Research Paper

Process Mining and Synthetic Health Data: Reflections and Lessons Learnt

Alistair Bullward[1], Abdulaziz Aljebreen[2][0000-0002-4746-3446], Alexander Coles[2], Ciarán McInerney[2][0000-0001-7620-7110] and Owen Johnson[2][0000-0003-3998-541X]

[1] NHS Digital, UK
alistair.bullward1@nhs.net
[2] University of Leeds, UK
{ml17asa, scadc, c.mcinerney, o.a.johnson}@leeds.ac.uk

Abstract. Analysing the treatment pathways in real-world health data can provide valuable insight for clinicians and decision-makers. However, the procedures for acquiring real-world data for research can be restrictive, time-consuming and risks disclosing identifiable information. Synthetic data might enable representative analysis without direct access to sensitive data. In the first part of our paper, we propose an approach for grading synthetic data for process analysis based on its fidelity to relationships found in real-world data. In the second part, we apply our grading approach by assessing cancer patient pathways in a synthetic healthcare dataset (The Simulacrum provided by the English National Cancer Registration and Analysis Service) using process mining. Visualisations of the patient pathways within the synthetic data appear plausible, showing relationships between events confirmed in the underlying non-synthetic data. Data quality issues are also present within the synthetic data which reflect real-world problems and artefacts from the synthetic dataset's creation. Process mining of synthetic data in healthcare is an emerging field with novel challenges. We conclude that researchers should be aware of the risks when extrapolating results produced from research on synthetic data to real-world scenarios and assess findings with analysts who are able to view the underlying data.

Keywords: Process Mining, Synthetic Data, Simulacrum, Data Grading, Taxonomy.

1 Introduction

A care pathway is "a complex intervention for the mutual decision-making and organisation of care processes for a well-defined group of patients during a well-defined period" [1]. Care pathways describe ideal patient journeys and the extent to which individual patients follow this ideal can be explored through analysis of data extracted from healthcare information systems. Such data can include patient-level events like

admissions, investigations, diagnoses, and treatments. Process-mining of healthcare data can help clinicians, hospitals and policy makers understand where care pathways are helping and hindering patient care [2]. However, healthcare data is sensitive and identifiable data, which necessitates strong information governance to protect patients' privacy. This necessary governance can make it difficult to access healthcare data for beneficial analysis and research (especially for process discovery where a clear purpose is harder to pin down and hence link to a legal basis). One solution is to make highly-aggregated open datasets available. For example, NHS Digital publishes open data across 130+ publications spanning key care domains. However, such datasets are often not sufficiently detailed for patient-level process mining of care pathways. Consequently, there has been a growth in synthetic or simulated data that attempt to mirror aspects of the real, patient-level data without disclosing patient-identifiable information [3].

Generating synthetic data from real world data sets can be achieved via a number of methods. An example of synthetic healthcare dataset from the USA is SyntheticMass which is an unrestricted artificial publicly available healthcare dataset containing 1 million records generated using Synthea [4]. This dataset was generated using public healthcare statistics, clinical guidelines on care maps format and realistic properties inheritance methods. Another example from the UK is a project developed by NHSx AI Lab Skunkworks called Synthetic Data Generation [5]. In this project, a model previously developed by NHS called SynthVAE has been adopted to be used with publicly accessible healthcare dataset MIMIC-III in order to read the data (inputs), train the model then generate the synthetic data and check the data through a chained pipeline. A third example is synthetic datasets generated using Bayesian networks [6] have demonstrated good-to-high fidelity [7] and can be coupled with disclosure control measures [8] to provide complex, representative data without compromising patient privacy.

Regardless of generation method, rigorous evaluation of synthetic data is needed to assure and ensure representativeness, usefulness and minimal disclosivity. Approaches to evaluation include using generative adversarial networks that incorporate privacy checks within the data-generation process [9]; discrepancy, distance and distinguishability metrics applied to specific analysis goals [10]; meaningful identity disclosure risk [11]; multivariate inferential statistical tests of whether real and synthetic datasets are similar [12]; conditional attribute disclosure and membership disclosure [12]; and others [13]. What has not been suggested to date are approaches to evaluation that are specific to process mining, we hypothesise that process mining of health care pathways has a set of specific data requirements that may not be easily satisfied by current approaches to synthetic healthcare data creation. To explore this, we present a taxonomy for synthetic data in healthcare to help evaluate and grade synthetic datasets to identify those that would be useful for process mining. We apply our taxonomy to a case study of the Simulacrum cancer dataset, which is an openly available dataset of cancer treatment data based on the English National Cancer Registration and Analysis Service [14].

2 Method

Our methods are presented in four parts. In part 1, we propose a taxonomy of synthetic data for process mining in healthcare. In part 2, we define a set of tests to classify synthetic data against the taxonomy. In part 3, we describe the Simulacrum dataset that we use in our case study. Finally, in part 4, we evaluate the Simulacrum dataset using the tests from part 2, and classify the dataset according to our taxonomy from part 1.

2.1 Part 1: A Taxonomy for Synthetic Data in Healthcare

We present a 3-grade taxonomy to help classify the fidelity of a synthetic dataset. By fidelity, we refer to the extent to which synthetic data represents the real data it is attempting to replace. Random data presented in the format of the real data has low fidelity but might have functional value for testing analysis pipelines because it has the "right shape". If synthetic data also mirrors statistical relationships within variables, then it has greater fidelity and has some inferential value following analysis. Greater fidelity would be demonstrated by a synthetic dataset that mirrors the real data's statistical relationships between variables.

More formally, we define a minimum grade 1 synthetic dataset as one in which the format of the synthetic data matches that of the original dataset from which it was derived. The types of features represented in the original dataset must be faithfully represented. Examples for healthcare data include time-stamped events, patient identifiers, and treatment codes. Grade 1 synthetic datasets are not expected to retain any statistical or clinically-meaningful relationships within or between columns. From the perspective of process mining, we expect to be able to produce a process model but the sequences of events depicted in the model are not expected to be realistic, nor are the event and transition metadata (e.g. event counts or inter-event duration).

We define a grade 2 synthetic dataset as one in which the independent distributional properties of each synthetic variable are similar (statistically or clinically) from the same properties of each variable in the original dataset. Grade 2 datasets are not expected to retain any statistical or clinically-meaningful relationships between features. From the perspective of process mining, we expect to be able to produce a process model and for the event and transition metadata to be realistic, but we do not expect the sequences of events depicted in the model to be realistic.

We define a grade 3 synthetic dataset as one in which the multivariate distributional properties of all synthetic variables are similar (statistically or clinically) from the same properties of all variables in the original dataset. From the perspective of process mining, we expect to be able to produce a process model, for the event and transition metadata to be realistic, and for the sequences of events depicted in the model to be realistic. This paper focuses on assessing grade 3 for analytical process mining.

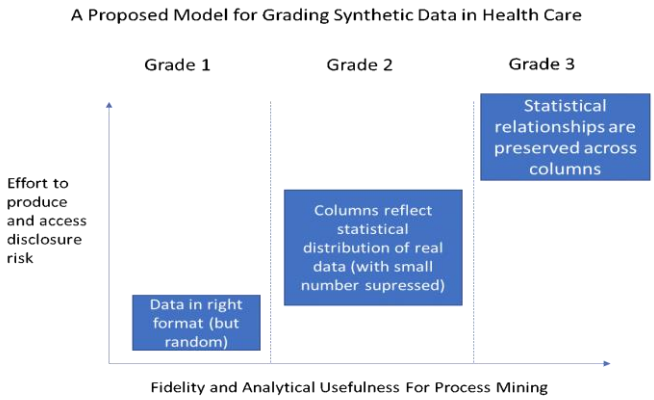


Fig. 1. A proposed model for grading synthetic data in healthcare

2.2 Part 2: Criteria for Grading Simulated Data

In part 1, we presented a 3-grade taxonomy to help classify the fidelity of a synthetic dataset in healthcare. Below, we present a set of criteria that would identify the grade of a given synthetic healthcare dataset. Taken together, the taxonomy and the criteria provide a framework for evaluating the suitability of a synthetic dataset for process mining. We suggest some tests against these criteria but we encourage analysts to design, implement and share their own tests in keeping with the principles of the criteria, below.

Criterion 1: The variables within the real dataset are present within the synthetic dataset and are of the correct data type.

A sufficient test of this criterion is a basic one-to-one mapping of variable names and data types. If a process model can be derived from the synthetic dataset, then this criterion is also met.

Criterion 2: Each synthetic variable's typical value, range, and distribution are statistically- or clinically-meaningful similar to the relative variable in the real dataset.

If a statistical approach is preferred, then candidate tests of this criterion are null-hypothesis significance tests for similarity of, for example, each variable's mean or median. Importantly, each of these null-hypothesis tests would not be sufficient to meet this criterion if they are conducted in isolation. This is because these tests do not test all distributional parameters. Even tests of distributions like the Komolgorov-Smirnov test only test the minimum largest difference between two distributions rather than the entire distribution.

If a clinical approach is preferred, then clinical and administrative domain experts can audit distribution summary statistics. This is in keeping with the ethos of PM2 methodology where domain experts are involved in the process mining [15].

Criterion 3: The sequential, temporal, and correlational relationships between all variables are statistically- or clinically-meaningfully similar to those present in the real dataset.

Correlational relationships can be tested using a multivariate null-hypothesis statistical test for similarity but are subject to the same limitations as similar tests applied to Criterion 2. This criterion might also be satisfied if it is possible to progress with iterative, process-mining methodology involving the production, evaluation and review of event logs and process models. One could also meet this criterion by testing if a process model derived from the synthetic dataset passes tests of conformance with a process model derived from the real dataset.

2.3 Part 3: Case study of the Simulacrum cancer dataset

The Simulacrum is a synthetic dataset derived from the data held securely by the National Cancer Registration and Analysis Service (NCRAS) within Public Health England [14]. NCRAS holds data on all cancer diagnoses in England and links them to other datasets collected by the English National Health Service. The Simulacrum uses a Bayesian network to provide synthetic data on patient demographics, diagnoses and treatments based on real patient data between 2013 and 2017. Table 1 shows a sample of the variables available in the Simulacrum that are relevant to process mining.

Table 1. Summary of Activity Data Available in the Simulacrum Dataset for 2,200,626 patients.

Activity	**Count of events across all cancers**	**Summary of information available for event**
Diagnosis Date	2,741,065	Site of neoplasm Morphology Stage grade of tumour age at diagnosis Sex cancer registry catchment area oestrogen receptor HERs status of the tumour Clinical nurse specialist Gleason Patterns Date of first surgical event Laterality Index of multiple deprivation
Decision to treat (Regimen)	749,721	Decision to treat date (Drug regimen)
First Surgery	1,736,082	Date of first surgical event linked to this tumour recorded in the Cancer Registration treatment table
Start Date on Regimen	828,980	Patient's height (metres (m)) Patient's weight (kilograms (kg)) Drug treatment intent Decision to treat date (Drug regimen) Start date (Drug regimen)

		Maximally granular mapped regimen Clinical trial indicator Chemo-radiation indicator Regimen grouping (benchmark reports)
SACT Cycle Start	2,561,679	Pseudonymised cycle ID Pseudonymised regimen ID Cycle identifier Start date (Cycle) Primary procedure (OPCS) Performance Status
Deaths	652,418	Date of Death

The Simulacrum dataset contains synthetic treatment events and associated variables for multiple cancers. We selected data from for malignant neoplasms of the brain (identified by the 3-character ICD10 code C71)

2.4 Part 4: Evaluation

We did not have access to the real world data on which the Simulacrum was based. We reviewed the Simulacrum for the presence of variables relevant to the brain cancer and leukaemia care pathways, and checked that the data types were appropriate, e.g. timestamp was a datetime data type. We assumed that the variables in the Simulacrum were also present in the real world. Regarding grade 2 fidelity, the producer of the Simulacrum synthetic dataset provided evidence that the distributions of each variable in the datasets were similar to those of the real dataset [16]

To test grade 3 fidelity, we sought to derive a process model of brain cancer from the Simulacrum synthetic data by applying process discovery to relevant variables. Patient ID was used as the case identifier, clinical events were used as the activity, and each event had an associated timestamp to produce an event log. PM4PY [17] packages were used to produce the process models and the PRoM was used to discover the processes [18]. Trace variants were extracted from the event log and reviewed by clinical experts for reasonableness.

To aid conformance checking, a normative model representing the expected pathway to be followed for brain cancer using available activities in Simulacrum was informed by brain tumour patient guides from the Brain Trust [19]. Conformance was quantified as the fitness of the synthetic event log when replayed on a petri net of the expected pathway [20]. This replay fitness provides a 0-1 measure of how many traces in the synthetic data's event log can be reproduced in a process model defined by the expected pathway, with penalties for skips and insertions.

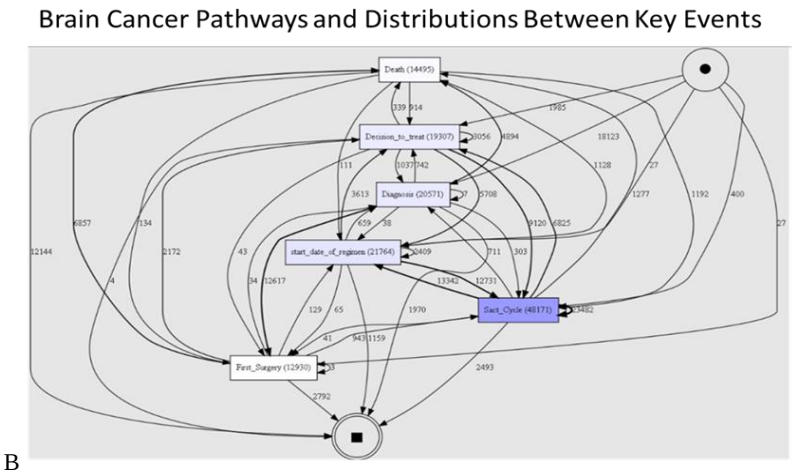
The distributions of durations between diagnosis and first surgery was also reviewed in the synthetic and the real dataset with the assistance of the producers of the Simulacrum synthetic dataset. This permitted a simple evaluation of the reasonableness of the temporal relationship between variables.

3 Results

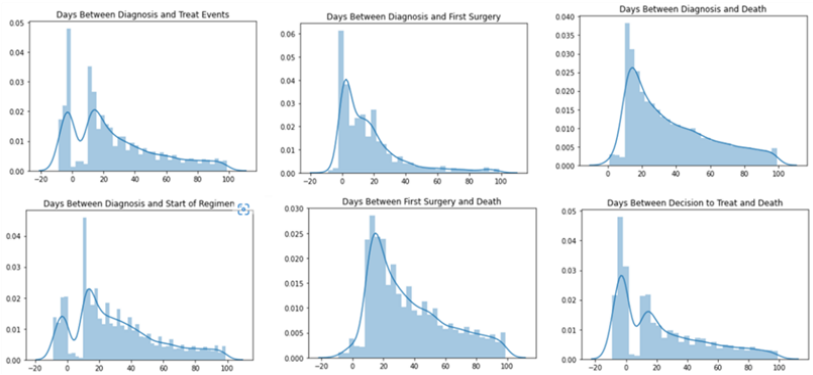
The fields required to inform the care pathway for brain cancers and leukaemia were all present and variables' data types were all correct. The discovered process model for

brain cancer shows a substantial variety of sequences that differ from the care pathway derived from the Brain Trust (Figure 2). Replay fitness of the synthetic event log on the expected pathway was 46%.

A



B



C

Fig. 2. A. The expected care pathway for brain cancer. *B.* Process discovery on Brain Cancer Pathways (ICD10 code C71). *C.* Histograms of durations between a sample of event pairs.

Of the 20,562 traces in the Simulacrum's brain cancer dataset, there were 4,080 trace variants (Figure 3). Most variants were unique traces (n_1 = 3,889) and there were relatively few variants matching only two traces (n_2 = 89). The four-most-common variants represented 75.9% of traces (15,608 / 20,562). In 122 spurious traces, the "Death" event occurred before the "First_Surgery" event. Figure 4 presents the transition matrix

between events with the care pathway being represented by the diagonal starting at the second cell from the top left, i.e. Start-Diagnosis date = 18,123.

Fig. 3. The seven most common trace variants for brain cancer, accounting for over 77% of all trace variants.

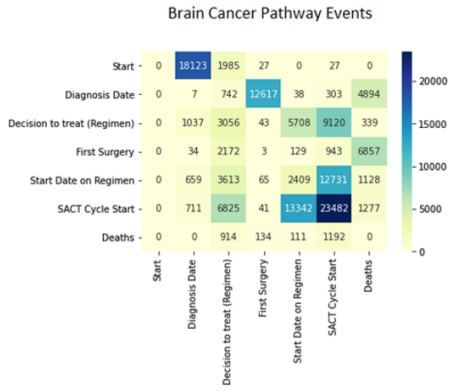


Fig. 4. Brain Cancer Event Summary

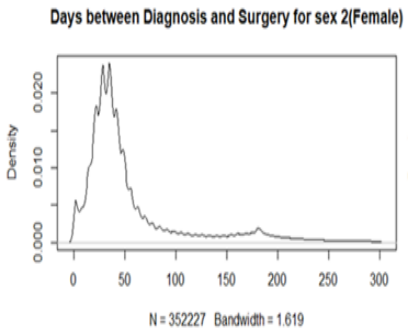


Fig. 5. Distribution of days between diagnosis and first surgery for all cancers for Females

Figure 5 shows the distribution of computed duration between date of diagnosis and first surgery, in the female sub cohort. There is a typical value of approximately 35 days but a long skew duration in the low hundreds of days There also appears to be a regular signal with a period of approximately 7-10 days.

4 Discussion

Care pathways are increasingly key in analysing health data. The aim of this paper was to present a taxonomy for synthetic data in healthcare to help evaluate and grade

synthetic datasets to identify those that would be useful for process mining. We conducted an example evaluation on the Simulacrum dataset.

According to our tests, we conclude that the Simulacrum meets the grade 3 criterion of our taxonomy. Grade 1 was met by our finding that the fields required to inform the care pathway for brain cancer were all present and variables' data types were all correct. Grade 2 was evidenced by the Simulacrum's producer's assuring that the distributions of each variable in the datasets were similar to those of the real dataset [16]. Grade 3 was evidenced by our ability to progress with an iterative, process-mining approach that involved the production of a process model and event log summary statistics that were reviewed with clinical experts and the producer of the synthetic dataset. In the remaining sections, we provide further details of the discussions with the producers of the Simulacrum synthetic dataset.

4.1 Meeting the grade 3 criterion

Our criterion for meeting grade 3 fidelity is if the sequential, temporal, and correlational relationships between all variables are statistically- or clinically-meaningfully similar to those present in the real dataset. We tested this criterion by progressing with an iterative, process-mining methodology and by testing if a process model derived from the synthetic dataset passes tests of conformance with a process model derived from the real dataset.

The Simulacrum synthetic dataset was able to produce a process model and trace variants that were similar to portions of the ideal care pathway.

The reasonableness of the synthetic dataset was also evidenced by our analysis of the distribution of days between diagnosis and first surgery, in female patients (Figure 5). Figure 5 also shows what appears to be a regular signal with a period of approximately 7-10 days. Discussions with the producers of the Simulacrum synthetic dataset confirmed that this regular signal reflects the underlying non-synthetic data. Collaborative discussions suggested the signal reflects weekly patterns for booking surgery - for example non-urgent surgery tends to be booked on weekdays - but we have yet to test this hypothesis. Such analysis and representativeness would not be possible with synthetic datasets lower than grade 3.

Regarding a formal check of conformance, a replay fitness of 46% is considered low, suggesting that the expected care pathway does not represent the behaviour observed in the synthetic data's event log well [20]. It is not clear whether the poor replay fitness represents poor adherence to guideline care pathways or poor fidelity of the Simulacrum data set. Guideline care pathways represent ideal patient journeys but real-life cancer treatment is known to be complex [21]. For example, process models discovered for endometrial cancer show good replay fitness but require more-complex processes [22]. The replay fitness of our discovered process model for brain cancer was 46%, which, assuming the Simulacrum data is representative, suggests that the care pathways for brain cancer are more complex than what is presented in the idealised care pathways.

4.2 Data quality

According to the ideal care pathway, we would expect all patients to experience all events that were selected from the Simulacrum synthetic dataset, and in the order specified by the ideal care pathway. On the contrary, Figure 4 shows substantial deviation from the ideal care pathway. This is indicated partly by non-zero diagonal counts that indicate direct repeats of events (though repeated SACT cycles are not unexpected). Deviation from the ideal care pathway is also partly indicated by non-zero counts anywhere beyond the diagonal starting at the second cell from the top left. For example, there were 1,037 synthetic patient records that showed a patient receiving a decision to treat before a diagnosis date. These deviations could be accounted for if patients were diagnosed with multiple genetically-distinct cancers. For example, it is plausible that the 34 synthetic patients that underwent cancer-related surgery before diagnosis were undergoing diagnostic surgery, or were patients undergoing curative or debulking surgery and in whom an additional, genetically-unique cancer was discovered following analysis of the biopsy.

However, the observation that 1,192 synthetic patient records show a patient has died before their SACT cycle started cannot be explained by the real-life complexity of healthcare delivery. Alternative explanations for these cases include administrative errors or spurious simulation during the data generating process. Our collaborative discussions with the producers of the synthetic dataset revealed that this anomaly was a known feature of the generation of the synthetic data rather than being a feature of the real data.

4.3 Collaboration with producers of the synthetic dataset

During the course of this work we have collaborated with the producers of the synthetic dataset under study. We felt that this was a crucial activity to aid in the efficient and effective use of the dataset. For example, without communication with producers of the synthetic datasets, it might not be possible to tell if a data quality issue is a result of the synthetic data generation or representative of the underlying data.

We have already presented two examples of the benefits of collaborating with the producers of synthetic datasets. The first was our analysis of the durations between date of diagnosis and first surgery (Figure 5). It was only through discussion with the producers of the synthetic data that we were able to check that the distribution of computed durations was representative of real world data, and that we were able to collaboratively hypothesise an explanation for the regular 7-10 day signal. The second example was our ability to conclude that the anomalous transitions between death and SACT cycle were an artefact of the Simulacrum's data-generating process.

4.4 Recommendations

We make the following recommendations to analysts who intending to apply process mining to synthetic datasets:

1. Producers of synthetic health data should grade it and produce evidence using test cases that will help users determine whether the data is relevant to their study.
2. Consumers of synthetic data should expect to liaise with the producer. In particular, they should:
   a. Ask how the data were generated.
   b. Ask what tests of representativeness, usefulness and disclosivity were conducted.
   c. Apply our taxonomy to grade the dataset.
   d. Have a line of communication open to discuss data quality issues.

5 Conclusions

In conclusion, process mining of care pathways is an important approach for improving healthcare but accessing patient event based records is often burdensome. Synthetic data can potentially reduce this burden by making data more openly available to researchers, however the quality of the synthetic data for process mining needs to be assessed. We propose an evaluation framework and demonstrated this framework using the openly available Simulacrum Cancer data set and identified this data set can be thought of as grade 3 which makes it useful for process mining. Although researchers may be able to explore synthetic data and generate hypotheses, we argue that they will need to work with producers with access to the real data to confirm findings. This paper makes a number of recommendations for producers and consumers of synthetic data sets and highlights potential further work on the taxonomy to subdivide different types of grade 3 data.

References

1. Vanhaecht, K.: The Impact of Clinical Pathways on the Organisation of Care Processes. Doctoral dissertation, (2007). Last accessed 24/08/2022.
2. Schrijvers, G., Hoorn, A. van . and Huiskes, N.: The Care Pathway Concept: concepts and theories: an introduction. International Journal of Integrated Care, 12(6), p.None. DOI: http://doi.org/10.5334/ijic.812 (2012)
3. The NHS X Analytics Unit homepage, https://nhsx.github.io/AnalyticsUnit/synthetic.html, last accessed 24/08/2022.
4. Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S.: Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. In: Journal of the American Medical Informatics Association "JAMIA", 25(3), 230–238 (2018).
5. AI Skunkworks projects homepage, https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/skunkworks/ai-skunkworks-projects, last accessed 24/08/2022.
6. Kaur, D., Sobiesk, M., Patil, S., Liu, J., Bhagat, P., Gupta, A., & Markuzon, N.: Application of Bayesian networks to generate synthetic health data. In: Journal of the American Medical Informatics Association "JAMIA", 28(4), 801–811 (2021).
7. Shen, Y., Zhang, L., Zhang, J., Yang, M., Tang, B., Li, Y., & Lei, K.: CBN: Constructing a clinical Bayesian network based on data from the electronic medical record. In: Journal of biomedical informatics, 88, 1–10 (2018).

8. Sweeney, L.: Computational disclosure control: A primer on data privacy protection. Doctoral dissertation, Massachusetts Institute of Technology, (2001). Last accessed 24/08/2022.
9. Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P.: Generation and evaluation of privacy preserving synthetic health data. In: Neurocomputing, 416, 244-255 (2020).
10. El Emam, K., Mosquera, L., Fang, X., & El-Hussuna, A.: Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study. In: JMIR medical informatics, 10(4), (2022).
11. El Emam, K., Mosquera, L., & Bass, J.: Evaluating identity disclosure risk in fully synthetic health data: model development and validation. In: Journal of medical Internet research, 22(11), (2020).
12. El Emam, K., Mosquera, L., Jonker, E., & Sood, H.: Evaluating the utility of synthetic COVID-19 case data. In: JAMIA open, 4(1), (2021).
13. El Emam, K.: Seven ways to evaluate the utility of synthetic data. In: IEEE Security & Privacy, 18(4), 56-59, (2020).
14. Health Data Insight, The Simulacrum homepage, https://healthdatainsight.org.uk/project/the-simulacrum, last accessed 24/08/2022.
15. Eck, M. L. V., Lu, X., Leemans, S. J., & Van Der Aalst, W. M.: PM^2: a process mining project methodology. In: International conference on advanced information systems engineering, pp. 297-313. Springer, Cham. (2015)
16. Health Data Insight, Testing the Simulacrum homepage, https://healthdatainsight.org.uk/project/testing-the-simulacrum, last accessed 24/08/2022.
17. Fraunhofer Institute for Applied Information Technology (FIT), PM4PY (2.2.24) [Software]. (2022)
18. Van der Aalst, W. M., van Dongen, B. F., Günther, C. W., Rozinat, A., Verbeek, E., & Weijters, T.: ProM: The process mining toolkit. In: BPM (Demos), 489(31), 2, (2009).
19. Brain trust homepage, https://brainstrust.org.uk, last accessed 24/08/2022.
20. Buijs, J. C., Dongen, B. F. V., & van Der Aalst, W. M.: On the role of fitness, precision, generalization and simplicity in process discovery. In: OTM Confederated International Conferences "On the Move to Meaningful Internet Systems", pp. 305-322. Springer, Berlin, Heidelberg. (2012).
21. Baker, K., Dunwoodie, E., Jones, R. G., Newsham, A., Johnson, O., Price, C. P., ... & Hall, G.: Process mining routinely collected electronic health records to define real-life clinical pathways during chemotherapy. In: International journal of medical informatics, 103, 32-41 (2017).
22. Kurniati, A. P., Rojas, E., Zucker, K., Hall, G., Hogg, D., & Johnson, O.: Process Mining to Explore Variations in Endometrial Cancer Pathways from GP Referral to First Treatment. In: Studies in health technology and informatics, 281, 769–773 (2021).