BioData Mining

RESEARCH

Open Access

Enhancing clinical outcome predictions through effective sample size evaluation in graph-based digital twin modeling

Xi Li[1†], Jui-Hsuan Chang[1†], Mythreye Venkatesan[1†], Zhiping Paul Wang[1] and Jason H. Moore[1*]

[†]Xi Li, Jui-Hsuan Chang and Mythreye Venkatesan contributed equally to this work.

*Correspondence: jason.moore@csmc.edu

[1] Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA 90069, USA

Abstract

Digital twins in healthcare offer an innovative approach to precision diagnosis, prognosis, and treatment. SynTwin, a novel computational methodology to generate digital twins using synthetic data and network science, has previously shown promise for improving prediction of breast cancer mortality. In this study, we validate SynTwin using population-level data for different cancer types from the Surveillance, Epidemiology, and End Results (SEER) program from the National Cancer Institute (USA). We assess its predictive accuracy across cancer types of varying sample sizes ($n = 1{,}000$ to 30,000 records), mortality rates (35% to 60%), and study designs, revealing insights into the strengths and limitations of digital twins derived from synthetic data in mortality prediction. We also evaluate the effect of sample size ($n = 1{,}000$ to 70,000 records) on predictive accuracy for selected cancers (non-Hodgkin lymphoma, bladder, and colorectal cancers). Our results indicate that for larger datasets ($n > 10{,}000$) including digital twins in the nearest network neighbor prediction model significantly improves the performance compared to using real patients alone. Specifically, AUROCs ranged from 0.828 to 0.884 for cancers such as cervix uteri and ovarian cancer with digital twins, compared to 0.720 to 0.858 when using real patient data. Similarly, among the selected three cancers, AUROCs using digital twins exceeded AUROCs using real patients alone by at least 0.06 with narrowing variance in performance as the sample size increased. These results highlight the benefit of network-based digital twins, while emphasizing the importance of considering effective sample size when developing predictive models like SynTwin.

Keywords: Digital twins, Synthetic data, Effective sample size, Reproducibility, Precision medicine

Introduction

Digital twins represent an innovative approach to healthcare, encompassing management and delivery, disease treatment and prevention, and health well-being maintenance, ultimately enhancing human life [8]. This concept is particularly appealing for precision medicine, as it enables the creation of virtual representations of patients for personalized diagnosis, prognosis, and treatment. Notable examples include

MRI-based digital models predicting treatment responses in triple-negative breast cancer [16], virtual clinical trials identifying patients who may benefit from new treatments [11], personalized atrial digital twins predicting iatrogenic scar-related atrial tachycardia in patients with persistent atrial fibrillation [13], and patient-physician digital twin dyads optimizing chemotherapy and radiation regimens in head and neck cancer patients [15].

A novel approach, SynTwin, combines synthetic patient data with network science to create digital twins for precision medicine [9]. The SynTwin methodology leverages patient similarity networks to enhance the prediction of clinical endpoints. First, a network is constructed where patients are nodes with edges based on feature distances, defining patient communities. Synthetic patients are then generated to model real-world correlations in the data, and digital twins are selected from this synthetic population to predict mortality in real patients within the patient communities. Applied to a large cancer registry dataset ($n = 87{,}674$) from the SEER program, the SynTwin method significantly outperformed mortality predictions for breast cancer, with digital twins providing better accuracy than using real data alone. These findings suggest that a network-based digital twin strategy incorporating synthetic patients could improve precision medicine efforts.

Despite the promising potential of SynTwin, two critical challenges remain unresolved and require further investigation. The first challenge is reproducibility. Research indicates that over 70% of scientists have attempted and failed to reproduce another researcher's experiments, and more than half have failed to reproduce their own experiments, highlighting a significant reproducibility crisis in scientific research [2]. The second challenge is determining the effective sample size. In medical research, an adequate sample size is crucial to control the risk of false-negative findings (Type II errors) and to ensure precise estimates of experimental outcomes [3]. Studies with small sample sizes tend to have wide confidence intervals, whereas larger samples can provide more precise estimates [7]. Given that extracting, cleaning, and curating data is often costly and time-consuming [1], determining the minimum effective sample size for an algorithm can lead to more efficient use of resources in medical research.

In this study, we further explored the contribution of digital twin strategies in medicine by evaluating the SynTwin approach in two specific scenarios: 1) assessing whether the performance of SynTwin in predicting breast cancer mortality can be reproduced for cancers with different mortality rates, and 2) evaluating the effect of varying sample sizes on prediction performance.

We applied the SynTwin approach to predicting mortality in population-based cancer registries from the Surveillance, Epidemiology, and End Results (SEER) program from the National Cancer Institute (USA). Seven cancers were selected, with sample sizes ranging from $n = 1{,}000$ to 30,000, to assess whether prior results with SynTwin translated to other cancer type. Our results show that for sample sizes exceeding $n = 10{,}000$, mortality predictions using digital twins were significantly more accurate than those based solely on real data across various cancer types. These findings indicate that the SynTwin approach is effective in multiple cancer domains, regardless of mortality rates, and achieves optimal performance with a proper size of samples being used for prediction.

Methods

The study employed the SynTwin algorithm [9] in two experiments. In the first experiment, the algorithm was applied to a dataset with sample sizes ranging from $n = 1{,}000$ to 30,000 records, with seven cancers: ovary, cervix uteri, small intestine, hepatic flexure, nasopharynx, oropharynx, and other urinary organs. In the second experiment, SynTwin was used to analyze datasets from three cancers: colorectal, non-Hodgkin lymphoma, and bladder, across varying sample sizes from $n = 1{,}000$ to 70,000 for each cancer type. This section provides the specifics of the data used and the experimental setup employed to derive the results.

Dataset

The Surveillance, Epidemiology, and End Results (SEER) database, administered by the National Cancer Institute (NCI), collects cancer incidence data from population-based cancer registries in the United States. For our study, we utilized the SEER*Stat Version 8.4.1 software to access the SEER Research Data: *Incidence—SEER Research Data, 17 Registries, Nov 2022 Sub (2000–2020)—Linked To County Attributes—Time Dependent (1990–2021) Income/Rurality, 1969–2021 Counties*. We applied specific filter criteria to extract relevant patient data,

{Site and Morphology.Site recode ICD-O-3/WHO 2008} = 'All Sites'

AND {Race, Sex, Year Dx.Year of diagnosis} = '2010','2011','2012','2013','2014','2015'

AND ({Cause of Death (COD) and Follow-up.Vital status recode (study cutoff used)} = 'Alive'

OR ({Cause of Death (COD) and Follow-up.Vital status recode (study cutoff used)} = 'Dead'

AND {Cause of Death (COD) and Follow-up.SEER cause-specific death classification} = 'Dead (attributable to this cancer dx)'))

Our filtering yielded a dataset of 2,044,665 unique cases across 80 different cancer sites. This range includes the largest sample size, with 318,134 unique cases for breast cancer, and the smallest sample size, with only 122 unique cases for pleural cancer. The previous study focused solely on the largest sample size: breast cancer. To ensure a balanced representation for predictive modeling, they employed a stratified sampling approach based on vital status. In contrast, our study utilized a criterion based on sample size, ranging from $n = 1{,}000$ to 30,000 records, with a deceased percentage between approximately 35% and 60%. This criterion was designed to ensure a diverse representation across different sample sizes and a balanced distribution of deceased cases. Specifically, we included seven distinct cancers: ovary ($n = 30{,}699$), cervix uteri ($n = 17{,}855$), small intestine ($n = 9{,}942$), hepatic flexure ($n = 5{,}204$), nasopharynx ($n = 2{,}990$), oropharynx ($n = 2{,}050$), and other urinary organs ($n = 1{,}308$). Additionally, to compare the effects of different sample sizes within the same cancer type using SynTwin, we selected three distinct cancers: colorectal ($n = 169{,}064$), non-Hodgkin lymphoma ($n = 81{,}161$), and bladder ($n = 72{,}928$), each with deceased percentage between 30 and 50%. These cancer types were chosen from the 12 most common cancers, each with more than 70,000 cases, based on data from the SEER cancer statistics: common cancer sites [10]. We then performed stratified sampling with varying sample sizes of $n = 1{,}000$, 5,000,

10,000, 20,000, 50,000, and 70,000 to assess the performance of the model across different dataset sizes. We partitioned each sample into two sets: a training dataset comprising approximately two-thirds of the sample, and a validation dataset comprising the remaining one-third. The training data was used to generate the digital twins, while the validation dataset was reserved for making predictions on real patient data by constructing the network and communities. The data processing steps are outlined by the flowchart in Fig. 1.

The features we included are age, year of diagnosis, sex, grade, sequence number, combined summary stage, race, ICDO3, laterality, primary site, diagnostic confirmation, and ICCC site. These features are common across different types of cancers, which allows for a standardized approach in cancer research. We utilized the binary outcome of vital status (alive or dead) as a prediction variable. To comply with previous studies, these features have been widely accepted as cancer-related and are essential for comparing and analyzing various cancer types consistently. Based on this data selection and processing, we constructed a retrospective cohort study to evaluate predictive performance across various cancer types and sample sizes.

Study design

The syntwin algorithm

The SynTwin algorithm [9] facilitates the creation of digital twins through generating the synthetic data and constructing of a community network structure to predict patient outcomes. Initially, Gower's distances [6] are computed between features. Gower's distance is a similarity measure that effectively handles both categorical and continuous variables by calculating separate distances for each feature type and then averaging these to obtain an overall distance score. This makes Gower's distance particularly robust for datasets with diverse feature types, as demonstrated in our prior studies. Subsequently, a graph network is established, where each data point represents a node, and distances serve as weights. Edge connections are filtered by determining the inflection point of a sigmoidal curve fitted to the relationship between distance thresholds and the number of connections.

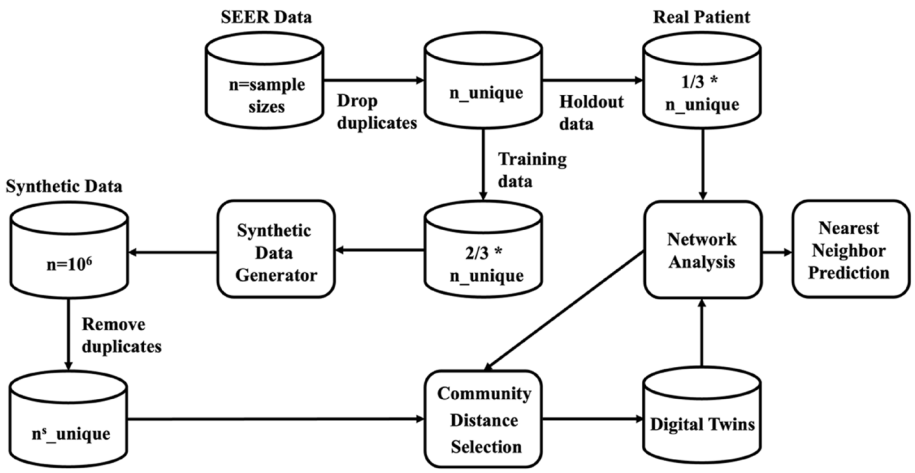

Fig. 1 Flowchart for data processing and analysis

In the third phase, SynTwin employs the multilevel algorithm [4] to detect large communities, adjusting resolution parameter settings to maximize the number of communities containing at least 10 nodes. This threshold was selected to balance granularity and statistical stability: higher resolution values generally yield more, smaller communities, while lower values result in fewer, larger ones. The fourth step utilizes the mixture of product of multinomials (MPoM) [5], a probability-based model chosen for its superior performance compared to other synthetic algorithms in our previous experiments, to generate synthetic data.

Next, digital twins are selected for each community based on a distance criterion. Specifically, for each community, we first identify the central real patient node and then calculate the distance from this central node to the farthest real patient node within the same community. Digital twins are assigned to the community if their distance to the central node is less than this maximum within-community distance. Finally, SynTwin employs a majority vote approach to predict mortality using features from both real patients and digital twins. The performance is evaluated using the area under the receiver operating characteristic curve (AUROC), calculated across 1000 bootstrapped samples from each large community's validation dataset.

Experiment setup

In our experimental design, we first compare seven different size cancers: ovary, cervix uteri, small intestine, hepatic flexure, nasopharynx, oropharynx, and other urinary organs to the baseline, breast cancer, in the previous result. Second, we compare different size within following three cancers: colorectal, non-Hodgkin lymphoma, and bladder. For all cancers, if a community is built on n_r real patients, we evaluate mortality prediction in target patients using the following setups: (A) the remaining $n_r - 1$ real patients; (B) all n_s digital twins from the same community; (C) a combination of the $n_r - 1$ real patients and n_s digital twins; (D) the $n_r - 1$ closest digital twins selected from the n_s; (E) those same $n_r - 1$ closest digital twins together with the $n_r - 1$ real patients; and (F) $n_r - 1$ real patients sampled from outside the community. Figure 2 presents one representative example to illustrate the six experimental setups.

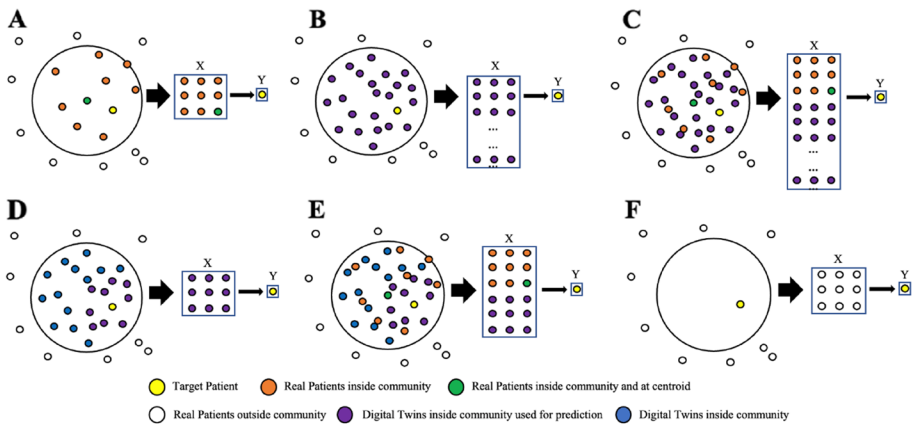

Fig. 2 Study design for comparing outcome prediction using real patients and/or digital twins. The large circles represent a community within the patient network

Results

Table 1 presents an overview the AUROC performance of SynTwin for predicting mortality in the holdout data across various cancer sites for each study design (A-F). As sample size increased, the range of confidence intervals narrowed down, indicating improved precision in the estimates. This trend was particularly evident in cancers with larger sample sizes (small intestine, cervix uteri, ovary, and breast), where AUROC values stabilized with ranges of 0.02–0.04, allowing for more confident comparisons between study designs. For cancers with more than 10,000 cases, significantly higher AUROCs were observed for study designs D and E compared to the other designs, ranging from 0.835 and 0.828 for cervix uteri (0.720 for A) to 0.882 and 0.884 for ovarian cancer (0.858 for A), respectively. Cancers with fewer than 10,000 cases (hepatic flexure, nasopharynx, oropharynx, and other urinary organs) had wider confidence intervals, ranging from 0.06–0.22, with overlapping performances across study designs.

Figure 3 shows the performance of different study designs (A-F) as the dataset size varies for each cancer site. Study designs D and E consistently perform better than the other designs, with a minimum improvement in AUROC of 0.06, when compared to study design A. The gap between the confidence intervals of different study designs increases as the sample size increases while the range of confidence intervals decreases, as shown in the average variance plot (Fig. 4). These results are consistent with the trends observed across multiple cancer sites in Table 1.

The comparable performance of SynTwin across different study designs in smaller sample sizes highlights the challenges associated with predicting mortality in less extensive datasets. Factors such as data heterogeneity and the inherent complexity of smaller datasets may contribute to these results. Interestingly, SynTwin exhibited consistent performance across cancers with large sample sizes, irrespective of the variation in the

Table 1 Comparison of SynTwin performance as measured by AUROC on various cancer datasets

Approximate sample size (Mortality rate %)	Cancer site	Design[a]	A	B	C	D	E	F
90,000 (50%)	Breast [9]	mean	0.791	0.784	0.783	**0.864**	**0.852**	0.494
		95% CI	[0.781, 0.800]	[0.774, 0.794]	[0.773, 0.793]	**[0.857, 0.872]**	**[0.844, 0.860]**	[0.482, 0.507]
30,000 (58.53%)	Ovary	mean	0.858	0.848	0.842	**0.882**	**0.884**	0.431
		95% CI	[0.843,0.872]	[0.833,0.864]	[0.827,0.858]	**[0.868,0.895]**	**[0.871,0.897]**	[0.411,0.452]
20,000 (34.65%)	Cervix uteri	mean	0.720	0.708	0.702	**0.835**	**0.828**	0.496
		95% CI	[0.700, 0.741]	[0.687, 0.729]	[0.680, 0.723]	**[0.819, 0.851]**	**[0.811, 0.846]**	[0.473, 0.519]
10,000 (38.49%)	Small intestine	mean	0.783	0.772	0.761	**0.872**	**0.861**	0.468
		95% CI	[0.761,0.804]	[0.751,0.794]	[0.738,0.783]	**[0.856,0.888]**	**[0.843,0.878]**	[0.442,0.493]
5,000 (46.25%)	Hepatic flexure	mean	0.768	0.789	0.780	0.852	0.844	0.554
		95% CI	[0.730,0.807]	[0.752,0.826]	[0.743,0.818]	[0.821,0.883]	[0.812,0.876]	[0.509,0.599]
3,000 (44.31%)	Naso-pharynx	mean	0.597	0.594	0.578	**0.708**	**0.709**	0.552
		95% CI	[0.540,0.654]	[0.541,0.648]	[0.524,0.632]	**[0.657,0.759]**	**[0.660,0.759]**	[0.493,0.612]
2,000 (56.83%)	Oropharynx	mean	0.671	0.683	0.664	0.746	0.768	0.438
		95% CI	[0.608,0.733]	[0.620,0.746]	[0.599,0.728]	[0.687,0.804]	[0.712,0.820]	[0.368,0.508]
1,000 (59.56%)	Other urinary organs	mean	0.467	0.530	0.466	0.670	0.635	0.522
		95% CI	[0.356,0.578]	[0.419,0.640]	[0.355,0.578]	[0.562,0.778]	[0.522,0.749]	[0.419,0.626]

[a] Real patients (A), digital twins (B), real patients and digital twins (C), closest digital twins (D), real patients and closest digital twins (E), and real patients outside the community (F). Bolded metric values are significantly better than the others

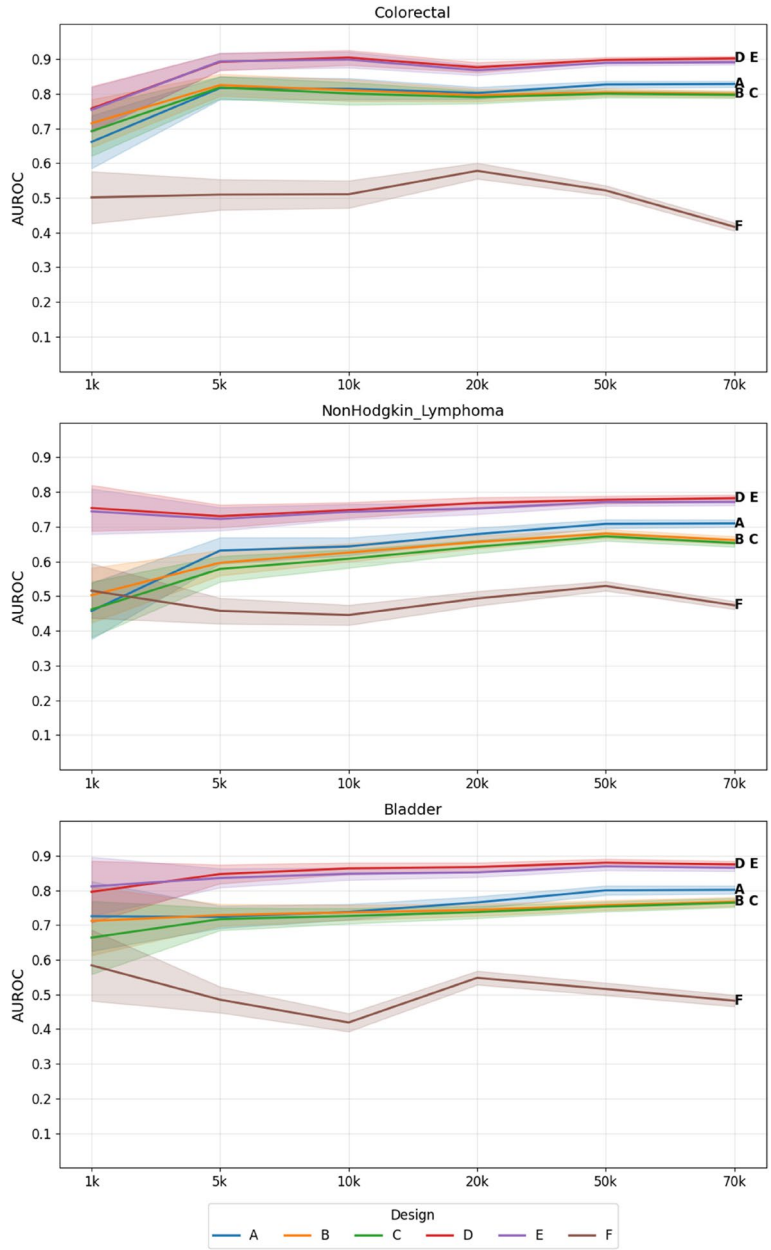

Fig. 3 SynTwin performance as measured by AUROC across different sample sizes and study designs for each cancer site. *Real patients (**A**), digital twins (**B**), real patients and digital twins (**C**), closest digital twins (**D**), real patients and closest digital twins (**E**), and real patients outside the community (**F**)

percentage of mortality among different cancer types, from 35% for small intestine to 58% for ovarian cancer.

Discussion

We have demonstrated the application of SynTwin, a digital twin approach that uses network science and synthetic data, to different cancer sites and evaluated the method's effective sample size by predicting the mortality of cancer patients. By analyzing the performance across different sample sizes and study designs, we gained a deeper

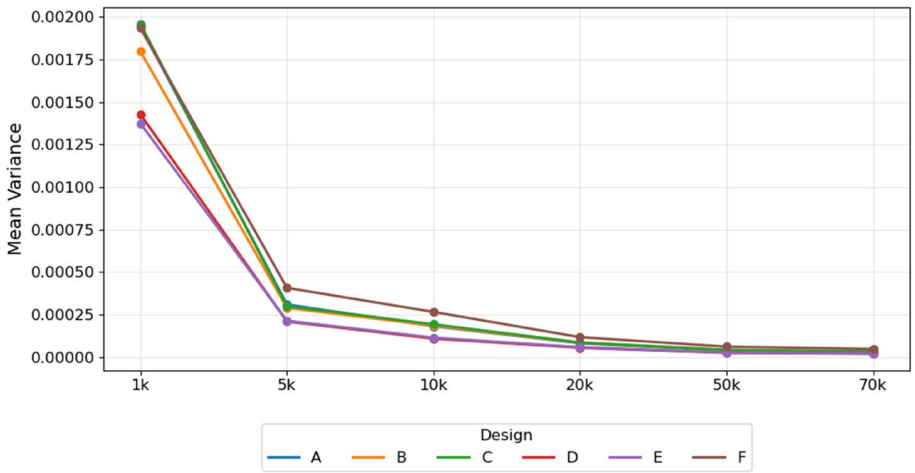

Fig. 4 Average variance of AUROC across the three cancer sites by study design and sample size. *Real patients (**A**), digital twins (**B**), real patients and digital twins (**C**), closest digital twins (**D**), real patients and closest digital twins (**E**), and real patients outside the community (**F**)

understanding of the strengths and limitations of digital twins in mortality prediction across diverse cancer contexts. Despite variations in mortality rates, SynTwin consistently performed well across cancers with sample sizes exceeding 10,000 cases, highlighting its robustness in mortality prediction. Our findings support the hypothesis that network-based digital twins improve predictive accuracy by closely resembling the target patient, rather than relying solely on real patient data [9].

This study underscores the importance of considering sample size and mortality estimates when developing and evaluating predictive models like SynTwin, as the method maintained robust performance despite varying sample sizes and mortality ranges. However, several considerations for future research remain. One key next step is expanding the set of clinical variables to further validate SynTwin's predictive capabilities across different disease domains and patient populations. The synthetic dataset generator used in this study (MPoM) only models categorical features, and therefore cannot be directly applied to numerical features. Moreover, scaling SynTwin to large datasets with more features might be computationally intensive, making it essential to optimize the method for efficiency. Artificial intelligence and deep learning techniques could refine synthetic data generation, improving SynTwin's adaptability to a broader range of datasets.

As SynTwin is applied to diverse datasets, feature selection will become an increasingly important challenge. The current study, using the SEER data, relied on a small set of features deemed most relevant for cancer mortality prediction. However, selecting the most informative features is essential for maintaining predictive accuracy and computational efficiency in larger datasets. A promising approach is integrating expert knowledge from biomedical knowledge bases (KB) for feature selection [14], which could refine patient similarity networks and improve synthetic patient generation. Prior work, such as our study on a knowledge graph for Alzheimer's disease (AlzKB), has demonstrated how KB-driven approaches can prioritize meaningful features in biomedical modeling [12]. Incorporating KBs into SynTwin could improve both interpretability and predictive performance of the model by leveraging curated medical insights for patient

similarity calculations and synthetic patient generation. Additionally, large language models (LLMs) could further enhance SynTwin by integrating insights from medical literature and KBs, extracting patient- and disease-specific information, and supporting clinical decision-making. Future work should explore these avenues to improve the precision, interpretability, and clinical utility of SynTwin in precision medicine.

SynTwin shows promise in predicting mortality among cancer patients, with consistent performance observed in larger datasets and across various cancer types. The trends in performance reported in this study highlights the importance of using adequate sample size in predictive modeling. Optimizing sample size enhances the adaptability of SynTwin, thereby guiding future research and facilitating its application in clinical settings. Expanding its use beyond cancer mortality prediction could open new opportunities for personalized, data-driven decision-making in healthcare.

Authors' contributions
X.L., J.C., and M.V. wrote the main manuscript text and prepared result tables and figures. Z.W. substantially contributed to the study design. J.M. provided leadership and conducted a critical review and refinement of the study. All authors reviewed the manuscript.

Funding
This work was supported by NIH grants LM010098 and AG066833.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 February 2025 Accepted: 8 April 2025
Published online: 15 April 2025

References

1. Awrahman BJ, Aziz Fatah C, Hamaamin MY. A Review of the Role and Challenges of Big Data in Healthcare Informatics and Analytics. Comput Intell Neurosci. 2022;2022:5317760.
2. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533:452–4.
3. Biau, D. J., Kernéis, S. & Porcher, R. Statistics in Brief: The Importance of Sample Size in the Planning and Interpretation of Medical Research. Clinical Orthopaedics and Related Research® 466, 2282 (2008).
4. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;2008:P10008.
5. Dunson DB, Xing C. Nonparametric Bayes Modeling of Multivariate Categorical Data. J Am Stat Assoc. 2009;104:1042–51.
6. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics. 1971;27:857–71.
7. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J. 2003;20:453–8.
8. Katsoulakis, E. et al. Digital twins for health: a scoping review. npj Digit. Med. 7, 1–11 (2024).
9. Moore JH, et al. SynTwin: A graph-based approach for predicting clinical outcomes using digital twins derived from synthetic patients. Pac Symp Biocomput. 2024;29:96–107.
10. National Cancer Institute. Cancer stat facts: common cancer sites. https://seer.cancer.gov/statfacts/html/common.html (2024).
11. Qi, T. & Cao, Y. Virtual clinical trials: A tool for predicting patients who may benefit from treatment beyond progression with pembrolizumab in non-small cell lung cancer. CPT: Pharmacometrics & Systems Pharmacology 12, 236–249 (2023).
12. Romano JD, et al. The Alzheimer's Knowledge Base: A Knowledge Graph for Alzheimer Disease Research. J Med Internet Res. 2024;26: e46777.
13. Sakata K, et al. Optimizing the Distribution of Ablation Lesions to Prevent Postablation Atrial Tachycardia: A Personalized Digital-Twin Study. JACC Clin Electrophysiol. 2024;10:2347–58.
14. Shao, S., Henrique Ribeiro, P., Ramirez, C. M. & Moore, J. H. A review of feature selection strategies utilizing graph data structures and Knowledge Graphs. Brief Bioinform 25, bbae521 (2024).

15. Tardini E, et al. Optimal Treatment Selection in Sequential Systemic and Locoregional Therapy of Oropharyngeal Squamous Carcinomas: Deep Q-Learning With a Patient-Physician Digital Twin Dyad. J Med Internet Res. 2022;24: e29455.
16. Wu C, et al. MRI-Based Digital Models Forecast Patient-Specific Treatment Responses to Neoadjuvant Chemotherapy in Triple-Negative Breast Cancer. Can Res. 2022;82:3394–404.

Publisher's Note