much smaller than the MIMIC-III database in size, thus making it easier to randomly match patient information. Regardless, the new risk is still much lower than the standard threshold of 9% and hence the security of the synthetic dataset is still very strong.

Our readers should further note that it is not meaningful to consider the population-to-sample risk when an attacker has access to both the database and the patient selection criteria. When both information are present, the attacker already has everything required to identify the vulnerable patients and thus nothing extra can be learned from further matching the real dataset to the synthetic dataset.

4 Usage Notes

The two datasets mentioned in this paper will be made available through PhysioNet (Goldberger et al., 2000), a repository that hosts and shares medical data managed by the MIT Laboratory for Computational Physiology. In addition, the datasets will be released under the PhysioNet Credentialed Health Data License 1.5.0.

[1] We thank our reviewer for pointing out the significance of this additional security validation.

References

Khaled El Emam, Lucy Mosquera, and Jason Bass. **Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation**. *Journal of Medical Internet Research*, 2020.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. **PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals**. *Circulation*, 2000.

Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. **Generation and Evaluation of Synthetic Patient Data**. *BMC Medical Research Methodology*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. **Generative Adversarial Nets**. In *the Advances in Neural Information Processing Systems*, 2014.

Omer Gottesman, Joseph Futoma, Yao Liu, Sonali Parbhoo, Leo Celi, Emma Brunskill, and Finale Doshi-Velez. **Interpretable Off-Policy Evaluation in Reinforcement Learning by Highlighting Influential Transitions**. In *the International Conference on Machine Learning*, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. **Improved Training of Wasserstein GANs**. In *the Advances in Neural Information Processing Systems*, 2017.

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*, 2016.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. **The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care**. *Nature Medicine*, 2018.

European Medicines Agency. **European Medicines Agency Policy on Publication of Clinical Data for Medicinal Products for Human Use**. 2014. URL http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.

Health Canada. **Guidance Document on Public Release of Clinical Information**. 2014. URL https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html.