Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210122

Desiderata for a Synthetic Clinical Data Generator

Joshua WIEDEKOPF[a,1], Hannes ULRICH[a], Andrea ESSENWANGER[b], Alexander KIEL[c,d], Ann-Kristin KOCK-SCHOPPENHAUER[a] and Josef INGENERF[a,e]
[a]*IT Center for Clinical Research, University of Lübeck, Germany*
[b]*Berlin Institute of Health (BIH), Germany*
[c]*LIFE – Leipzig Research Centre for Civilization Disease, University of Leipzig, Germany*
[d]*Federated Information Systems, German Cancer Research Center (DKFZ)*
[e]*Institute for Medical Informatics, University of Lübeck, Germany*

Abstract. The current movement in Medical Informatics towards comprehensive Electronic Health Records (EHRs) has enabled a wide range of secondary use cases for this data. However, due to a number of well-justified concerns and barriers, especially with regards to information privacy, access to real medical records by researchers is often not possible, and indeed not always required. An appealing alternative to the use of real patient data is the employment of a generator for realistic, yet synthetic, EHRs. However, we have identified a number of shortcomings in prior works, especially with regards to the adaptability of the projects to the requirements of the German healthcare system. Based on three case studies, we define a non-exhaustive list of requirements for an ideal generator project that can be used in a wide range of localities and settings, to address and enable future work in this regard.

Keywords. Realistic Synthetic Electronic Health Records, Synthetic Data, RS-EHR, Requirements, Secondary Use

1. Introduction

The role of healthcare informatics is ever-increasing. Electronic Health Records (EHRs) are rapidly becoming the standard for healthcare providers, and many jurisdictions have established comprehensive programs to further this process by defining meaningful secondary uses of the routinely captured data, such as the Medical Informatics Initiative (MII) in Germany [1].

Due to complex, well-justified privacy and patient consent requirements, testing these approaches with real data within academic environments is often very challenging. Hence, a clear need for authentic *synthetic* patient data has been identified across the participating consortia within the MII.

In evaluating several approaches for obtaining such synthetic data, it became clear that no existing tool can support all use cases adequately. Inspired by Cimino's paper on *Desiderata for a Controlled Medical Vocabulary* [2], we have identified several use

[1] Corresponding Author: Joshua Wiedekopf, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany; E-Mail: j.wiedekopf@uni-luebeck.de.

cases for a corpus of realistic, but synthetic, Electronic Health Records (henceforth called RS-EHRs).

We examine existing approaches and concepts and identify shortcomings of these solutions. Based on a common set of use cases, we define requirements for an ideal data generation algorithm that will be broadly and internationally applicable.

2. Prior Work

A lot of EHR research is possible by using existing, real corpora of EHR data that have been anonymised by the maintaining organisations, such as the *MIMIC-III* dataset [3]. While the contained data can often be transformed into the desired data structures [4], they are generally highly specific to the jurisdiction that they were recorded in. For example, prescribed medicines may be coded using the US-specific *RxNorm* or *National Drug Code* coding systems, which are not used in Europe. The use of static corpora of anonymised EHR data is thus clearly not adequate for all possible use cases of EHR data [5].

Another very important drawback to the use of pseudonymised or even anonymised databases is the possibility of disclosure of private data from the existing corpus. Computationally feasible re-identification of anonymised health data has already been demonstrated in the literature [3, 6].

By generating RS-EHRs, such considerations can be taken into account, making these generators an ideal tool to alleviate these concerns. Broadly speaking, there are two general approaches, namely *data driven* and *model-based*. The first approaches transfer real patient data into synthetic data by learning the patterns of disease and care in real *background* EHRs. The *EMERGE* framework [7] and machine learning-based approaches such as *medGAN* [8] or statistical approaches [9] fall into this category. While these algorithms require a comparatively minor amount of manual input from subject matter experts, they will inevitably make severe errors when capturing the disease progression and careflow models, compromising the realism of the output data [7, 8]. Additionally, these approaches can only produce features that exist in the background data, so that, for example, the occurrence of diseases is limited to common diseases. To add missing features, modelling-based approaches have to be integrated into a data-driven generator [7]. They also require access to a background EHR corpus, which is subject to privacy laws and may also lead to inadvertent disclosure of protected health information from the real patient data [5, 6, 10].

The *model-based* approaches do not require this access but rely on a number of models for disease progression and patient care instead, which have to be created and curated by subject matter experts. The principle is exemplified in the *PADARSER* approach [10], of which the *Synthea* project [5] is one implementation. This method, however, is considerably more labour-intensive, but will all but guarantee the realism of the generated EHRs, assuming the correctness of the underlying models. *Synthea* is widely used outside of the United States due to its ease of use and output in standard formats (HL7 FHIR and HL7 CDA, among others), but falls short of the ideal software in a number of ways that will become clear in the next sections.

3. Case Studies

Our user-based requirements analysis is based on the following case studies whose use cases an ideal RS-EHR generator should support:

- *Medical Students:* Case reports are an essential learning tool for medical students. By synthesising complete EHRs with prior admissions and pre-existing conditions, this learning experience can be greatly enhanced by adding variety [10].
- *Healthcare Informatics researchers:* As stated above, gaining access to real patient data for secondary use is often not possible for researchers in Healthcare Informatics [5, 7, 10]. Since there is often significant political pressure on this research community, RS-EHRs can greatly accelerate research in this area.
- *Terminologists and Data Stewards:* The support of rich terminologies like SNOMED CT by application systems is still quite limited, leading to low acceptance by end-users. RS-EHRs can readily incorporate these standards, leading to greater adoption, while supporting the work of terminology content creators by demonstrating real-world applications [11].

4. Requirements

We have identified several functional requirements that are shared by the user groups we have selected for our examination. They are stated here in alphabetical order, since we value most requirements to the same degree.

4.1. Free and Open Ecosystem and Adaptation of Software Engineering Best Practices

Implementing a RS-EHR tool can only be a collaborative effort due to the massive amounts of work required in the modelling of disease and treatment processes, but also regarding the implementation of the system itself. It can thus only be satisfactorily achieved in a free and open process via collaboration environments, such as public source code repositories. For maintaining control over the development process, and for ensuring the quality of the product, best practices such as the adherence to a common code style, ensuring good documentation and testing coverage, etc. will have to be embraced.

4.2. Input from Domain Experts

While many aspects of generator development fall squarely in the domain of healthcare IT researchers, the modelling of disease and treatment processes, as well as epidemiological and demographical modelling and related activities, always has to be supported by domain experts. The ideal generator will thus provide modelling tools adequate for use by these experts with a high degree of usability, but also high expressivity (c.f. requirement 4.3). Synthea's Generic Modelling Framework [5] may serve as an example of such a toolkit.

4.3. Maximal Extensibility, Localisability and Configurability

Existing frameworks and toolkits are often very domain-specific and very hard to adapt for other use-cases and localities. For example, Synthea [5] is heavily influenced by the healthcare system of the United States, which makes its use in other localities challenging. This can best be seen when examining the very US-specific modelling of insurance, which plays a major role in Synthea's modelling of treatment but is entirely inadequate when applied to European healthcare systems. It is not integrated within the module framework (where adaptation to other contexts and localities are easily possible) but buried deep within the generation engine code.

By *making these underlying assumptions explicit*, however, the adaptation of the software can be greatly simplified, while enabling the introduction of entirely new functionality, such as the generation of free-text medical notes, images or signals, can be greatly improved and simplified.

Also falling within this requirement is a call for support of diverse and rich output formats and semantic standards, which should be easily adaptable. In the context of requirement 4.6, this implies a need for a support of mechanisms such as HL7 FHIR profiles and extensions, as well as for multiple, locale-specific, codings of concepts.

4.4. No Dependency on Background EHRs

A *hard* dependency on background EHRs is undesirable. It allows for generating only a limited number of RS-EHRs and comes with significant privacy concerns. On the other hand, it may be very beneficial for healthcare providers to be able to access such a background repository that is then used for deriving more precise demographical data or for fine-tuning the occurrence of disease to the specific distribution within the particular healthcare provider's pool of patients. In our opinion, existing model-based approaches like the *PADARSER* framework [10] could be feasibly adapted to satisfy this goal.

4.5. Realism and Clinical Validity

A rather obvious requirement is of course that of realism, which is closely related to the clinical validity of the models. It can be satisfied by a reliance on subject matter experts (c.f. requirement 4.2) and a reliance on evidence-based medicine, which can almost guarantee a very high degree of realism.

However, with real clinical data rarely being perfect [12], a capability for introducing a user-definable amount of common errors may provide a more realistic corpus of data than a perfect process.

4.6. Support of Semantic and Syntactic Standards for Interoperability

Real EHRs come in a plethora of formats and semantic interpretations. Many EHR systems store data in relational tables with mostly incompatible schemata. Additionally, the coded medical content in real EHRs is expressed in a multitude of coding standards, such as the ICD family of classifications, or rich terminologies such as SNOMED CT and LOINC.

However, the primary output format of the ideal generator should be an established interoperability standard such as HL7 FHIR, since these standards can generally be

transformed to other representations without tremendous information loss [4]. Similarly, SNOMED CT and LOINC should be primarily used for clinical concepts, while alternative codings should be supported using (HL7 FHIR) terminology server translation queries at run-time.

Because of fundamental differences in the locale-specific, and often legally prescribed coding of many concept classes (e.g., medicinal products), the importance of this requirement should not be underestimated. Similarly, structural concepts such as encounters are modelled very differently across healthcare systems.

5. Conclusion

The above list of requirements is far from complete. We present them in this paper as a possible ground for discussion in the further development of RS-EHR generators.

Because of the ever-increasing relevance of electronic health records, and use cases that build on these data, in combination with an ever-increasing drive for patient privacy and self-determination, the need for RS-EHRs will also steadily increase.

Acknowledgements

This work was supported by the HiGHmed project within the Medical Informatics Initiative, funded by the Federal Ministry of Education and Research (grant number 01ZZ1802Z).

References

[1] Semler SC, Wissing F, and Heyder R. German Medical Informatics Initiative, Methods Inf. Med. 2018; 57:e50–e56.
[2] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century, Methods Inf. Med. 1998; 37:394–403.
[3] Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database, Sci. Data. 2016; 3:160035.
[4] Ververs S, Ulrich H, Kock A-K, and Ingenerf J. Konvertierung von MIMIC-III-Daten zu FHIR, 63 Jahrestag. GMDS. 2018; 256.
[5] 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; 25:230–238.
[6] Emam KE, Dankar FK, Vaillancourt R, Roffey T, and Lysyk M. Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records, Can. J. Hosp. Pharm. 2009; 62.
[7] Buczak AL, Babin S, and Moniz L. Data-driven approach for creating synthetic electronic medical records, BMC Med. Inform. Decis. Mak. 2010;10.
[8] Choi E, Biswal S, Malin B, Duke J, Stewart WF, and Sun J. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks, ArXiv Prepr. 2017.
[9] Zhang L, Ding X, Ma Y, Muthu N, Ajmal I, Moore JH, et al. A maximum likelihood approach to electronic health record phenotyping using positive and unlabeled patients, J. Am. Med. Inform. Assoc. 2020; 27:119–126.
[10] Dube K, and Gallagher T. Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use, in: J. Gibbons, and W. MacCaull (Eds.), Found. Health Inf. Eng. Syst., Springer, Berlin, Heidelberg, 2014; 69–86.
[11] Cangioli G, Chronaki C, Gooeg KR, Højen AR, Karlsson D, Jaulent MC, et al. Current and Future Use of SNOMED CT. Deliverable in the ASSESS CT project, 2016.
[12] Bell SK, Delbanco T, Elmore JG, Fitzgerald PS, Fossa A, Harcourt K, et al. Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes, JAMA Netw. Open. 2020; 3:e205867.