The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures

Article *in* BMC Medical Informatics and Decision Making · March 2019

DOI: 10.1186/s12911-019-0793-0

CITATION
1

READS
179

5 authors, including:

Junqiao Chen
Evolent Health
18 PUBLICATIONS **28** CITATIONS
SEE PROFILE

David Chun
Cerner Corporation
6 PUBLICATIONS **2** CITATIONS
SEE PROFILE

BMC Medical Informatics and Decision Making

RESEARCH ARTICLE Open Access

The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures

Junqiao Chen[1], David Chun[2*], Milesh Patel[1], Epson Chiang[1] and Jesse James[1]

Abstract

Background: Clinical data synthesis aims at generating realistic data for healthcare research, system implementation and training. It protects patient confidentiality, deepens our understanding of the complexity in healthcare, and is a promising tool for situations where real world data is difficult to obtain or unnecessary. However, its validity has not been fully examined, and no previous study has validated it from the perspective of healthcare quality, a critical aspect of a healthcare system. This study fills this gap by calculating clinical quality measures using synthetic data.

Methods: We examined an open-source well-documented synthetic data generator Synthea, which was composed of the key advancements in this emerging technique. We selected a representative 1.2-million Massachusetts patient cohort generated by Synthea. Four quality measures, Colorectal Cancer Screening, Chronic Obstructive Pulmonary Disease (COPD) 30-Day Mortality, Rate of Complications after Hip/Knee Replacement, and Controlling High Blood Pressure, were selected based on clinical significance. Calculated rates were then compared with publicly reported rates based on real-world data of Massachusetts and United States.

Results: Of the total Synthea Massachusetts population ($n = 1,193,439$), 394,476 were eligible for the "colorectal cancer screening" quality measure, and 248,433 (63%) were considered compliant, compared to the publicly reported Massachusetts and national rates being 77.3 and 69.8%, respectively. Of the 409 eligible patients, 0.7% of died within 30 days after COPD exacerbation, versus 7% reported in Massachusetts and 8% nationally. Using an expanded logic, this rate increased to 5.7%. No Synthea residents had complications after Hip/Knee Replacement (Massachusetts: 2.9%, national: 2.8%) or had their blood pressure controlled after being diagnosed with hypertension (Massachusetts: 74.52%, national: 69.7%). Results show that Synthea is quite reliable in modeling demographics and probabilities of services being offered in an average healthcare setting. However, its capabilities to model heterogeneous health outcomes post services are limited.

Conclusions: Synthea and other synthetic patient generators do not currently model for deviations in care and the potential outcomes that may result from care deviations. To output a more realistic data set, we propose that synthetic data generators should consider important quality measures in their logic and model when clinicians may deviate from standard practice.

Keywords: Synthetic clinical data, Clinical quality measures, Model validation, Synthetic patient data

* Correspondence: dchun2@gmu.edu
[2]Health Administration and Policy, George Mason University, 4400 University Drive, Fairfax, Virginia 22030, USA
Full list of author information is available at the end of the article

Background

Clinical data synthesis is an emerging technique that has the potential to boost clinical research, system implementation and training, while protecting patient privacy [1]. However, its validity has not been fully examined, which poses questions for its broader adoption.

Access to data is essential for research, implementation and training across disciplines. However, obtaining real-world data can be costly and often presents ethical challenges such as privacy concerns. This is particularly challenging in healthcare, where health records contain highly sensitive information and are strictly protected by laws and organizational policies [1].

To circumvent these challenges, some organizations and individuals have developed different approaches to synthesize clinical data. These approaches are usually based on some probability-based logic and completely bypass the use of real patient-level data. By doing so, it imposes no risk for revealing personally identifiable information. Example products include Patient Generator [2], EMRbots [3], and Synthea [1]. Synthea, for example, emphasizes the use of publicly available health statistics (e.g., census) and clinical guidelines, and attempts to make the synthetic data sufficiently realistic but not real [1].

Before wider use of synthetic data, its validity needs to be tested, e.g. how closely synthetic data is equivalent to real data [4]. Researchers in this field usually referred to this property as the "realism" of synthetic data [5], and researchers in the broader simulation community called it as operational validity and the process of assuring this property as operational validation [6], which consists of a variety of methods. A recent article [5] found that there was no consensus on the methods most appropriate for operational validation of synthetic clinical data, and only few studies have actually done it. This is not a unique shortcoming of synthetic clinical data. Earlier review found that validation of healthcare simulation in general was lacking [6, 7], which consists of a variety of methods.

Because quality of care is one of the primary goals and characteristics of a healthcare system [8], we consider it critically important to have synthetic data presenting the same level of care quality as real data. Therefore, we will use clinical quality measures to validate the synthetic clinical data. Quality measures are evidence-based metrics to quantify the processes and outcomes of healthcare. They are widely used to indicate the level of effectiveness, safety and timeliness of the services that a healthcare provider or organization offers [9].

After reviewing the three synthetic data products, we decided to focus on Synthea because it is open-source, well-documented in peer-reviewed journal articles and online documentation. Patient Generator is a commercial product that builds its core modules based upon Synthea, so our understanding of Synthea would largely apply to Patient Generator. We determined EMRBots as ineligible for our study because, as described later, most of the quality measures chosen focus on health outcomes, which is not an aspect that EMRBots considers in its design. According to the creator of EMRBots, this is because it doesn't model time-dependent interactions between patient factors and clinical outcomes [3]. Synthea models care processes after clinical guidelines and models care outcomes after literature and clinical expertise. Synthea currently models 38 clinical conditions and their progressions; simulating patient-provider encounters, lab data, medication prescription and more [1]. In this aspect, we find Synthea to be the most comprehensive, open-source synthetic patient generator that is freely available for our validation study. Quality measures might be effective to uncover some unrealistic aspects of Synthea because Synthea models mainly after clinical guidelines, which describe what ideally should happen, while quality measures are "the other side of the coin" to spotlight suboptimal care. So far, although it has been suggested [10], quality measures have never been used in the existing few validation studies [5]. We present the first study using this method.

By doing this study, we hope to contribute to the healthcare community from three perspectives. From the perspective of synthetic data developers, we hope to provide an external validation on a representative product, shedding lights on potential areas of improvement. From the perspective of synthetic data users (researchers, system implementers, teachers, health policy developers), we hope to provide some insights on for what use cases that synthetic data would be a reliable replacement of real data and for what use cases it is not. Lastly, from the perspective of the broader healthcare community, we argue that improving a general-purpose synthetic data generator such as Synthea would essentially improve our understanding of how the healthcare system works. As mentioned above, healthcare quality is an important pillar of a healthcare system. To explain any difference between the quality scores derived from synthetic data and the ones from real-world data could help us better understand the contributing factors to real-world healthcare quality.

Methods

The SyntheticMass dataset

The dataset we used is called SyntheticMass, which contains more than one million "synthetic residents" of Massachusetts pre-generated using Synthea and ready for free download [11]. The goal of this synthetic population was to statistically mirror the Massachusetts population regarding demographics, disease burdens,

vaccinations, medical visits and social determinants [11]. To achieve this goal, the Synthea model was initiated by real demographics data of Massachusetts residents on the census track, town and county levels. Demographic variables included population, percentage of difference races, median age, median household income, and percentage of college graduates. After the synthetic patients was created, they went through their clinical journeys per disease modules. Explained in detail below, a disease module essential simulates patients' through a series of clinical processes per recommendations from clinical guidelines and projects their care outcomes per findings from literature or input from clinical experts. If a disease module is set up correctly, it should imitate real-world health care phenomena, including the quality of care.

We found this population to be the most appropriate for our study because of two reasons. Firstly, it attempts to mimic the characteristics of the entire population of Massachusetts, which would make our quality measure results comparable to those that are publicly reported. Secondly, because Synthea adopts Monte Carlo simulation technique, it generates a slightly different population every time the software is run [10]. Using a large, representative, pre-generated population on the other hand, would facilitate other researchers to replicate our work.

SyntheticMass dataset contains a series of tables to mimic typical information from an electronic health record system. Within these tables, we mainly focused on the "encounter", "condition" and "patient" tables. The encounter table entails patients' encounters to health facilities, such as the service date, encounter type and principal diagnosis. The condition table provides information on onset and end dates for clinical conditions (signs, symptoms and diagnoses). The condition table accumulates all identifiable conditions that a patient has, even a condition is not the main reason why a patient seeks care in a particular encounter. The patient table provides demographic information such as identifiers, address, birth date, death date (if applicable), and gender. Modeled after electronic health records, diagnoses and procedures in Synthea are coded using Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED-CT).

Disease modules and quality measures

We started with selecting quality measures relevant to the clinical modules available in Synthea. Then we obtained the publicly reported rates of the selected measures for Massachusetts and United States as our real-world reference. We then obtained the specifications of those measures and calculated rates using SyntheticMass datasets, so that we could compare the results from real data to those from synthetic data.

A clinical module is the basic unit in Synthea to model "clinical" and "control" events (or "states" in technical terms) in a clinical domain. "Clinical states" effect disease progression and care, while "control states" effect flow control. Figure 1 is a simplified example of children ear infection provided by Synthea. Children have varying likelihoods of developing ear infection based on their age, which then triggers a non-urgent pediatric admission. During the admission a patient has certain chance of taking either an anti-biotic or painkiller. The example stops here but for other modules, there is usually a process to model outcomes after the treatment (e.g., certain chance of recovery). However, as discussed later, the modeling of outcomes might be indeed a shortcoming in Synthea. Currently there are 38 modules in Synthea, ranging from allergies, chronic diseases (e.g. Asthma), to social circumstances (e.g. homelessness).

Our selection criteria was that a quality measure needed to correspond to a clinical module available in Synthea, and was also publicly reported in quality reporting

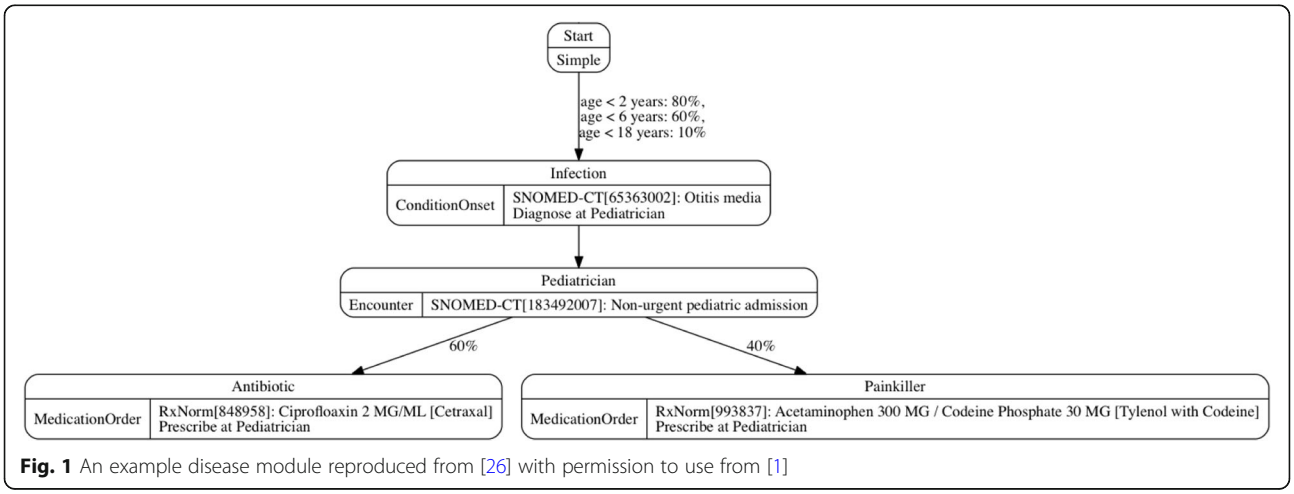

Fig. 1 An example disease module reproduced from [26] with permission to use from [1]

Table 1 Information on Selected Measures

Measure Name	Eligibility Criteria	Type of Measure	Reporting Organizations	Related Synthea Modules
Colorectal Cancer Screening	Patients between 50 and 75 years old who have had a colorectal cancer screening in a specified time frame (time frame dependent on last test taken).	Process measure	Centers for Medicare and Medicaid Services	Colorectal Cancer
COPD 30-day mortality	Patients who have had an admission with principal diagnosis of COPD exacerbation and died within 30 days of index admission.	Outcome measure	Centers for Medicare and Medicaid Services	COPD
Complications after Hip/Knee Replacement	Patients who have had either Total Knee Arthroplasty or Total Hip Arthroplasty and have had a complication within a specified time frame (time frame dependent on complication).	Outcome Measure	Centers for Medicare and Medicaid Services	Total Joint Replacement
Controlling High Blood Pressure	Patients diagnosed with hypertension that have had their blood pressure at or below the target blood pressure in subsequent blood pressure readings.	Outcome Measure	National Quality Assurance Committee	Metabolic Syndrome Disease

programs. After reviewing measures in the Healthcare Effectiveness Data and Information Set (HEDIS), Hospital Compare and Star Ratings, we found four measures eligible with Synthea's modules: Colorectal Cancer Screening, Chronic Obstructive Pulmonary Disease (COPD) 30-Day Mortality, Complications after Hip/Knee Replacement, and Controlling High Blood Pressure. Table 1 provides information on each of the selected measures. HEDIS is operated by the U.S. National Quality Assurance Committee (NCQA) and is collection of performance measures geared towards health insurance plans [12]. Hospital Compare [13] is a website operated by the U.S. federal agency Centers for Medicare and Medicaid Services (CMS) that provides performance data from participating hospitals. Star Ratings is also operated by CMS and is a rating tool that uses mostly health insurance claims data to grade Medicare Advantage plans based on quality measures and other metrics [14]. They are both public programs operated by the U.S. federal agency Centers for Medicare and Medicaid Services designed for patients to be able to compare hospitals and health plans in their area with others. Although Hospital Compare and Star Ratings use a variety of data sources, the measures that we selected for this study all use administrative claims data [15]. These measures correspond to the Colorectal Cancer, COPD, Total Joint Replacement, and Metabolic Syndrome Disease Modules in Synthea [16]. The clinical significance of each measure will be elaborated below. As a set, they cover both preventive service and chronic disease management, both ambulatory care and hospital care, and both medical service and surgical service. The first quality measure we calculate using synthetic data is Colorectal Cancer Screening, which requires patients aged 50 to 75 to have appropriate screening for colorectal cancer. This measure is important because treatment for colorectal cancer in its earliest stage can lead to high survival rate, colorectal cancer screening for adults in the 50–75 age group can help detect potentially cancerous polyps or colorectal cancer early [17]. We included patients who were alive in 2015 and 2016. Five tests were considered appropriate in this measure: Colonoscopy (recommended every 10 years), Flexible Sigmoidoscopy (every 5 years), Computed Tomography Colonography (every 5 years), Fecal Occult Blood Test (every year), and Stool DNA Test (every 3 years).

COPD 30-day mortality

The COPD 30-day mortality measure is defined as patients who died during or within 30 days of index admission with a principal diagnosis of COPD exacerbation. This measure is important because patients hospitalized after COPD exacerbations have had their mortality rate significantly affected by the quality of care given. Mortality is used because it is an indicator of the overall efficacy of more difficult to measure individual processes [17].

As shown later in the result section, the denominator identified by strictly following the measure specification is small. For sensitivity analysis, we expanded the definition and examined how that might influence the result. The strictly-defined calculation only used encounters with a principal diagnosis of COPD as part of the denominator/numerator criteria. The expanded calculation included other encounters where the COPD condition was active at the time of the encounter date.

Complication rate for hip/knee replacement

This measure looks for the occurrence of 8 complications within specific time periods after the hip or knee replacement surgery. With an aging population with high rates of osteoarthritis, Total Hip Replacement and Total Knee Replacement complications have been identified as a priority area for outcome measure development [18]. Heart attack, pneumonia, sepsis, septicemia or shock would be counted in numerator if it happens within 7 days of the admission; surgical site bleeding, pulmonary embolism, or death within 30 days; or

mechanical complications, periprosthetic joint infection or wound infection within 90 days of admission. The index admission date is defined as the encounter date with one of two SNOMED-CT codes: Total Knee Replacement (609588000) and Total Hip Replacement (52734007).

Controlling high blood pressure

The measure for controlling high blood pressure looks for patients with a diagnosis of hypertension who have had their subsequent blood pressure measurements below 140/90 mmHg (for patients aged 18–59 or 60–85 with diagnosis of diabetes) or 150/90 mmHg (for patients aged 60–85 with no diagnosis of diabetes). This measure is important for population health because hypertension increases a patient's risk for heart disease and stroke, both of which are leading causes of death in the United States [19]. In this calculation, we found patients in the synthetic population with an active condition of hypertension (SNOMED-CT: 38341003), and have had their blood pressure adequately controlled to below the defined limits after the condition onset date.

Calculated rate using Synthea

A quality measure consists of value sets and clinical logic. A value set is a list of medical codes used to determine numerator and denominator eligibility for the selected measures used value sets, a set of medical codes used in administrative claims and electronic medical records to define diagnoses and procedures, to identify numerator and denominator eligibility for each clinical quality measure. For example, a value set for influenza may contain ICD-10 codes that are related to an influenza diagnosis. The value sets for each quality measure are defined in the technical specifications of each measure, and were obtained from the Value Set Authority Center (VSAC). Mapping between coding systems or terminologies is necessary because these measures were mostly developed for administrative data, thus International Classification of Diseases Version 10 (ICD-10) and Current Procedure Terminology (CPT) was used to codify diseases and procedures, respectively. The ICD-10 and CPT codes were mapped to SNOMED-CT in accordance to documentation obtained from VSAC. When needed, we mapped ICD-10 or CPT codes to SNOMED-CT, which, as mentioned above, is the terminology used in Synthea.

After obtaining quality measure compliance rates for the SyntheticMass population, we carried out statistical bootstrapping in SPSS (IBM Armonk, New York) with 1000 resamples to obtain a 95% confidence interval for each measure.

Publicly reported rate

For the COPD 30-Day Mortality measure and Complications after Hip/Knee Replacement measure, we used publicly reported rates found on Hospital Compare as the reference population for comparison. For the Colorectal Cancer Screening and Controlling High Blood Pressure measure, the real-world rates of Massachusetts were aggregated from publicly reported Star Rating data of nine health insurance companies based in Massachusetts. National measure data was obtained by using rates from the 2017 Healthcare Effectiveness Data and Information Set report [20, 21].

Results

As shown in Table 2, SyntheticMass has a population size of around one sixth of the real Massachusetts population. Apart from average weight and BMI, the synthetic population has comparable demographic characteristics as the Massachusetts population. The likeness of demographic data between the synthetically generated population and Massachusetts population might be attributed to the use of the real-world population as a reference for calibration when developing the Synthea's data generation process. However, we observed that the SyntheticMass population is much more obese (BMI) on average compared its real-world counterpart.

Our calculation for Colorectal Cancer Screening identified 314,355 eligible members (denominator), with 215,919 of them numerator compliant (68.7%). This is below the average for Massachusetts population (77.3%) but closer to the nationally reported rate (69.8%). An interesting observation is that Synthea only modules two out of the five eligible tests in their modules: Colonoscopy and Fecal Occult Blood Test.

For the COPD Mortality 30-day measure, our calculation under strictly-defined specification returned a total of 409 encounters with a principal diagnosis of COPD during the measurement year of 2016. Of the 409 eligible admissions, three had associated patient deaths within 30 days of the index admission date (0.7%). With the concern that this logic might be too stringent, we expanded the inclusion criteria to include all admissions from patients with COPD related conditions. Using the expanded specification, the number of encounters in the denominator increased significantly to 181,458, of which 10,373 deaths occurred within 30 days. The expanded rate for this measure is 5.7%, which is lower than the national rate of 8% and state average for Massachusetts hospitals, 7% (1233/17636, range 5.2–9.3%). Another interesting observation is that only two SNOMED-CT codes that are related COPD were used in Synthea: 185086009 (Chronic Obstructive Bronchitis) and 84,733,001 (Pulmonary Emphysema), compared to the 20 ICD-10 diagnosis codes included in

Table 2 Comparison of Demographics between SyntheticMass and Real World Populations [27–29]

	SyntheticMass	Reference Population Estimates
Total Population	1,193,439	6,547,629
Age		
< 18	19.2%	21.7%
18–64	64.5%	64.5%
≥ 65	16.3%	13.8%
Average	40	39.1
Female	51%	51.6%
BMI (Adult)		
< 18.5	2%	N/A[b]
18.5–24.99	15.4%	N/A[b]
25–29.99	20%	35.6%
≥ 30	62.6%	23.3%
Male Average	31	28.7
Female Average	32	29.2
Average Height (cm)		
Male	176.77	175.77[a]
Female	163.33	161.80[a]
Average Weight (kg)		
Male	97.98	88.77[a]
Female	86.04	76.43[a]
Adults ≥20 years old with High Blood Pressure	30.92%	33.4%

[a]Nationally Reported Rate, all other rates shown reflect Massachusetts population
[b]No publically reported data found for reference

the Hospital Compare value sets, Synthea might not be sufficient to describe all the different types and nuances of COPD.

Our Hip/Knee complication measure identified zero patients who met the numerator criteria. Within the entire SyntheticMass population, only 207 synthetic patients had a hip/knee replacement during the measurement period of 2016 (denominator). None of them had an admission with heart attack or pneumonia 7 days after the procedure; or died within 30 days after the procedure. Even after expanding parameters to include any patients who had a condition onset date 7 or 30 days after procedure, our calculation yielded no patient for numerator. The Massachusetts average rate calculated from real data is 2.92% (700/23949, range 1.9–4.4%). Although 0 and 2.92% is a small difference arithmetically, it triggers serious concern over Synthea's capability to model postoperative complications. A potential explanation is that of the 8 complications, there was only one code (Pneumonia, SNOMED-CT 233604007) directly being used in Synthea. We tried every effort to identify these complications using different codes, and did try to use Myocardial Infarction (SNOMED-CT 22298006) to replace Acute Myocardial Infarction in the original specification, we still could not identify even one numerator case.

As shown in Table 3, the Colorectal Cancer Screening rate of SyntheticMass is very close to the national rate. This is expected since this measure is a procedure-based measure, and Synthea's strength is exactly that it could model the probability of certain services offered in different phases of care. However, it is much lower than the Massachusetts rate. This infers that Synthea could not model regional variation in quality. The other two outcome-based measures, on the other hand, are much more complex to model as there are many factors involved in the services that may impact the outcomes. Even with expanded logic, the rates of SyntheticMass are still much lower than Massachusetts or national rates.

Our measure for Controlling High Blood Pressure resulted in zero patients meeting numerator criteria. Of the total SyntheticMass population, 241,311 synthetic adult patients had hypertension (29.91%). We hypothesized that a certain percentage of hypertensive could have their blood pressure controlled post treatment. After analyzing Synthea's data, this doesn't seem to be the case. This may indicate that although Synthea is modeled to simulate a realistic percent of population with hypertension, it does not realistically model the outcomes that may occur post-diagnosis of hypertension as a result of intervention.

Table 3 Quality Measure Rates of SyntheticMass versus State/National Rates

Measure	SyntheticMass Rate	Massachusetts Rate	National Rate
Colorectal Cancer Screening	68.7% (68.5, 68.9%) (215,919/314,355)	77.3%	69.8%[a]
COPD 30-Day Mortality: Strict	0.7% (0, 1.7%) (3/409)	7.0% (1233/17636)	8.0%
COPD 30-Day Mortality: Expanded	4.7% (4.6, 4.8%) (8612/181,458)	7.0% (1233/17636)	8.0%
Complications of Hip/Knee Replacement	0% (0/207)	2.9% (700/23949)	2.8%
Controlling High Blood Pressure	0% (0/241,311)	74.52%	69.7[a]%

Numbers in parentheses represent lower and upper limits for 95% confidence interval, respectively
[a]Medicare PPO Reported Rate: Measure reported rates not available from HospitalCompare, rate acquired from HEDIS Medicare PPO rate. HEDIS reports rates for Medicare PPO and HMO, Medicaid PPO, Commercial PPO and HMO. Reported compliance rates for Colorectal Cancer Screening range from 58.3% to 69.8 and for Controlling High Blood Pressure range from 54.5 to 69.7%

Discussion

In this paper, we attempt to validate the realism of Synthea, a synthetic clinical data generator, by calculating clinical quality measures and comparing the results with real-world rates.

Our analysis shows that, apart from average weight and BMI, the synthetic population has comparable demographic characteristics as the Massachusetts population. We speculate two reasons behind this. Firstly, similar to the issue of hypertension (once a patient becomes hypertensive, the condition controlled), once a patient becomes obese, he or she might not lose weight, which inflates the overall obesity rate in the entire population. Secondly, it could also be due to the lack of reference data to accurately calibrate the model, or a difficulty in simulating height/weight interactions simultaneously to create an accurate BMI distribution.

Results of validation using quality measures indicate Synthea has both strengths and weaknesses in its approach. On one hand, as evident in the Colorectal Cancer Screening result, Synthea presents a high level of reliability in modeling the probabilities of certain services being offered in an average healthcare setting. On the other hand, as evident in other outcomes measures and a variance analysis between hospitals, its capabilities to model heterogeneous post-intervention health outcomes limited. We are inconclusive on Synthea's possible limitations in modeling for regional variances in quality, due to the little variance between the national and state reported rates for all measures but Colorectal Cancer Screening. This is indeed a difficult task, testified by creators of another synthetic data generator EMRBots [3].

This is the first study that uses quality measures to validate synthetic data. Results highlight the importance of incorporating quality measures in the synthesis logic, which, as far as we know, has not been considered in any existing products. The assumption that all care processes will follow clinical guidelines is not realistic. In real-world clinical practices, noncompliance with clinical guidelines is very common and diverse variations in healthcare utilization have been observed for decades [22]. In the long term, an ideal synthesis logic should account for all the factors that influence compliance with clinical guidelines, including the guideline's quality itself, clinicians' attitude to behavioral changes, an organization's resources and many more [23]. Quality measures could serve as one way to explore and verify our understanding of these factors. Such a thinking process could also apply to other aspects of realism in synthetic data besides "quality", such as "cost" or "access".

This leads to an important viewpoint on the contribution of synthetic data in general. It is not merely creating another source of data we could safely play with. By researching all the underlying mechanisms that could increase its realism, we could gradually parameterize the factors and interactions that make our health system the way it is now. Quoting Epstein's famous article *Why Model?* (Epstein, 2008, [24]), the development and calibration of a simulation model could offer explicit explanation of real-world phenomena, guide data collection, illuminate core dynamics, raise new questions and more. All of these are critical to enhance our understanding of the complexity in health care [25].

Our study has a few limitations. Firstly, we could only identify publicly reported quality measures for four clinical modules in Synthea. This might undervalue Synthea, which might present higher realism in other, less complicated, modules. Secondly, for the selected clinical modules, we only had one quality measure each for validation, which is not optimal since "quality" is a multi-faceted concept. However, the quality measures examined in this paper have been widely adopted in national quality programs and represent fundamental facets of quality in those clinical domains. We believe they are the basic ones that Synthea needs to model after to improve its realism. Thirdly, the original specification of these quality measures are designed mostly for administrative data, which is different from the clinical data (electronic health records) that Synthea generates. Although this may have an impact on the calculated rates, the differences between the synthetic rates and real-world rates are so big for three measures that we believe it could not be solely attributed to the features of different data sources.

Conclusion

In order to spread the use of synthetic clinical data, its realism needs to be tested. Clinical quality measures could serve as an effective validation tool because it is critical that synthetic data presents the same level of care quality as real data. After applying quality measures in Synthea, its strength and weakness have been uncovered, especially its limited capability to model heterogeneous health outcomes after major interventions. To improve its realism, Synthea and other synthetic data generators needs to model factors that make clinical practice deviate from standard guidelines and introduce variations in healthcare quality. Doing so could contribute to our overall understanding of the complexity in healthcare. Future validation studies should continue to identify eligible quality measures to validate new modules available in Synthea, and identify publicly reported rates based on electronic medical records. If Synthea and other synthetic data generators could be continuously improved, expanded and rigorously validated with variations in health care quality in mind, we are optimistic about the future of synthetic clinical data.

Abbreviations

CMS: Centers for Medicare and Medicaid Services; COPD: Chronic Obstructive Pulmonary Disease; CPT: Current Procedural Terminology; HEDIS: Healthcare Effectiveness Data and Information Set; ICD-10: International Classification of Diseases and Related Health Problems, revision 10; SNOMED-CT: Systematized Nomenclature of Medicine -- Clinical Terms; VSAC: Value Set Authority Center

Acknowledgements

We would like to thank Mr. Jason Walonoski for helping us understand how to use the Synthea software and datasets.

Availability for data and materials

The dataset supporting the conclusions of this article is available in the SyntheticMass website, https://syntheticmass.mitre.org/download.html.

Funding

Not Applicable.

Authors' contributions

Study Conception and Design: JC, DC. Acquisition of Data: JC, DC. Analysis and Interpretation of Data: JC, DC. Drafting of Manuscript: JC, DC. Critical Revision: JC, DC, MP, EC, JJ. All authors read and approved the final manuscript.

Ethics approval

Not Applicable.

Consent for publication

Not Applicable.

Competing interests

Evolent Health is a commercial company that engages in the provision of health care delivery and payment services. It is not a pharmaceutical company and does not sponsor clinical trials. The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Clinical Informatics, Evolent Health, Arlington, USA. [2]Health Administration and Policy, George Mason University, 4400 University Drive, Fairfax, Virginia 22030, USA.

Received: 10 October 2018 Accepted: 6 March 2019
Published online: 14 March 2019

References

1. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: an approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230–8. https://doi.org/10.1093/jamia/ocx079.
2. Pletcher, T. (2017). MiHIN annual report 2017. Retrieved from The Michigan Health Information Network Shared Services: https://mihin.org/wp-content/uploads/2017/12/Final-Annual-Report-2017.pdf
3. Kartoun, U. (2016). A methodology to generate virtual patient repositories. Computing Research Repository.
4. McLachlan S. Realism in synthetic data. Palmerston North: Massey University; 2017.
5. McLachlan S, Dube K, Gallagher T, Daley B, Walonoski J. The ATEN Framework for creating the realistic synthetic electronic health record. In: 11th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2018). Funchal: Science and Technology Publications, LDA; 2018. p. 220–30.
6. Sargent R. Verification And Validation Of Simulation Models. In: Proceedings of the 2010 Winter Simulation Conference; 2010. p. 167–83.
7. Fone D, Hollinghurst S, Temple M, Round A, Lester N, Weightman AL, Palmer S. Systematic review of the use and value of computer simulation modelling in population health and health care delivery. J Public Health Med. 2003;25(4):325–35.
8. Berwick DM, Nolan TW, Whittington J. The Triple Aim: Care,Health, and Cost. Health Aff. 2008;27(3):759–69. https://doi.org/10.1377/hlthaff.27.3.759.
9. U.S. Centers for Medicare and Medicaid Services. (2017). Quality measures. Retrieved from Quality Measures: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/QualityMeasures/index.html
10. Open Source Electronic Health Record Agent. (2018). Synthetic Patient Data Project Group. Video File. Retrieved from https://www.youtube.com/watch?v=0qGhjVPfW9w&t=794s
11. SyntheticMass. Retrieved October 10, 2018 from https://syntheticmass.mitre.org.
12. HEDIS and Performance Measurement. Retrieved October 10, 2018 from https://www.ncqa.org/hedis/.
13. What is Hospital Compare. Retrieved October 10, 2018 from https://www.medicare.gov/hospitalcompare/About/What-Is-HOS.html.
14. Star Ratings. Retrieved October 10, 2018 from https://www.medicare.gov/find-a-plan/staticpages/rating/planrating-help.aspx.
15. Centers for Medicare and Medicaid Services. (n.d.). *Hospital Compare: Data Sources*. Retrieved October 10, 2018 from Medicare.gov: The Offical U.S. Government Site for Medicare: https://www.medicare.gov/hospitalcompare/Data/Data-Sources.html
16. Hall, D. (2018). Module Gallery. Retrieved from Synthea Github: https://github.com/synthetichealth/synthea/wiki/Module-Gallery
17. National Quality Measures Clearinghouse. Chronic obstructive pulmonary disease (COPD): hospital 30-day, all-cause, risk-standardized mortality rate following COPD hospitalization. Rockville: Agency of Healthcare Quality Research. Retrieved from National Quality Measures Clearinghouse; 2017.
18. Centers for Medicare & Medicaid Services. (2018). #1550 hospital-level risk-standardized complication rate (RSCR) following elective primary total hip arthroplasty (THA). National Quality Forum.
19. Centers for Diesase Control and Prevention. (2018). About high blood pressure (hypertension). Retrieved from Centers for Diesase Control and Prevention: https://www.cdc.gov/bloodpressure/about.htmhttps://www.cdc.gov/bloodpressure/about.htm.
20. National Committee of Quality Assurance. (2017). Colorectal Cancer screening. Retrieved from health care accreditation, Health Plan Accreditation Organization - NCQA: http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2017-table-of-contents/colorectal-cancer
21. National Committee of Quality Assurance. (2017). Controlling high blood pressure. Retrieved from health care accreditation, Health Plan

Accreditation Organization - NCQA: http://www.ncqa.org/report-cards/health-plans/state-of-health-care-quality/2017-table-of-contents/controlling-high-blood-pressure
22. Wennberg J, Gittelsohn A. Small area variations in health care delivery: a population-based health information system can guide planning and regulatory decision-making. Science. 1973;182(4117):1102–8.
23. Quaglini S. Compliance with clinical practice guidelines. Studies in Health Technology and Informatics. 2008;139:160–79.
24. Epstein JM. Why model? Journal of Artificial Societies and Social Simulation. 2008;11(4):12.
25. Plsek PE, Greenhalgh T. The challenge of complexity in health care. BMJ. 2001;323:625.
26. Generic Module Framework. (2018). Retrieved from Synthea Github: https://github.com/syntheticheath/synthea/wiki/Generic-Module-Frameworkhttps://github.com/syntheticheath/synthea/wiki/Generic-Module-Framework.
27. U.S. Census Bureau. (2018). 2016 American community survey 1-year estimates. Retrieved from American Fact Finder: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_1YR_S0101&prodType=table
28. Centers for Disease Control and Prevention. (2017). Massachusetts state nutrition, physical activity, and obesity profile. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/nccdphp/dnpao/state-local-programs/profiles/massachusetts.html
29. National Center of Health Statistics. Health, United States, 2016: with Chartbook on long-term trends in health. Hyattsville: Centers for Disease Control and Prevention; 2017.

Advancing informatics with electronic medical records bots (EMRBots)

Article · September 2019

DOI: 10.1016/j.simpa.2019.100006

1 author:

Uri Kartoun
IBM Research
84 PUBLICATIONS **240** CITATIONS

SEE PROFILE

Software Impacts 2 (2019) 100006

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Advancing informatics with electronic medical records bots (EMRBots)

Uri Kartoun

Center for Computational Health, IBM Research, Cambridge, MA, USA

ARTICLE INFO

Keywords:
Simulated medical records
Synthetic medical data
Electronic medical records
Machine-learning
Data-mining
Artificial intelligence

ABSTRACT

Electronic medical records (EMRs) contain sensitive and detailed documentation on a variety of conditions at the individual level. Because EMRs are subject to confidentiality requirements, access to them is limited. In an attempt to address privacy limitations, knowledge-driven experimental artificially generated electronic medical records (EMRBots) have been introduced. EMRBot repositories have been used in a variety of scenarios to advance teaching, enhance student dissertations, facilitate hackathons, and produce R packages. In addition to describing its methodology, the manuscript reviews EMRBot use cases published by independent researchers.

Code metadata

Current Code version	*v1.0*
Permanent link to code / repository used of this code version	https://github.com/SoftwareImpacts/SIMPAC-2019-8
Legal Code License	*CC BY 4.0*
Code Versioning system used	*none*
Software Code Language used	*C#, SQL*
Compilation requirements, Operating environments & dependencies	*All operating systems.*
If available Link to developer documentation / manual	https://arxiv.org/abs/1608.00570
Support email for questions	uri.kartoun@ibm.com

1. Impact overview

Sampling data by combining distributions of events (e.g., admission dates, laboratory dates), distributions of values (e.g., admission lengths of stay, laboratory values) and distributions of covariates (e.g., admission primary diagnoses) has helped generate confidentiality-free virtual patient repositories. The aim of this manuscript is to review their use as published by independent researchers. Unlike other methods, the proposed methodology not only generates detailed data types (e.g., demographics, admissions, chief complaints, comorbidities, laboratory values) but is also ultimately invulnerable in terms of confidentiality because it does not rely on real data elements pulled from an existing electronic medical record (EMR) repository; therefore, it is not associated with privacy concerns regarding individuals' sensitive data.

In April 2015, the three generated cohorts (i.e., 100 [1]; 10,000 [2]; and 100,000 [3] virtual patients) were made publicly available for download at the dedicated website www.emrbots.org [4]. Since its launch, the site has been visited by more than 11,500 individuals; approximately 6000 visitors registered with their full names, institution names, and e-mail addresses. To date, the cohorts have been used by the scientific community as a primary data source to publish 4 journal manuscripts [5–8] and 9 conference papers [9–17]. Furthermore, the cohorts served as the primary data source to form 3 PhD dissertations [18–20]. Other than universities, institutions who have used the cohorts include start-up and pharmaceutical companies as well as governmental agencies, such as the Centers for Disease Control and Prevention, U.S. Consumer Product Safety Commission, U.S. Department of Energy, National Institute of Standards and Technology, and the National Institutes of Health. Internet traffic analytics for the website are available in Supplementary Content 1.

The EMRBots patient repositories can be instantaneously downloaded to allow the use of EMR-like data. Use of the repositories does not require installation of any software. The repositories contain raw data only and are provided as textual flat files; thus, they are independent of any specific operating system and may be used with a wide variety of database systems. Further, the repositories can be used even on low-performance computers, as they can be run on

E-mail address: uri.kartoun@ibm.com.

https://doi.org/10.1016/j.simpa.2019.100006
Received 21 June 2019; Received in revised form 24 August 2019; Accepted 4 September 2019
Available online xxxx

open-source software products, such as R and MySQL, which have minimal installment requirements. With no confidentiality restrictions, the cohorts can be distributed to a class of any size. For example, faculty members could distribute the records to a class of 100 students, all of whom could then use them to practice algorithms and build models.

2. Use of EMRBots to advance the development of novel computational methods

A novel long short-term memory (LSTM) unit, the time-aware LSTM (T-LSTM), which can handle reliably irregular elapsed times between successive elements of sequential data, was developed by incorporating EMRBots [12]. The T-LSTM, proven superior to the widely used long short-term memory neural network [21], has received substantial attention from researchers since becoming publicly available in August 2017 [e.g.,22–28]. The coauthors of citing papers are affiliated with a variety of universities as well as commercial organizations (e.g., Adobe Research, DeepMind, Google Research, IBM Research, Microsoft Research).

The codevelopers of the T-LSTM reported further on the use of EMRBots to develop health-time-aware-modulars (Health-ATM), a deep architecture to uncover patient health information given the heterogeneity of EMR data [13]. The authors compared the performance of Health-ATM with its reduced models, including the recurrent neural network (RNN) [29], convolutional RNN, attentive convolutional RNN, and time-aware convolutional RNN, as well as with other commonly used methods, including logistic regression, RETAIN [30], CNN NIPS [31], and CNN SDM [32]. Results of disease prediction from both real-world and artificial data demonstrate the superior performance of Health-ATM in comparison to all baselines. The broad attention of the scientific community to T-LSTM and the high performance achieved by Health-ATM prove the usefulness of EMRBots. Moreover, the reported uses highlight EMRBots' potential to further serve as a fundamental building block for the development of any new computational method focused on EMR, with additional examples found in [e.g.,6–8,10,11,15,17].

3. Use of EMRBots to advance education

Members of the Center for Research Informatics at the University of Chicago have used EMRBots to teach statistics and machine learning [33]. Their workshop documented in a well-detailed tutorial illustrated a use case of identifying risk factors for malignant neoplasm and was split into 4 tasks; all of the tasks required the students to program using R. The first task was to identify the outcome (having the disease) using ICD-10-CM codes. This was achieved by using basic R commands (e.g., reading data into a data frame) as well as by applying more advanced functionalities, such as mutating, filtering, selecting, and merging. The second task included the identification of patient admissions with the outcome as well as patient admissions that were not associated with the outcome. The students had to apply R code to extract features such as patient age, gender, and length of stay. Furthermore, the workshop taught basic statistics, such as applying the t-test on the mean age of populations, Chi-squared to compare gender prevalences, and the nonparametric Wilcoxon rank sum test to compare medians of length of stay. The third task included the extraction of features relative to each admission. In addition to age, lengths of stay, and gender, features included a representative set of laboratory tests (albumin, hematocrit, sodium). As a final task, the students were taught how to apply logistic regression considering the covariates and outcomes that were defined in the preceding tasks. In March 2019, Mayampurath and Johnson's tutorial applied to EMRBots was used to enhance "Computationally-Enabled Medicine", a course directed by Prof. Paul Avillach and Prof. Isaac Kohane of the Department of Biomedical Informatics in Harvard Medical School [https://github.com/kartoun/IBM-Harvard-Workshop/].

Another example for enhancing education with EMRBots is the Recommender Systems course given by Prof. Kostas Stefanidis at the Faculty of Natural Sciences of the University of Tampere, Finland [34]. Prof. Stefanidis and colleagues have further used the repositories to develop a model for group recommendations incorporating the notion of fairness, and additionally define a novel approach to measure similarities between patients [35].

The courses specified above illustrate how EMRBots data sets enhance education. Due to the lack of confidentiality and privacy requirements, the lecturers could analyze data in class and distribute data [1–3], source code [36], and slides among students. Furthermore, the courses' organizers provided complete slide decks of their lectures, including code samples. These resources combined with actual EMRBots data could be useful for lecturers and students at other institutions. Moreover, given the lack of concern regarding distributing related content, the EMRBots data sets could also be used beyond lectures, for example, to enhance hackathons. A first use of EMRBots in a hackathon was reported by Carnegie Mellon University in April 2018 [37].

In addition to offering students hands-on experience with EMR-like data, the repositories have been crucial to advancing PhD dissertations. For example, the repositories were used to evaluate the performance of a new type of cloud platform [18], an effort that yielded a publication at the *IEEE HealthCom* conference [9]. EMRBots were also used to assess PRIIME, a framework for interactive personalized interesting pattern discovery [19]. Experiments in the dissertation included the evaluation of PRIIME with EMRBots and the comparison of results with several publicly available resources not related to health care. The PhD resulted in a publication at the *IEEE International Conference on Big Data* [14].

EMRBots can also be made available for integration with open source packages developed by the scientific community. Pioneering such use yielded the creation of the new R package "comoRbidity", which is capable of providing analyses of disease comorbidities from both clinical and molecular perspectives. The package yielded the creation of a manuscript published in *Bioinformatics* [5].

4. Data records

Table 1 summarizes several characteristics of the largest cohort, which comprises 100,000 virtual patients. Each virtual patient in the cohort is associated with 1 to 10 admissions; each admission lasts from 1 to 20 days and is associated with a single chief complaint. All values are associated with a date and time. Each admission record also includes multiple measurements of common laboratory tests [4]. The number of admissions per patient, the length of stay per admission, and the laboratory values are randomly generated; however, they are sampled from predefined ranges of values. For example, a patient's age could not exceed 95 years as of January 1, 2015, and a glucose measurement could only be in the range of 60–140 mg/dL. In total, the database contains 1.4 GB of data, representing 100,000 virtual patients associated with 361,760 admissions and 107,535,387 total laboratory measurements.

5. Limitations

Although the cohorts are useful for practicing or developing new machine-learning algorithms, they cannot serve as resources to assess real patient outcome scenarios (e.g., 30-day readmission prediction or disease prognosis) because their creation process does not take into account the complex time-dependent interactions between the factors associated with real patients. Developing algorithms to create virtual patient repositories that reliably mimic real EMR is a tremendous challenge because it necessitates approaches that populate databases with a combination of linear and nonlinear associations between all medical elements, as well as with random associations. The algorithms required would comprise both individual-level and population-level assumptions and apply an intelligent functionality to assign acceptable temporal

Table 1
A virtual patient repository of 100,000 patients.

Variable and category	Patients (n = 100,000)
Mean age as of 1/1/2015, years (SD)	57.8 (17.3)
Gender (%)	
Female	52.0
Ethnicity (%)	
White	49.0
Asian	23.0
African American	15.0
Unknown	13.0
Mean number of admissions per patient (SD)	3.6 (1.5)
Mean length of stay, days (SD)	11.0 (5.2)
% Population with length of follow-up (years)	
0–9	13.1
10–15	9.3
>15	77.6
Population below poverty (%)	21.6
Comorbidities; Prevalence (%)	
Malignant neoplasm	41.4
Rheumatoid arthritis	25.6
Diabetes (type I or II)	24.4
Renal complications	17.0
Coronary artery disease	7.0
Laboratory values (Mean; SD)	
Blood urea nitrogen (mg/dL)	17.5; 7.2
Platelets (k/cumm)	284.9; 95.3
Creatinine (mg/dL)	0.9; 0.2
Albumin (gm/dL)	4.2; 1.0
Lymphocytes (%)	25; 5.8

differences among all medical events. For instance, a virtual patient repository of nonalcoholic fatty liver disease patients would need to include assumptions about inverse correlations of albumin levels and sodium levels with cardiovascular disease [38], while virtual congestive heart failure (CHF) patients would need to be associated with a high prevalence of diuretic use, advanced age, and a high prevalence of associated comorbidities, such as renal failure [39].

Code availability

Full access to the source code is available [36]: https://github.com/kartoun/emrbots/. All users are welcomed to use the databases and code and apply modifications. The databases and code should primarily be used to publish scientific manuscripts, enhance open source libraries, facilitate hackathons, and lecture. In addition to a direct download of the databases [1–3], documentation of the source code is available in Supplementary Content 2.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The project was self-funded by the author (Uri Kartoun PhD).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.simpa.2019.100006.

References

[1] U. Kartoun, EMRBots: A 100-patient database - Figshare, 2018, http://figshare.com/articles/A_100-patient_database/7040039.
[2] U. Kartoun, EMRBots: A 10,000-patient database - Figshare, 2018, http://figshare.com/articles/A_10_000-patient_database/7040060.
[3] U. Kartoun, EMRBots: A 100,000-patient database - Figshare, 2018, http://figshare.com/articles/EMRBots_a_100_000-patient_database/7040198.
[4] U. Kartoun, A methodology to generate virtual patient repositories, 2016, CoRR abs/1608.00570.
[5] A. Gutiérrez-Sacristán, À. Bravo, A. Giannoula, M.A. Mayer, F. Sanz, L.I. Furlong, Comorbidity: An R package for the systematic analysis of disease comorbidities, Bioinformatics (2018) http://dx.doi.org/10.1093/bioinformatics/bty315.
[6] M. Nithya, T. Sheela, Predictive delimiter for multiple sensitive attribute publishing, Cluster Comput. (2018) http://dx.doi.org/10.1007/s10586-017-1612-y.
[7] M. Nithya, T. Sheela, Relational forecast limiter algorithm for ICD based EMRs, Int. J. Eng. Technol. 7 (334) (2018) 103–106, https://www.sciencepubco.com/index.php/ijet/article/view/18782/8582.
[8] J. Chen, D. Chun, M. Patel, E. Chiang, J. James, The validity of synthetic clinical data: A validation study of a leading synthetic data generator (Synthea) using clinical quality measures, BMC Med. Inform. Decis. Mak. 19 (1) (2019) 44, http://dx.doi.org/10.1186/s12911-019-0793-0.
[9] M. Bahrami, Singhal M., A dynamic cloud computing platform for eHealth systems, in: 2015 IEEE 17th International Conference on e-Health Networking, Applications and Services (Healthcom): Short and Demo Papers. Oct. 2015, Boston, MA, USA, 2015, pp. 14–7.
[10] S. Janaswamy, R.D. Kent, Semantic interoperability and data mapping in EHR systems, in: 2016 IEEE 6th International Conference on Advanced Computing, IACC, 27–8 Feb, 2016, Bhimavaram, India.
[11] L.S. Kumar, A. Padmapriya, Evidence based subsequent disease extraction from EMR health record by grade measure, in: 2016 Online International Conference on Green Engineering and Technologies (IC-GET) 19 Nov. 2016, Coimbatore, India.
[12] I.M. Baytas, C. Xiao, X. Zhang, F. Wang, A.K. Jain, J. Zhou, Patient subtyping via Time-Aware LSTM networks, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13–7 August, 2017, Halifax, NS, Canada.
[13] Ma Tengfei, Xiao Cao, Wang Fei, Health-ATM: A deep architecture for multi-faceted patient health record representation and risk prediction, in: Proceedings of the 2018 SIAM International Conference on Data Mining.
[14] A.B. Mansurul, M.A.I. Hasan, PRIIME: A generic framework for Interactive Personalized Interesting Pattern Discovery, in: 2016 IEEE International Conference on Big Data (Big Data), 5–8 December 2016, Washington, DC, USA.
[15] Y. Ji, C. Xu, J. Xu, H. Hu, vABS: Towards verifiable attribute-based search over shared cloud data, in: 2019 IEEE 35th International Conference on Data Engineering, ICDE, 8–11 April 2019, Macau, China.
[16] N. Li, C. Tsigkanos, Z. Jin, S. Dustdar, Z. Hu, Ghezzi C., POET: Privacy on the edge with bidirectional data transformations, in: 2019 IEEE International Conference on Pervasive Computing and Communications, PerCom, Kyoto, Japan, 2019, pp. 1–10, http://dx.doi.org/10.1109/PERCOM.2019.8767395.
[17] M. Vardalachakis, H. Kondylakis, L. Koumakis, A. Kouroubali, D. Katehakis, ShinyAnonymizer: A tool for anonymizing health data, in: International Conference on Information and Communication Technologies for Ageing Well and e-Health, ICT4AWE, 2019.
[18] M. Bahrami, A Dynamic Cloud with Data Privacy Preservation (Ph.D. dissertation), Department of Electrical Engineering and Computer Science at University of California, Merced, 2016.
[19] A.B. Mansurul, Generic Frameworks for Interactive Personalized Interesting Pattern Discovery (Ph.D. dissertation), Department of Computer Science at Purdue University, 2016.
[20] I.M. Baytas, Contributions to Machine Learning in Biomedical Informatics (Ph.D. dissertation), Computer Science Department at Michigan State University, 2019.
[21] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[22] Z. Che, Y. Liu, Deep learning solutions to computational phenotyping in health care, in: 2017 IEEE International Conference on Data Mining Workshops, New Orleans, LA, USA, 18-21 Nov. 2017.
[23] B. Jin, H. Yang, L. Sun, C. Liu, Y. Qu, J. Tong, A treatment engine by predicting next-period prescriptions, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, ACM, New York, NY, USA, 2018, pp. 1608–1616.
[24] X. Teng, M. Yan, A.M. Ertugrul, Y.R. Lin, Deep into hypersphere: Robust and unsupervised anomaly discovery in dynamic networks, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 2724–2730.
[25] F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, A. Zhang, Risk prediction on electronic health records with prior medical knowledge, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, ACM, New York, NY, USA, 2018, pp. 1910–1919.

[26] B. Bai, S. Zhang, B.L. Egleston, S. Vucetic, Interpretable representation learning for healthcare via capturing disease progression through time, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, ACM, New York, NY, USA, 2018, pp. 43–51.

[27] S. Wu, S. Liu, S. Sohn, S. Moon, C.I. Wi, Y. Juhn, H. Liu, Modeling asynchronous event sequences with RNNs, J. Biomed. Inform 83 (2018) 167–177.

[28] B. Liu, S. Shi, Y. Wu, D. Thomas, L. Symul, E. Pierson, J. Leskovec, Predicting pregnancy using large-scale data from a women's health tracking mobile application, in: Ling Liu, Ryen White (Eds.), The World Wide Web Conference, WWW '19, ACM, New York, NY, USA, 2019, pp. 2999–3005, http://dx.doi.org/10.1145/3308558.3313512.

[29] E. Choi, A. Schuetz, W.F. Stewart, J. Sun, Using recurrent neural network models for early detection of heart failure onset, J. Am Med. Inform. Assoc. (2016).

[30] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, Adv. Neural Inf. Process. Syst. (2016) 3504–3512, http://arxiv.org/abs/1608.05745.

[31] Z. Che, Y. Cheng, Z. Sun, Y. Liu, Exploiting convolutional neural network for risk prediction with medical feature embedding, in: NIPS 2016 Workshop on Machine Learning for Health, NIPS ML4HC, 2016.

[32] Y. Cheng, F. Wang, P. Zhang, Hu J., Risk prediction with electronic health records: A deep learning approach, in: Proceedings of the 2016 SIAM International Conference on Data Mining, 2016.

[33] A. Mayampurath, J. Johnson, Statistical Modeling of Clinical Data, Center for Research Informatics at the University of Chicago, 2018.

[34] K. Stefanidis, Fairness in group recommendations in the health domain, in: Recommender Systems (TIETS43), 2017.

[35] M. Stratigi, H. Kondylakis, Fairgrecs: Fair group recommendations by exploiting personal health information, in: Database and Expert Systems Applications. DEXA 2018, in: Lecture Notes in Computer Science, 11030, Springer, Cham, 2018.

[36] Uri Kartoun, EMRBots: Full source code., 2019, GitHub.

[37] T. Gebert, J. Shuli, S. Jiaxian, Characterizing Allegheny county opioid overdoses with an interactive data explorer and synthetic prediction tool. 2018. arXiv:1804.08830.

[38] K.E. Corey, U. Kartoun, H. Zheng, R.T. Chung, S.Y. Shaw, Using an electronic medical records database to identify non-traditional cardiovascular risk factors in nonalcoholic fatty liver disease, Am. J. Gastroenterol 111 (5) (2016) 671–676, https://www.ncbi.nlm.nih.gov/pubmed/26925881.

[39] L.A. Allen, K.E. Smoyer Tomic, D.M. Smith, et al., Rates and predictors of 30-day re-admission among commercially insured and medicaid-enrolled patients hospitalized with systolic heart failure, Circ. Heart Fail. 5 (6) (2012) 672–679, http://www.ncbi.nlm.nih.gov/pubmed/23072736.