Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240739

SIMpat: A Synthetic Benchmark for Similarity Metrics on Patient Representations

Jean-Virgile VOEGELI*[a,b], Mina BJELOGRLIC*[a,b,1],
Christophe GAUDET-BLAVIGNAC*[a,b], Richard DUBOS [*a,b],
Myriam ZIMMERMANN [a,b], Adel BENSAHLA TALET [a,b], Yuanyuan ZHENG [a,b],
Julien EHRSAM[a,b] and Christian LOVIS [a,b]
[a] *Division of Medical Information Sciences, Geneva University Hospitals, Geneva, Switzerland*
[b] *Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland*
*These authors contributed equally to this work
ORCiD ID: Mina Bjelogrlic https://orcid.org/0000-0002-6922-3283

Abstract. Similarity and clustering tasks based on data extracted from electronic health records on the patient level suffer from the curse of dimensionality and the lack of inter-patient data comparability. Indeed, for many health institutions, there are many more variables, and ways of expressing those variables to represent patients than patients sharing the same set of data. To lower redundancy and increase interoperability one strategy is to map data to semantic-driven representations through medical knowledge graphs such as SNOMED-CT. However, patient similarity metrics based on this knowledge-graph information lack quantitative evaluation and comparisons with pure data-driven methods. The reasons are twofold, firstly, it is hard to conceptually assess and formalize a gold-standard similarity between patients resulting in poor inter-annotator agreement in qualitative evaluations. Secondly, the community has been lacking a clear benchmark to compare existing metrics developed by scientific communities coming from various fields such as ontology, data science, and medical informatics. This study proposes to leverage the known challenges of evaluating patient similarities by proposing SIMpat, a synthetic benchmark to quantitatively evaluate available metrics, based on controlled cohorts, which could later be used to assess their sensibility regarding aspects such as the sparsity of variables or specificities of patient disease patterns.

Keywords. Benchmark, Patient representations, similarity metrics

1. Introduction

Research directly or indirectly relying on patient similarity measures based on patient representations, could benefit from common and key indicators for fair and useful comparison [1]. Melton et al. [2] proposed to differentiate between the *semantic distance* which measures the relative closeness between two concepts of interest in a taxonomy from the *clinical distance*, which is the amount of relative evidence for closeness from

[1] Corresponding Author: Mina Bjelogrlic; E-mail: mina.bjelogrlic@unige.ch.

an inter-patient distance perspective when comparing a single concept in one case with the nearest concept in a second case. They report three important considerations: i) Inter-patient distance is not determining if cases are identical (two identical EHR cases do not mean that patients are totally similar); ii) Semantic distance between two concepts is different than clinical distance between two case features (concepts carry different amount of clinically relevant information); iii) Clinical distance from a concept in one case to the nearest concept in another case can be calculated using defined relationships to find the minimal-cost path.

Additionally, the proposed metrics in the literature come from different scientific communities coming from various fields such as ontology, data science, and medical informatics, and the community has been lacking a robust benchmark [3] to compare existing metrics to build up on common conclusions [4] and translate valuable qualitative evaluations from different clinical experts into new metric propositions. The proposed benchmark, SIMpat, has the limitations of being synthetically generated, but brings to the community a reasonably large-scale dataset with matched cohorts and a framework for fair comparison of existing metrics from the literature.

2. Materials and Methods

Synthea is an open-source synthetic patient generator. It creates patients starting from their birth and model their medical records until their death, in an independent manner. Their diseases, conditions and medical care are defined by one or more generic modules: each module models events that could occur in a real patient's life with a progression of states and a description of transition between them. Full details about the Synthea generator and their modules withs pre-established probabilities for patient generation can be found in [5]. We selected 6 different diseases encoded with SNOMED-CT concepts (SCT) [6], that were deemed by a medical professional as “different enough”:

- Cerebral Palsy (SCT 128188000);
- Colorectal Cancer (SCT 93761005);
- Dialysis (SCT 265764009). Dialysis is a condition and not a disease, but is used here as a proxy for renal issue;
- Hypertension (SCT 59621000);
- Breast Cancer (SCT 254837009);
- Prostate Cancer (SCT 126906006). All men with prostate cancer are Veterans in Synthea.

The methodology for the construction of the synthetic dataset is presented in Fig. 1).

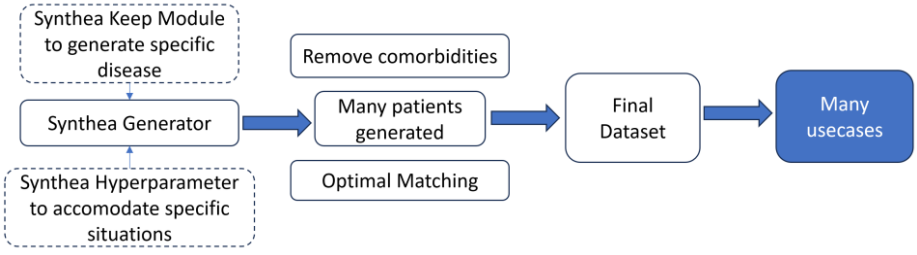

Figure 1. Methodology for the construction of the synthetic dataset.

The parameters of the data generation are the following: each individual was set to be between the age of 18 and 80 years old, located in the default location, Massachusetts, with a medical history of 10 years. Except for specific sex-disease such as breast cancer and prostate cancer, all cohorts contain both male and female individuals. For the age parameter, note that Synthea modules sometimes specify a minimum age to onset a certain condition / disease: For example, colorectal cancer can only onset after 50 years old, and prostate cancer after 60 years old. Although limiting, it reflects nonetheless the consensus around the distribution of considered diseases by age.

For each patient, Synthea generates a lifetime medical history and keeps it if and only if a specified disease (here, a specific SCT code) has appeared at least once. If not the case, Synthea tries again for a fixed number of times. In our case, each Synthea run was set to try 10.000 times. Since Synthea generates the whole life of a patient, this allows for comorbidity when generating a patient. This means that even though a patient is in a specific cohort and has a specific disease, he or she can also have another disease from another cohort. The cohorts are matched on age, sex, BMI, smoking status, and revenue (proxy for socio-economical status) using optimal matching and Mahalanobis distance. Patients are matched 2 by 2, then joined together. The standardized Mean Difference (SMD) is computed on the overall matched population to assess the effect size (less than 0.1 is ideal, between 0.1 and 0.2 is acceptable).

3. Results

The rows in Table 1 represent the cohort generated by Synthea, while the columns represent the diseases.

Table 1. Disease matrix for all individuals

Cohort	Breast Cancer	Cerebral Palsy	Colorectal Cancer	Dialysis	Hypertension	Prostate Cancer
Breast Cancer	187	0	0	0	94	0
Cerebral Palsy	0	346	0	0	170	0
Colorectal Cancer	0	0	346	0	202	0
Dialysis	0	0	0	346	337	0
Hypertension	0	0	0	0	346	0
Prostate Cancer	0	0	0	0	90	159
Total	187	346	346	346	1239	159

Table 2. Descriptive Table of the cohorts. Median [IQR], Pearson's Chi-squared test; Kruskal-Wallis rank sum test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates). Diastolic Blood Pressure (DBP). Systolic Blood Pressure (SBP). Blood Sugar (BS). High School (HS) Distinct (Dist.)

Characteristic	Breast Cancer N = 187	Cerebral Palsy N = 346	Colorectal Cancer N = 346	Dialysis N = 346	Hypertension N = 346	Prostate Cancer N = 159
Gender pv <0.001						
F	100%	54%	54%	54%	54%	0%
M	0%	46%	46%	46%	46%	100%
Age pv <0.001	62[54-70]	66[52-79]	66[57-74]	61[51-75]	65[52-79]	73[67-78]
Smoker pv = 0.39						
Ex-smoker	23%	28%	28%	27%	28%	34%

Education pv = 0.98						
HS	27%	30%	33%	30%	29%	28
no answer	0.5%	0.9%	0.9%	0.9%	1.2%	1.3%
< HS	12%	12%	11%	12%	12%	8.2%
>HS	61%	57%	55%	57%	58%	62%
Sal. (k$) pv>0.99	65[33-119]	66[28-122]	67[31-118]	69[32-119]	68[32-116]	58[32-113]
DBP mmHg pv <0.001	78[71-86]	77[69-85]	77[69-86]	78[69-88]	81[73-90]	75[68-84]
SBP mmHg pv = 0.003	118[108-126]	115[104-127]	117 [104-127]	118 [107-129]	119 [106-129]	113 [102-123]
BS mg/dL pv = 0.015	86[74-96]	86[75-95]	81[72-91]	82[73-93]	83[74-93]	80[71-92]
Dist. SCT pv <0.001	51[43-64]	53[42-66]	51[43-65]	59[50-73]	46[36-61]	56[47-67]
SCT codes pv <0.001	180 [144-221]	143 [118-196]	148 [114-203]	677 [502-865]	143 [108-216]	207 [139-443]

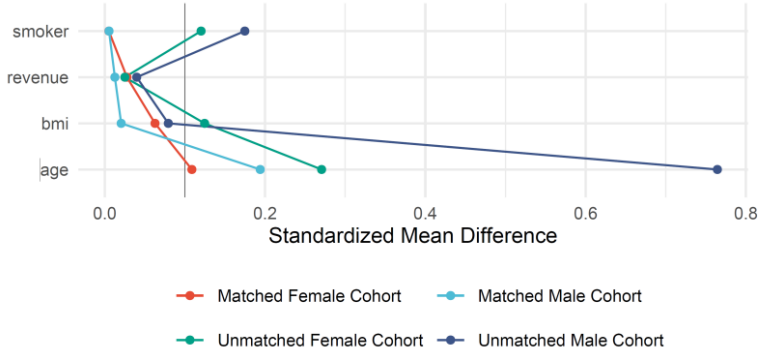

Figure 2. Left: Covariate Balance for matched variables: age, smoker, BMI and revenue.

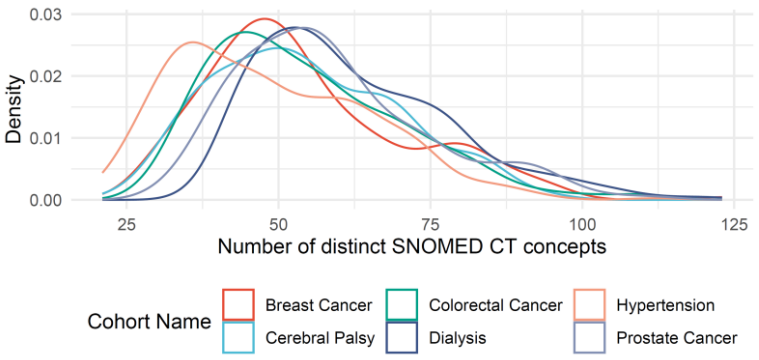

Figure 3. Density of distinct SCT per cohort.

4. Discussion

We can see that one of the most present comorbidities is Hypertension, which is present in all cohorts. As well, Dialysis is also a fairly common comorbidity. The matching solved most of the imbalance between covariates (with a SMD below 0.1, for more details see Fig. 2) according to the main variables, namely age, smoker, BMI, and revenue. Overall, we have 317.99 (sd = 309.51) SCT concepts per generated patients. Minimum number of concepts for a patient is 43 and maximum is 2495. Patients have an average of 55.45 distinct SCT concepts (16.49) (see Fig.3).

In this paper we propose a novel benchmark for evaluation of existing and future similarity metrics for patient representations. The contributions are an open-source implementations of State-of-The-Art metrics [7], an open-source dataset with 6 cohort, and the framework to construct more cohorts following the same strategies, with full details publicly available in [8]. Overall, the proposed dataset is a starting point to fairly evaluate patient similarity, quantitatively assess advantages and disadvantages, and raise fundamental questions of current knowledge and data-driven strategies to represent patients.

Further work will include advanced statistical tests to assess the discriminability of the different patient representations using 12 State-Of-The-Art different distance metrics to separate patients: i) 4 *knowledge-graph* based metrics from section 2.3 from [2] with our open-source implementation already available in [7]; ii) 4 *text embedding* methods; and iii) 4 *graph embedding* methods reviewed by [9] and now publicly available for the SIMpat dataset [8]. As a first impression it appears that the distribution on the distance and hence the represented similarity between patients varies drastically. We argue that testing on controlled cohorts composed by very distinct (to less distinct) cases, could give more insight on robustness and sensitivity on aspects such as the sparsity of variables and clinical specificities of patient disease patterns in order to propose more robust, clinically relevant, stratified and personalized patient representations.

References

[1.] Keszthelyi D, Gaudet-Blavignac C, Bjelogrlic M, Lovis C. Patient Information Summarization in Clinical Settings: Scoping Review. JMIR Med Inform. 2023 Nov 28;11(1):e44639.

[2] Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform. 2006;39(6):697–705.

[3] Röösli E, Bozkurt S, Hernandez-Boussard T. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. Sci Data. 2022 Dec;9(1):24.

[4] Mincu D, Roy S. Developing robust benchmarks for driving forward AI innovation in healthcare. Nat Mach Intell. 2022 Nov 15;4(11):916–21.

[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 JAMIA. 2018 Mar 1;25(3):230–8.

[6] SNOMED International. Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [Internet]. 1999 [cited 2022 Jun 27]. Available from: https://www.snomed.org/

[7] GitHub [Internet]. [cited 2024 Mar 18]. Jyvay/SIMpat. Available from: https://github.com/Jyvay/SIMpat

[8] Voegeli JV, Bjelogrlic M, Gaudet-Blavignac C, Dubos R, Zimmermann M, Ehrsam J, Bensahla Talet A, Zheng Y, & Lovis C. SIMpat: a synthetic benchmark for similarity metrics on patient representations [Data set]. (2024) Zenodo. https://doi.org/10.5281/zenodo.10830066

[9] Pattisapu N, Patil S, Palshikar G, Varma V. Medical Concept Normalization by Encoding Target Knowledge. In: Proceedings of the Machine Learning for Health NeurIPS Workshop [Internet]. PMLR; 2020 [cited 2023 Oct 25]. p. 246–59. Available from: https://proceedings.mlr.press/v116/pattisapu20a.html