Synthetic Longitudinal Patient Data Generation

Katariina Perkonoja, Department of Mathematics and Statistics, University of Turku, Finland

Introduction

The proliferation of data has expedited research and innovation activities. However, not all industries have benefited equally from the surge in data availability, partly due to legal restrictions on data usage and privacy regulations. One of the suggested solutions for tackling this problem is *synthetic data generation.*

This poster

1) Briefly introduces the basic concept of synthetic data
2) Describes the special features of longitudinal patient data
3) Gives a preview of the results of an ongoing systematic literature review on the topic

What are synthetic data?

- Synthetic data (SD) are artificially generated data that mimic the statistical properties and patterns of real-world data
- SD are created using algorithms or models to replicate the characteristics, structure, and relationships found in existing datasets
- Synthetic data can be used for various purposes, such as privacy protection, data augmentation, and testing machine learning models, without the need to access or expose actual sensitive or confidential information
- The terms synthetic and simulated data are sometimes used interchangeably, but unlike SD, simulated data are not always based on existing data

Figure 1. Real and synthetic human faces randomly sampled from a dataset[1]. Can you spot the difference? The correct answers can be found in the lower right corner of this poster.

Special features of longitudinal patient data

Longitudinal patient data can take different forms depending on the selected study design and analysis methods. The common forms and typical characteristics are illustrated in Figure 2 below.

Balanced longitudinal data

Patient ID	Sex	Race	Intervention	Age	Height	Weight	CCI	Albumin	CRP	QoL	⋯	Visit
1	Male	Black	Case	40	187	80	5	37.3	3.8	88	⋯	Baseline
1	Male	Black	Case	40	187	82	5	NA	NA	88	⋯	3 months
1	Male	Black	Case	40	187	81	5	37.4	3.8	89	⋯	6 months
2	Female	Caucasian	Control	27	164	58	12	42.1	4.5	100	⋯	Baseline
2	Female	Caucasian	Control	27	NA	NA	12	42.0	4.3	100	⋯	3 months
2	Female	Caucasian	Control	27	164	58	13	42.2	4.5	100	⋯	6 months
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
n	Female	Hispanic	Case	56	175	68	10	38.7	3.6	60	⋯	Baseline
n	Female	Hispanic	Case	56	175	70	10	38.5	3.7	NA	⋯	3 months
n	Female	Hispanic	Case	57	176	71	10	38.4	3.7	72	⋯	6 months

(a)

Unbalanced longitudinal data

Patient ID	Sex	Race	Family	⋯	Visit
1	Male	Black	1	⋯	Baseline
1	Male	Black	1	⋯	40 weeks
2	Female	Asian	1	⋯	Baseline
2	Female	Asian	1	⋯	12 weeks
2	Female	Asian	1	⋯	24 weeks
2	Female	Asian	1	⋯	40 weeks
⋮	⋮	⋮	⋮	⋱	⋮
n	Female	Caucasian	k	⋯	Baseline
n	Female	Caucasian	k	⋯	12 weeks
n	Female	Caucasian	k	⋯	24 weeks

(b)

Balanced longitudinal data

Patient ID	Sex	Race	Intervention	Age_B	Age_3	Age_6	Height_B	Height_3	Height_6	Weight_B	Weight_3	Weight_6	CCI_B	CCI_3	CCI_6	Albumin_B	⋯
1	Male	Black	Case	40	40	40	187	187	187	80	82	81	5	5	5	37.3	⋯
2	Female	Caucasian	Control	27	27	27	164	NA	164	58	NA	58	12	12	13	37.3	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱
n	Female	Hispanic	Case	56	56	56	175	175	176	68	70	71	10	10	10	37.3	⋯

(c)

Figure 2. The subfigure (a) shows *balanced,* and the subfigure (b) shows *unbalanced* longitudinal data in *long format*. The third subfigure (c) illustrates the same data as (a) but in *wide format.* Repeated measurements create a unique temporal structure that is essential to preserve when generating synthetic data. Moreover, missing data (NA), measurement errors (176 in (a)), and dropouts (last line in (b)) are common issues encountered in longitudinal data that can impede synthetic data generation.

Systematic literature review

To identify the existing methods for generating synthetic longitudinal patient data, we conducted a systematic literature review. The protocol[2] describing all procedures was registered in the PROSPERO database (CRD42021259232).

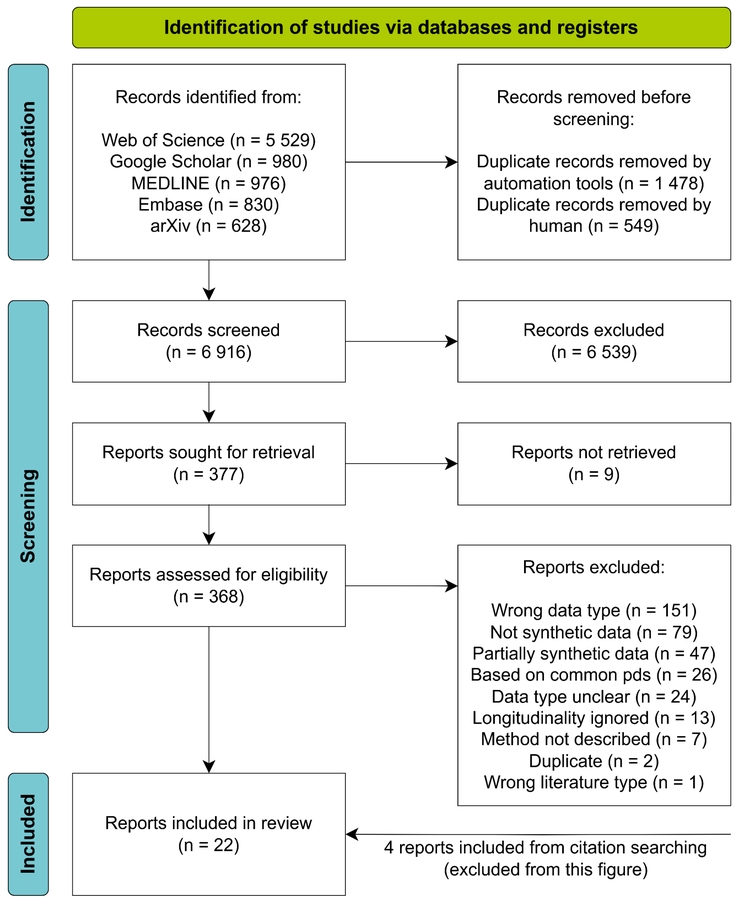

Figure 3. A total of 22 reports were identified. The literature covers English-language articles published in peer-reviewed journals, proceedings, pre-prints, books, book chapters and reviews published by the end of 2022.

Preliminary results from the ongoing systematic review

- The notions of synthetic and longitudinal data lack clarity, which poses challenges in identifying pertinent literature
- The predominant methods for creating synthetic longitudinal patient data are those based on generative adversarial networks and Bayesian networks
- Most methods require longitudinal data in a wide and balanced format and fail to account for missing data
- At present, there is a lack of consensus and standardization regarding the evaluation of synthetic data quality
- The assessment of privacy in synthetic data is difficult and often neglected
- There exist substantial disparities in the quality of methodological descriptions, which may hinder the reuse and development of the methods

References, Funding & Acknowledgements

1. 140k Real and Fake Faces dataset available at https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces
2. The systematic review protocol available at https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=259232

This work was funded by the Finnish Cultural Foundation (grant number 00220801)
A special thanks to the co-authors of the systematic literature review, Joni Virta and Kari Auranen

View PDF

LinkedIn

UNIVERSITY OF TURKU

Top left to right: real, fake, fake. Bottom left to right: real, real, fake.