Property-Driven Synthetic Data Engineering for Data-Scarce Software Systems: Reflections from the Breast Cancer Domain

Aurora Francesca Zanenga
aurora.zanenga@unibg.it
University of Bergamo
Bergamo, Italy

Andrea Bombarda
andrea.bombarda@unibg.it
University of Bergamo
Bergamo, Italy

Marsha Chechik
chechik@cs.toronto.edu
University of Toronto
Toronto, Canada

Saverio D'Amico
saverio.damico@humanitas.it
Humanitas Clinical and Research Center, IRCCS
Rozzano, Italy

Rita De Sanctis
rdesanctis@asst-pg23.it
ASST Papa Giovanni XXIII
Bergamo, Italy

Alberto Zambelli
azambelli@asst-pg23.it
ASST Papa Giovanni XXIII
Bergamo, Italy

Claudio Menghi
claudio.menghi@unibg.it
University of Bergamo
Bergamo, Italy
McMaster University
Hamilton, ON, Canada

arXiv:2607.06133v1 [cs.SE] 7 Jul 2026

Abstract

Modern software systems increasingly depend on data for analysis, prediction, testing, and decision-making. Yet many important domains, including medicine, safety-critical systems, and regulated industries, lack abundant, shareable, or representative data. Synthetic data generation is often proposed as a remedy, but our experience engineering software for intraoperative radiotherapy (IORT) in breast cancer treatment suggests that synthetic data shifts rather than solves the central engineering problem. The key challenge becomes deciding which properties synthetic data must preserve, how these properties should be elicited from stakeholders, how they can be validated under privacy constraints, and how they evolve. We call this problem *property-driven synthetic data engineering*. Drawing on a collaboration with oncologists and preliminary experiments with a sensitive IORT dataset, we identify challenges in requirements, validation, privacy, and pipeline evolution. We argue that automated software engineering research should develop methods and tools for eliciting, formalizing, checking, and evolving validity properties for synthetic data in data-scarce software systems.

CCS Concepts

• **Software and its engineering**;

Keywords

Property-Driven Synthetic Data, Breast Cancer, Data-Scarce Software Systems

1 Introduction

Data are driving modern software applications. Tasks such as prediction, classification, testing, and decision-making rely on the availability of large, accessible datasets, and the increasing adoption of artificial intelligence further reinforces this assumption. However, in many safety-critical and privacy-sensitive domains, this assumption does not hold. In domains such as medicine and regulated industries, data are often scarce, difficult to access, and subject to strict confidentiality constraints. As a result, software engineering (SE) practices that implicitly rely on abundant data become difficult to apply. This is the case with our partner, a large hospital with more than one million outpatient services and 100,000 emergency department visits. Despite the large number of patients, in many medical scenarios, the number of patients affected by certain pathologies is (fortunately) limited, and the data contain sensitive health information. For example, in the case of breast cancer, while the overall number of cases is significant (approximately 2.3 million new cases each year [13]), the therapeutic landscape [5] is rapidly evolving, and the number of patients within specific subpopulations (e.g., young women or minority groups) may be very small or even absent. Consequently, engineering software relying on such data introduces specific challenges that contrast with mainstream SE practices based on abundant and shareable datasets.

To address data scarcity, Synthetic Data Generation (SDG) techniques create artificial datasets that mimic properties (e.g., patterns, distributions, and correlations) of existing datasets. These techniques are widely proposed as a remedy in contexts where data are limited or cannot be shared due to privacy constraints [6]. For example, clinical datasets typically contain personal and identifiable information, making their sharing and reuse across institutions challenging. However, our experience suggests that synthetic data do not eliminate the core problem; rather, they shift it. The central challenge becomes determining which properties synthetic data must preserve, how these properties can be elicited from stakeholders, how they can be validated in the absence of ground truth, and how they evolve over time under changing requirements and constraints. We argue that data scarcity fundamentally breaks standard SE assumptions, requiring a shift from data-driven validation

to property-driven synthesis and validation. In such settings, developers cannot rely on large datasets for validation, testing, or model training. Instead, they must reason in terms of properties, constraints, and synthetic approximations. This shift introduces new challenges in requirements engineering, validation, and system design that are not addressed by existing methodologies. To address this gap, this paper proposes *property-driven synthetic data engineering* as a research agenda for automated SE, in which the central artifacts are stakeholder-specific validity properties that must be elicited, formalized, checked, and evolved.

We expose this problem through a case study in breast cancer treatment, namely intraoperative radiotherapy (IORT), which delivers (immediately after the tumor is removed) a high-concentration dose of radiation [18]. While highly valuable for clinical research, the data from our partner hospital cannot be freely shared due to strict privacy constraints; at the same time, enabling access to those data would significantly benefit the research community, allowing for more extensive experimentation and validation of medical and computational approaches. To overcome these challenges, our project envisions the application of SDG solutions and highlights the need to treat SDG not as a pre-processing step, but as an engineering process centered on validity properties.

In this paper, we reflect on our collaboration, preliminary experimentation, challenges we faced, and lessons we learned. We argue that SDG for software systems should be treated not as a data preprocessing technique, but as an engineering process whose central artifacts are stakeholder-specific validity properties. In data-scarce and privacy-constrained settings, developers must elicit, formalize, check, trade off, and evolve properties such as statistical fidelity, clinical plausibility, privacy protection, and task-specific utility. We call this problem *property-driven synthetic data engineering* and identify it as an emerging challenge for automated SE. In particular, the following are the contributions of this paper:

- It identifies data absence as a SE problem that affects requirements, validation, testing, and system evolution in data-driven software.
- It introduces a *property-driven* view of synthetic data engineering, in which synthetic datasets are evaluated against stakeholder-specific properties rather than only against global similarity metrics.
- It reports early lessons from an IORT breast-cancer case study and derives a research agenda for automated support for property elicitation, property checking, trade-off analysis, and pipeline evolution.

This paper is structured as follows. Section 2 presents our research problem. Section 3 presents our preliminary experimentation and results. Section 4 presents challenges and lessons learned. Section 5 presents related work. Section 6 presents our conclusions.

2 Research Problem and Context

This work is based on a collaboration with the oncologists of a major hospital specializing in breast cancer treatment. We aim to support physicians by designing a software system that that supports clinical data-analysis and decision-support pipeline, allowing oncologists to explore IORT outcomes, evaluate recurrence-related factors, and test analysis workflows without exposing identifiable patient data. Our effort in analyzing SDG is intended not to replace clinical evidence, but to support software development, testing, benchmarking, and controlled sharing of analysis pipelines. We received a dataset containing clinical and treatment-related variables collected during IORT procedures. The dataset consists of 1000 patients described by 64 variables (before data cleaning). This is a considerably large dataset for this domain, both in terms of samples (not all the patients affected by breast cancer are treated with IORT) and features (64). This dataset contains sensitive information (e.g., personal information that are highly confidential). The size of the dataset and the sensitivity of the contained information pose significant challenges for engineering software applications that can effectively support physicians. On the one hand, despite being a considerably large dataset for this domain, the limited number of samples hampers the effective use of existing AI solutions that need a significant amount of data to ensure the reliability of the results produced by the analysis. On the other hand, considering the sensitive nature of the data, the distribution of the dataset is forbidden. These constraints make it difficult not only to build models but also to define how results should be validated.

$$\begin{array}{ccccc} D & \longrightarrow & A(D) & \longrightarrow & R \\ f \,\vdots & & & & \vdots\, P \\ D' & \longrightarrow & A(D') & \longrightarrow & R' \end{array}$$

Figure 1: Synthetic data generation: A high-level overview.

SDG is a technique that may address these challenges. Figure 1 outlines our formalization of the *property-driven synthetic data* generation for software applications. Consider a scenario in which a (limited) set of data (D) leads to the results R when processed by an analysis (A). Examples of such an analysis are survival analysis, risk prediction, patient stratification, and testing of a clinical decision-support system. SDG concerns the development of a procedure f that generates a set of data (D') which leads to the results R' when processed by the same analysis (A). Examples of results are Kaplan-Meier survival curves, estimation of Cox model coefficients, model predictions, or test coverage of patient subgroups. We argue that SDG should ensure that R' satisfies a set of properties P, e.g., distributional similarity, clinically legal value ranges, plausible treatment sequences, low privacy leakage, or utility for downstream tasks. Such properties are stakeholder-dependent and potentially conflicting. For example, clinicians may require clinical plausibility, data scientists statistical fidelity, and privacy officers strong disclosure protection. As a result, satisfying one property (e.g., statistical similarity) may hamper the satisfaction of others (e.g., privacy preservation), requiring specific trade-offs. For instance, synthetic data may preserve correlations observed in the original data while producing clinically implausible values or unrealistic combinations of variables, or may expose sensitive information. This highlights that the SE challenge is not only generating synthetic data, but understanding which properties matter for a given task and how they can be validated in the absence of ground truth.

3 Preliminary Experimentation

In our preliminary experimentation, we conducted three activities: (a) cleaning our dataset, (b) selecting and applying an appropriate SDG technique, and (c) analyzing our results.

Dataset Cleaning. We manually cleaned our dataset to improve data quality. We removed columns that contained empty values for more than 80% of the rows (e.g., "Previous Tumor Date", "Chemotherapy Start Date", "Days from IORT to Chemotherapy Start", "Percentage Difference Between Administered and Prescribed Dose", "Mastectomy Date"). Since those columns had a high percentage of missing values, they were considered not relevant in most of the cases. Moreover, due to the large amount of missing data, it was not possible to reliably model their correlations with the other variables. Also, such variables could negatively affect the quality of the SDG process. We fixed typos in entry values (e.g., the word "fulfilled" was often written in different ways, such as "full filled" or "fulfilled"), standardized these entries, and removed rows containing values that did not fall within the set of valid values for a given column (e.g., when the allowed values were 0, 1, or 2, but a value of 5 was present). After pre-processing, the final dataset contained 709 patients and 58 variables (down from 64). We discussed our changes with oncologists to ensure the soundness of our cleaning activity (e.g., we verified that the columns that had been removed did not contain any relevant information). Such a process highlighted a critical challenge: data are not always ready to be used.

Synthetic Data Generation. We reviewed existing SDG techniques from a recent work [15] that could be suitable for our context and have been applied to the medical domain. We selected TabDDPM [9], CTGAN [1], Gaussian Copula [10], CopulaGAN [11], and Tabular Variational Autoencoder (TVAE) [8] since they can process tabular data and have been applied in similar contexts.

Analyzing our results. After some preliminary experiments, we evaluated the quality of the generated data by analyzing both the relationships between variables and their individual distributions. First, we analyzed the correlation matrices. The correlation matrices enable us to assess whether the relationships among variables were preserved. In addition, we analyzed the distributions of synthesized variables, which closely match those of the original variables. While these results suggest that these methods can preserve the statistical properties of the data, they do not necessarily guarantee that the generated data are clinically plausible. This observation further motivates our formalization of the SDG task for SE, where different properties may be considered by different stakeholders and different SDG approaches may be used. Additionally, our analysis showed that different evaluation views lead to different conclusions, so generator selection is a requirements- and validation-related problem, not only an ML model-selection problem.

We observed that TVAE [8], a deep learning generative model built on a Variational Autoencoder (VAE), achieved better results than the other approaches, given the characteristics of our dataset (i.e., a limited number of samples and a relatively high number of variables). Specifically, for TVAE, the two correlation matrices show a high similarity, showing that the model can capture and reproduce the main relationships between variables.

To analyze the clinical validity of the synthetic IORT dataset, we studied the Local Recurrence-Free Survival (LRFS). We compared

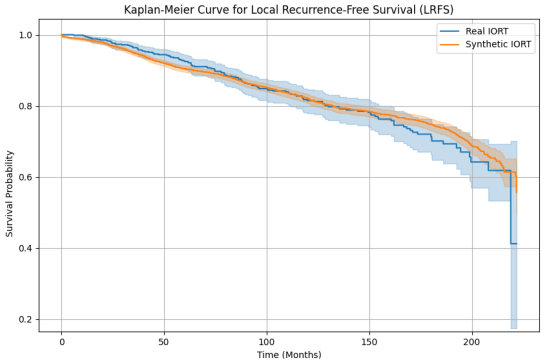

Figure 2: Survival Curves: IORT vs Synthetic IORT

the real and synthetic data using four steps: (I) Global Kaplan-Meier curves, (II) Stratified Kaplan-Meier curves, (III) Univariate Cox models, and (IV) Multivariate Cox models. These analyses enable the verification of patient survival patterns and clinical relationships. For example, we report the results of the Kaplan-Meier analysis, a method used to estimate survival over time. We compared the global Kaplan-Meier curves [14] between the real and synthetic data. The curves in Figure 2 showed a very similar trend over most of the follow-up period. In both datasets, the probability of remaining free from local recurrence gradually decreased over time. Some differences appeared at the end of the curves, where the number of patients was low, making the estimates less reliable. These differences may have clinical or SE implications. Thus, we do not claim that our synthetic data are clinically valid. Rather, our analyses illustrate that a generator can appear statistically plausible while still requiring domain-specific validation before use. Our preliminary experimentation confirms that engineering these systems is complex: It is not a mere data-analytic problem; it is a SE problem that involves multiple dimensions and stakeholders, where the properties P to be considered for the validation of the dataset assume a primary value.

4 Toward Property-Driven Synthetic Data Engineering

Our experience suggests that the central challenge is not choosing a better synthetic data generator, but engineering the validity conditions under which generated data can be trusted for a particular software task. We therefore frame SDG as a *property-driven* engineering process. In this process, stakeholders define the properties that synthetic data must satisfy; developers operationalize those properties as checks; tools compare generators and configurations against those checks; and pipelines evolve as data, tasks, and regulations change. This process provides an additional opportunity for automation for software engineers, i.e, to support developers in moving from informal stakeholder needs to executable validity checks. We report a possible mapping between stakeholders, properties, and automation opportunities in Table 1.

Thanks to our preliminary experimentation, we learned the following *lessons* (L):

L1. Data cleaning is engineering, not pre-processing. The removal of variables and rows changes what properties can be preserved and what conclusions can be drawn: Data cleaning decisions

Table 1: Mapping of stakeholders to properties, checks and automation opportunities

Stakeholder	Desired property	Example property/check	Automation opportunity
Clinician	Clinical plausibility	No impossible treatment/value combinations; survival curves clinically plausible	Constraint mining; rule checking; clinician-in-the-loop validation
Data scientist	Statistical fidelity	Marginal distributions, correlations, task utility	Metric selection and model comparison
Privacy officer	Disclosure protection	Low membership-inference risk; no near-duplicate patients	Privacy risk analysis; synthetic-data release gates
Software Engineer	Test adequacy/Pipeline reliability	Synthetic data exercises target scenarios and edge cases	Property-based test generation; regression testing of data pipelines
Maintainer	Evolution robustness	New data or hospital source does not invalidate properties	Monitoring; drift detection; impact analysis

become part of the system specification because they determine which properties can later be synthesized, validated, and trusted.

L2. Synthetic data validity is multi-property. Statistical similarity, clinical plausibility, privacy, and task utility may conflict.

L3. There is no universal best generator. The choice of generator depends on data characteristics and stakeholder properties.

L4. Synthetic-data pipelines evolve. As new patients, variables, hospitals, and tasks appear, properties must be revisited.

Based on these lessons, we identify three key SE *challenges*.

Requirements under uncertainty. Precisely understanding and eliciting the requirements of the system is challenging. Data may be used by different stakeholders (e.g., oncologists, other physicians, data managers, or legal offices), each of them with different expectations and requirements. For example, oncologists are interested in the clinical validity of the generated data; data managers are concerned about statistical validity; legal offices need to ensure the confidentiality of personal information. Designing solutions that meet all of these requirements is complicated (e.g., we experienced different tradeoffs between clinical validity and confidentiality). A further challenge is the formalization of these requirements. Providing a precise mathematical definition of the properties P is difficult, and small variations in their definition may significantly affect the evaluation of different synthesis solutions. Future tools could mine candidate properties from real data, encode clinical and privacy constraints, detect property conflicts, recommend appropriate synthetic data generators, and continuously monitor whether a synthetic-data pipeline remains valid as the real dataset evolves.

Validation without ground truth. To be valuable for the SE community, synthetic data must preserve specific characteristics that enable obtaining valuable information from their analysis. For example, in the absence of sufficient real data, it is difficult to define ground truth for validation. Privacy constraints prevent cross-institution validation, forcing synthetic data to act as a surrogate for ground truth, which fundamentally alters validation pipelines. We evaluated the generated data by comparing distributions and correlations to ensure that key statistical properties were preserved and performed clinical validation using survival curves [4]. From the software analytics perspective, defining appropriate metrics for assessing the quality of the generated data is a challenge. A single standard metric does not exist; Instead, multiple measures are required, including both statistical evaluation and clinical validation. However, these measures may not agree, as statistical similarity does not necessarily imply clinical validity. Moreover, the characteristics of the data (e.g., heterogeneity, missing values, and imbalance) further complicate the evaluation, as different generation models may behave inconsistently across different types of variables.

Evolution of synthetic pipelines. Data evolves as new patients are analyzed. (Synthesized) Data from other hospitals may need to be integrated. Also, the pathology and the population characteristics change over time (e.g., average age). Thus, new properties may be of interest for stakeholders. Synthetic data pipelines must adapt to these changes. In our context, some variables were removed due to missing values, but they may become relevant again as new data becomes available. This requires continuous adaptation of the synthesis process and validation criteria.

5 Related Work

Recent work has explored SDG techniques for tabular and medical datasets, focusing on improving data availability while preserving statistical and clinical properties. Zazzetti et al. [18] propose a framework for generating longitudinal synthetic breast cancer data using models such as GANs, VAEs, and language models, and introduce a comprehensive validation pipeline (SAFE) to assess fidelity, privacy, and clinical utility. They demonstrate how synthetic data can support clinical research, including predictive modeling and the creation of synthetic control groups. More broadly, several studies analyze generative approaches for tabular data, including GAN-based and VAE-based methods [8, 15]. These techniques aim to reproduce statistical patterns in the original data and are often evaluated through metrics such as distribution similarity, correlation preservation, and downstream task performance. Other work highlights the role of synthetic data in addressing data scarcity, imbalance, and privacy constraints in machine learning applications [2, 7]. While these approaches provide effective generation techniques and validation pipelines, they typically rely on predefined metrics or domain-specific evaluation procedures.

Synthetic data has also been widely used in SE to support testing [3], benchmarking, and empirical studies. Prior work uses synthetic datasets to evaluate security software, perform statistical testing [16], and support experimental reproducibility when real-world data are scarce or proprietary [17]. In these contexts, synthetic data is treated as a practical substitute for real data [12], enabling controlled experimentation and scalability of evaluations.

In contrast, our focus is not on proposing a new generator or validation metric. We argue that SDG introduces an SE problem: stakeholders must define validity properties, developers must operationalize them into checks, tools must reason about trade-offs and conflicts, and pipelines must evolve as data and use cases change.

6 Conclusion and Future Work

This paper reflects on our experience engineering software systems in a data-scarce and privacy-constrained medical domain for supporting oncologists of a major hospital. Our results suggest that data scarcity does not simply limit existing approaches but fundamentally challenges core assumptions of SE. In such settings, developers cannot rely on large datasets for validation or model construction. Instead, they must reason in terms of properties, constraints, and stakeholder-specific requirements. Our preliminary evaluation shows that SDG is not merely a technical solution to data unavailability, but a complex engineering problem which introduces the necessity of considering statistical fidelity, clinical plausibility, and privacy, requiring explicit trade-offs that current methodologies do not adequately support or address.

Our observations highlight a broader shift for the SE community working in data-scarce and data-sensitive domains: from data-driven to property-driven approaches. Validation, requirements engineering, and system evolution must be rethought to explicitly account for limited, sensitive, and evolving data. We therefore argue that future research should focus on methods for defining, balancing, and validating properties in synthetic data. More broadly, SE must move beyond the assumption of abundant data and develop foundations for building and validating systems in its absence.

Data Availability Statement

Due to the sensitivity of the dataset, neither the original nor the synthesized data can be shared at this time, as disclosure risks cannot yet be fully excluded.

References

[1] Halal Abdulrahman Ahmed, Juan A. Nepomuceno, Belén Vega-Márquez, and Isabel A. Nepomuceno-Chamorro. 2025. Synthetic Data Generation for Healthcare: Exploring Generative Adversarial Networks Variants for Medical Tabular Data. *International Journal of Data Science and Analytics* 20, 6 (May 2025), 5739–5754. doi:10.1007/s41060-025-00816-w

[2] Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, José I. Santamaría, A. Albahri, B. S. Al-dabbagh, M. Fadhel, M. Manoufali, Jinglan Zhang, Ali H. Al-timemy, Ye Duan, Amjed Abdullah, Laith Farhan, Yi Lu, Ashish Gupta, Felix Albu, Amin Abbosh, and Yuantong Gu. 2023. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data* 10 (2023), 1–82. doi:10.1186/s40537-023-00727-2

[3] Mohammad Hossein Amini and Shiva Nejati. 2024. Bridging the Gap between Real-world and Synthetic Images for Testing Autonomous Driving Systems. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (Sacramento, CA, USA) *(ASE '24)*. Association for Computing Machinery, New York, NY, USA, 732–744. doi:10.1145/3691620.3695067

[4] J Martin Bland and Douglas G Altman. 1998. Survival probabilities (the Kaplan-Meier method). *BMJ* 317, 7172 (Dec. 1998), 1572–1580. doi:10.1136/bmj.317.7172.1572

[5] Maria-Joao Cardoso, Philip Poortmans, Elżbieta Senkus, Oreste D Gentilini, and Nehmat Houssami. 2024. Breast cancer highlights from 2023: Knowledge to guide practice and future research. *The Breast* 74 (2024), 103674.

[6] Giacomo Fantino, Marco Rondina, Antonio Vetrò, and Juan Carlos De Martin. 2026. Quantifying Privacy Risks in Synthetic Data: A Study on Black-Box Membership Inference. In *Fundamental Approaches to Software Engineering*, Elvira Albert and Corina Pasareanu (Eds.). Springer Nature Switzerland, Cham, 86–106.

[7] Mandeep Goyal and Q. Mahmoud. 2024. A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics* (2024). doi:10.3390/electronics13173509

[8] A Kiran and S Saravana Kumar. 2023. A Comparative Analysis of GAN and VAE based Synthetic Data Generators for High Dimensional, Imbalanced Tabular data. In *2023 2nd International Conference for Innovation in Technology (INOCON)*. 1–6. doi:10.1109/INOCON57975.2023.10101315

[9] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. TabDDPM: Modelling Tabular Data with Diffusion Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 17564–17579. https://proceedings.mlr.press/v202/kotelnikov23a.html

[10] Zheng Li, Yue Zhao, and Jialin Fu. 2020. SynC: A Copula based Framework for Generating Synthetic Data from Aggregated Sources. In *2020 International Conference on Data Mining Workshops (ICDMW)*. 571–578. doi:10.1109/ICDMW51313.2020.00082

[11] Marko Miletic and Murat Sariyar. 2024. Challenges of Using Synthetic Data Generation Methods for Tabular Microdata. *Applied Sciences* 14, 14 (2024). doi:10.3390/app14145975

[12] Vittoriano Muttillo, Claudio Di Sipio, Riccardo Rubei, Luca Berardinelli, and MohammadHadi Dehghani. 2024. Towards Synthetic Trace Generation of Modeling Operations using In-Context Learning Approach. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (Sacramento, CA, USA) *(ASE '24)*. Association for Computing Machinery, New York, NY, USA, 619–630. doi:10.1145/3691620.3695058

[13] World Health Organization and World Health Organization. 2024. WHO Global Breast Cancer Initiative: breast cancer awareness month. *World Health Organization. https://www. who. int/news-room/events/detail/2024/10/01/default-calendar/who-global-breast-cancer-initiative–breast-cancer-awareness-month* (2024).

[14] Jason T. RichMD, J. Gail NeelyMD, Randal C. PanielloMD, Courtney C. J. VoelkerMD DPhil, Brian NussenbaumMD, and Eric W. WangMD. 2010. A practical guide to understanding Kaplan-Meier curves. *Otolaryngology–Head and Neck Surgery* 143, 3 (2010), 331–336. doi:10.1016/j.otohns.2010.05.007 arXiv:https://doi.org/10.1016/j.otohns.2010.05.007 PMID: 20723767.

[15] R. Shi, Y. Wang, M. Du, X. Shen, Y. Chang, and X. Wang. 2025. A Comprehensive Survey of Synthetic Tabular Data Generation. *arXiv preprint arXiv:2504.16506* (2025).

[16] Ghanem Soltana, Mehrdad Sabetzadeh, and Lionel C. Briand. 2017. Synthetic data generation for statistical testing. In *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*. 872–882. doi:10.1109/ASE.2017.8115698

[17] Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D. Mooney, and Bradley A. Malin. 2022. A Multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications* 13, 1 (Dec. 2022). doi:10.1038/s41467-022-35295-1

[18] Elena Zazzetti, Saverio D'Amico, Flavia Jacobs, Rita De Sanctis, Lorenzo Chiudinelli, Mariangela Gaudio, Gianluca Asti, Mattia Delleani, Elisabetta Sauta, Mirco Quintavalla, et al. 2025. Longitudinal synthetic data generation by artificial intelligence to accelerate clinical and translational research in breast cancer. *JCO Clinical Cancer Informatics* 9 (2025), e2500033.