TYPE Original Research
PUBLISHED 05 February 2025
DOI 10.3389/frai.2025.1533508

Check for updates

OPEN ACCESS

EDITED BY
Sokratis Makrogiannis,
Delaware State University, United States

REVIEWED BY
Rocco Zaccagnino,
University of Salerno, Italy
Nikos Kanakaris,
University of Southern California,
United States

*CORRESPONDENCE
Austin A. Barr
✉ austin.barr@ucalgary.ca

RECEIVED 24 November 2024
ACCEPTED 23 January 2025
PUBLISHED 05 February 2025

CITATION
Barr AA, Quan J, Guo E and Sezgin E (2025) Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data. *Front. Artif. Intell.* 8:1533508.
doi: 10.3389/frai.2025.1533508

Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data

Austin A. Barr[1]*, Joshua Quan[1], Eddie Guo[1] and Emre Sezgin[2,3]

[1]Cumming School of Medicine, University of Calgary, Calgary, AB, Canada, [2]The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH, United States, [3]Department of Pediatrics, The Ohio State University College of Medicine, Columbus, OH, United States

Background: Clinical data is instrumental to medical research, machine learning (ML) model development, and advancing surgical care, but access is often constrained by privacy regulations and missing data. Synthetic data offers a promising solution to preserve privacy while enabling broader data access. Recent advances in large language models (LLMs) provide an opportunity to generate synthetic data with reduced reliance on domain expertise, computational resources, and pre-training.

Objective: This study aims to assess the feasibility of generating realistic tabular clinical data with OpenAI's GPT-4o using zero-shot prompting, and evaluate the fidelity of LLM-generated data by comparing its statistical properties to the Vital Signs DataBase (VitalDB), a real-world open-source perioperative dataset.

Methods: In Phase 1, GPT-4o was prompted to generate a dataset with qualitative descriptions of 13 clinical parameters. The resultant data was assessed for general errors, plausibility of outputs, and cross-verification of related parameters. In Phase 2, GPT-4o was prompted to generate a dataset using descriptive statistics of the VitalDB dataset. Fidelity was assessed using two-sample *t*-tests, two-sample proportion tests, and 95% confidence interval (CI) overlap.

Results: In Phase 1, GPT-4o generated a complete and structured dataset comprising 6,166 case files. The dataset was plausible in range and correctly calculated body mass index for all case files based on respective heights and weights. Statistical comparison between the LLM-generated datasets and VitalDB revealed that Phase 2 data achieved significant fidelity. Phase 2 data demonstrated statistical similarity in 12/13 (92.31%) parameters, whereby no statistically significant differences were observed in 6/6 (100.0%) categorical/binary and 6/7 (85.71%) continuous parameters. Overlap of 95% CIs were observed in 6/7 (85.71%) continuous parameters.

Conclusion: Zero-shot prompting with GPT-4o can generate realistic tabular synthetic datasets, which can replicate key statistical properties of real-world perioperative data. This study highlights the potential of LLMs as a novel and accessible modality for synthetic data generation, which may address critical barriers in clinical data access and eliminate the need for technical expertise, extensive computational resources, and pre-training. Further research is warranted to enhance fidelity and investigate the use of LLMs to amplify and augment datasets, preserve multivariate relationships, and train robust ML models.

KEYWORDS

synthetic data, large language model, artificial intelligence, machine learning, ChatGPT, big data, education, surgery

1 Introduction

Clinical data is fundamental to advance medical research and enable the development of machine learning (ML) models. This is particularly relevant in surgical care, as procedural medicine increasingly relies on decision-making based on large-scale data (Maier-Hein et al., 2017). However, access to real-world (real) clinical data is constrained by ethical, legal, and logistical barriers (Pavlenko et al., 2020; Wartenberg and Thompson, 2010). Data requests often require institutional review board approval, data sharing agreements, and compliance with various data privacy regulations (e.g., HIPAA, GDPR, PIPEDA) (Bentzen et al., 2021; Ness, 2007). In many institutions, clinical datasets are proprietary and restricted to internal use. Clinical data also requires significant pre-processing and de-identification procedures which is resource intensive and can delay or hinder research projects—particularly for students, trainees, and early career researchers (Tudur et al., 2017; Willemink et al., 2020). These protections, while essential for patient privacy, limit the accessibility of real clinical data.

In addition to regulatory concerns, clinical data is often incomplete (Newgard and Lewis, 2015). Data collected in clinical settings may suffer from missing values, errors, or biases introduced during data entry. Furthermore, data de-identification processes commonly remove or obscure personal health information (e.g., date of birth, date of operation, demographic data, geographic data). These constraints limit the reliability of analyses as well as the accuracy and generalizability of ML models trained on this data. Collectively, these challenges underscore the need for alternative approaches to provide researchers, learners, and developers with necessary data while preserving patient privacy.

Synthetic clinical datasets, which are artificially generated rather than captured as real patient information, offer a potential solution to the challenges associated with accessing and using real patient data (Beaulieu-Jones et al., 2019; Bellovin et al., 2019; van Breugel and van der Schaar, 2023). Synthetic data can be shared, analyzed, and used freely, bypassing the regulatory and logistical obstacles associated with real data use (El Emam et al., 2020). Despite the potential of synthetic data, achieving high utility, fidelity, and privacy remains a significant challenge (Jordon et al., 2022). Current methods of synthetic data generation, including generative adversarial networks (GANs) (Goodfellow et al., 2020) and variational autoencoders (VAEs) (Kingma and Welling, 2022), have demonstrated utility in synthetic data generation (Goncalves et al., 2020; Jacobs et al., 2023; Rajotte et al., 2022). However, privacy concerns have been raised regarding generative models (Chen et al., 2020; Hayes et al., 2018) including data extractions, model inversions, and membership inference attacks (Rajotte et al., 2022). Furthermore, despite some workarounds, there are issues with mode collapse (Thanh-Tung and Tran, 2020) and the applicability of GANs toward generating categorical and binary data (Jacobs et al., 2023). The use of GANs and VAEs is also self-limiting to those with technical expertise (e.g., complex architecture, fine-tuning) as well as access to necessary computational resources and reference datasets for training.

In recent years, large language models (LLMs)—a computational model capable of language generation and other natural language processing tasks—offer new possibilities for generating text that is coherent and contextually relevant (Brown et al., 2020; Nazir and Wang, 2023). A prominent and publicly available LLM, OpenAI's ChatGPT, has shown utility in generating synthetic text-based data (Calvo-Lorenzo and Uriarte-Llano, 2024; Hämäläinen et al., 2023; Li et al., 2021). However, the potential of generating tabular synthetic clinical data with ChatGPT remains largely unexplored. Use of LLMs for synthetic data generation may offer an accessible alternative to GANs and VAEs, reducing the need for specialized knowledge and computational resources, which could broaden the reach of synthetic data use in research and ML model development.

This study aims to assess the feasibility of generating realistic tabular clinical data with OpenAI's GPT-4o (Hurst et al., 2024) using zero-shot prompting, and evaluate the fidelity of LLM-generated data by comparing its statistical properties to the Vital Signs DataBase (VitalDB) (Lee et al., 2022), a real open-source multi-parameter perioperative dataset.

2 Methods

2.1 Overview

We conducted a two-phase study to evaluate the feasibility of generating synthetic clinical datasets with GPT-4o using a single prompt and without pre-training. In both phases, GPT-4o was prompted to generate a synthetic dataset based on 13 clinical parameters derived from VitalDB. In Phase 1, GPT-4o was prompted with high-level qualitative descriptions of the 13 clinical parameters, to assess its ability to generate a complete and contextually relevant tabular dataset without guiding statistics. In Phase 2, GTP-4o was prompted to generate a synthetic dataset using descriptive statistics of the VitalDB dataset. Both Phase 1 and 2 datasets were statistically compared to VitalDB, with Phase 1 data serving as a baseline for comparisons.

2.2 Real dataset

The real clinical dataset used as a comparator in this study is the open-source VitalDB. The VitalDB dataset is a perioperative dataset consisting of multi-parameter data from surgery patients who underwent routine and emergency non-cardiac (general, thoracic, urological, and gynecological) operations at Seoul National University Hospital (Seoul, Korea) from August 2016 to June 2017 (Lee et al., 2022). The dataset included 6,388 de-identified cases encompassing a wide range of clinical parameters including demographic, preoperative, intraoperative, and postoperative parameters.

The VitalDB dataset was selected due to its open-source availability and data completeness for parameters spanning the entire perioperative period. The VitalDB dataset also included a variety of data formats (i.e., numerical, text), variable types (i.e., continuous, categorical, binary), and distributions (i.e., normal, skewed). These considerations ensured a comprehensive evaluation of GPT-4o's ability to generate and replicate statistical properties of a wide array of clinical data.

2.3 Parameter selection and data cleaning

The VitalDB dataset was reviewed for data completeness and parameters with missing data were excluded. Included parameters (*n* = 13) were chosen based on relevance to perioperative care and to represent a range of data formats and variable types. Remaining parameters with similar data formats, variable types, or clinical information were excluded for redundancy in the context of a feasibility study. Timepoint variables in the VitalDB dataset were recorded as the duration from an assigned case start time in seconds. Two time variables were included and converted to hours: operation duration (difference between operation end time and operation start time) and postoperative length of stay (difference between discharge time and operation end time). Selected parameters are presented in Table 1.

To further ensure quality and relevance of the data used for comparison, case files for patients younger than 18 (*n* = 57), older than 89 (*n* = 8), missing an American Society of Anesthesiologists (ASA) physical status classification (*n* = 130), and with negative discharge times (*n* = 27) were excluded from the dataset. In total, *n* = 6,166 cases were included.

2.4 Generation of synthetic datasets

Prior to the generation of Phase 1 and 2 synthetic datasets, GPT-4o was not pre-trained or provided any patient data from the VitalDB dataset. In Phase 1, GPT-4o was prompted with qualitative descriptions of 13 clinical parameters and asked to generate corresponding data for 6,166 patients. The prompt did not include descriptive statistics, definitions (e.g., ASA physical status classification), or formulas to calculate parameters (e.g., body mass index (BMI)). The prompt used to generate the Phase 1 synthetic dataset is presented in Box 1.

In Phase 2, GPT-4o was prompted with descriptive statistics of the VitalDB dataset. For continuous parameters (age, height, weight, operation duration, postoperative length of stay), descriptive statistics included mean, standard deviation, and range. Descriptive statistics were not provided for BMI, and GPT-4o was instructed to calculate this parameter using each case file's corresponding height and weight. Descriptive statistics for height were inputted to GPT-4o as centimeters, requiring the LLM to convert the height parameter to meters in order to calculate BMI—this transformation was not specifically instructed within the prompt. For categorical and binary parameters (ASA physical status classification, operation type, biological sex, preoperative hypertension, preoperative diabetes mellitus, intraoperative transfusion), corresponding proportions were provided. GPT-4o was instructed to assign ascending whole number values for each case ID. For time variables, natural log transformations were used to normalize skewed distributions and GPT-4o was provided with descriptive statistics of the log-transformed values. The prompt used to generate the Phase 2 synthetic dataset is presented in Box 2.

GPT-4o application programming interfaces (API) were not used in order to determine feasibility of data generation without further technical expertise and resources.

TABLE 1 Summary of selected parameters from the VitalDB dataset.

Category (*n*)	Parameters (units)
Demographic data (6)	Case ID, age (years), biological sex (M/F), height (cm), weight (kg), BMI (kg/m^2)
Preoperative morbidity (3)	ASA physical status classification (1–6), preoperative hypertension (yes/no), preoperative diabetes mellitus (yes/no)
Intraoperative data (3)	Operation type, operation duration (hours), intraoperative transfusion (yes/no)
Postoperative outcomes (1)	Postoperative length of stay (hours)

BMI, body mass index, ASA, American Society of Anesthesiologists.

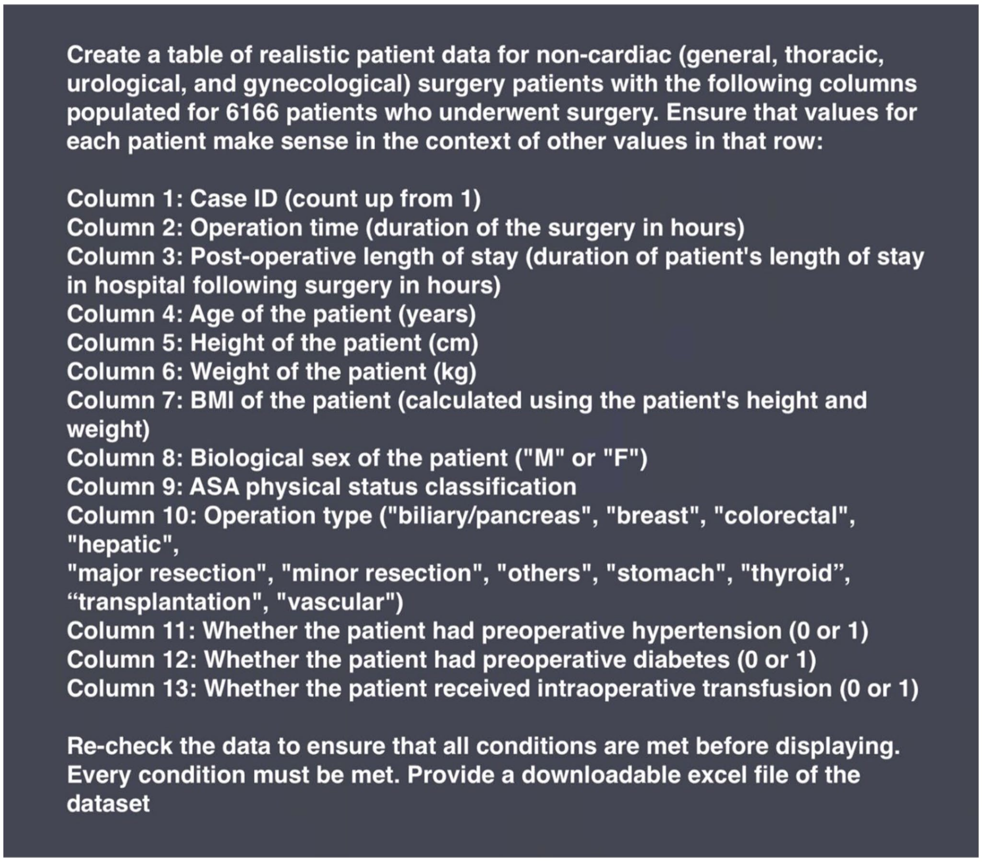
2.5 Dataset analysis

The Phase 1 dataset was assessed for general errors, plausibility of outputs, and cross-verification of related parameters. Assessment of general errors evaluated missing data, unexpected outputs, and formatting issues in the tabular data output. Plausibility of outputs involved evaluating time variables for positive values, ASA physical status classification for values between 1 and 6, and categorical and binary parameters for expected values (i.e., only including categories provided in the prompt, category proportions add to 100%). Cross-verification of related parameters involved confirming that all BMI values were appropriately calculated given the corresponding height and weight for each case file.

The Phase 1 and 2 datasets were compared to the VitalDB dataset for statistical similarity. Continuous variables were compared using two-sample t-tests and 95% CI overlap. Given the large sample size, we used parametric two-sample *t*-tests. The log-transformed values of operation duration and postoperative length of stay were used for statistical testing with the two-sample *t*-tests for Phase 2 data. For each continuous parameter, 95% CI overlap was calculated as the proportion of shared values compared to the entire range of values within both 95% CIs from LLM-generated and VitalDB datasets. The Python library Matplotlib was used to generate figures visualizing the overlap of 95% CI for continuous parameters and proportional alignment of categorical and binary parameters. Categorical and binary variables were compared using two-sample proportion tests. Statistical testing was performed using RStudio v.4.4.2 and statistical significance was set at 0.05. For two-sample t-tests and proportion tests, *p*-values above 0.05 indicated statistically insignificant differences in means and proportions, therefore representing an effective replication of descriptive statistical properties from the VitalDB reference dataset.

2.6 Ethical considerations

Datasets generated in this study solely represented fictitious patient data. Use of the VitalDB dataset was used in accordance with the requirements outlined by the study team. No data from the VitalDB dataset was inputted directly into GPT-4o for pre-training and no direct data is included in this paper. Furthermore, the synthetic



Create a table of realistic patient data for non-cardiac (general, thoracic, urological, and gynecological) surgery patients with the following columns populated for 6166 patients who underwent surgery. Ensure that values for each patient make sense in the context of other values in that row:

Column 1: Case ID (count up from 1)
Column 2: Operation time (duration of the surgery in hours)
Column 3: Post-operative length of stay (duration of patient's length of stay in hospital following surgery in hours)
Column 4: Age of the patient (years)
Column 5: Height of the patient (cm)
Column 6: Weight of the patient (kg)
Column 7: BMI of the patient (calculated using the patient's height and weight)
Column 8: Biological sex of the patient ("M" or "F")
Column 9: ASA physical status classification
Column 10: Operation type ("biliary/pancreas", "breast", "colorectal", "hepatic",
"major resection", "minor resection", "others", "stomach", "thyroid", “transplantation", "vascular")
Column 11: Whether the patient had preoperative hypertension (0 or 1)
Column 12: Whether the patient had preoperative diabetes (0 or 1)
Column 13: Whether the patient received intraoperative transfusion (0 or 1)

Re-check the data to ensure that all conditions are met before displaying. Every condition must be met. Provide a downloadable excel file of the dataset

BOX 1
Prompt input to generate the Phase 1 synthetic dataset with GPT-4o.

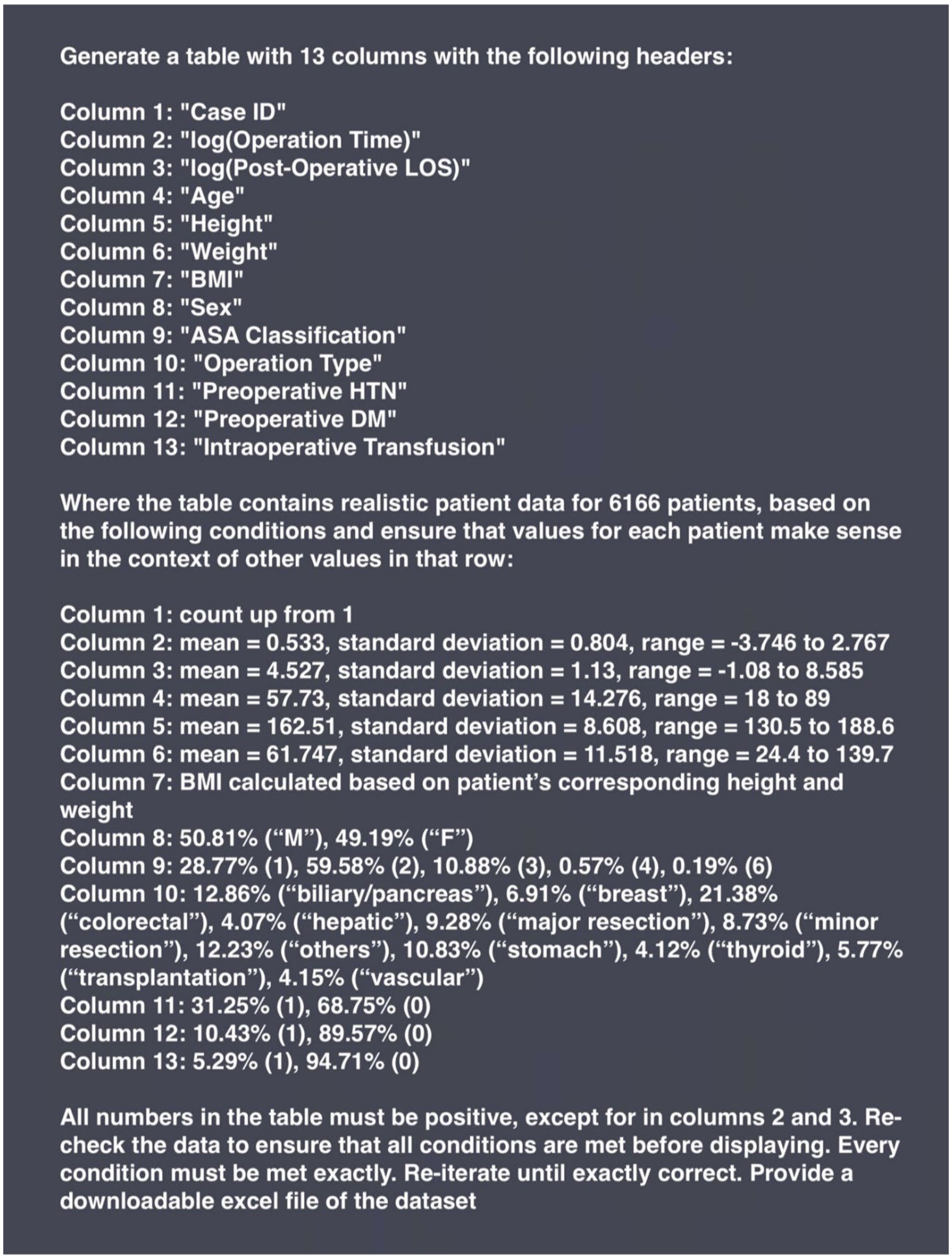
datasets generated by GPT-4o were evaluated solely for research purposes and not used in any form of clinical decision-making.

3 Results

In Phase 1, GPT-4o generated a complete and structured tabular dataset comprising 6,166 case files. All 13 expected columns were present, complete, and appropriately labeled; no missing data, unexpected outputs, or formatting issues were present within the generated dataset. Furthermore, all generated time variables were positive, ASA physical status classifications were within the appropriate range (1–6), and categorical and binary parameters only included expected values and proportions all added to 100%. Each case also included a correctly calculated BMI corresponding to the appropriate height and weight.

Review of calculated means and ranges for continuous variables included operation duration (6.46 h; 1.00–12.00), postoperative length of stay (154.84 h; 12.00–299.90), age (53.52 years; 18.00–89.00), height (174.51 cm; 150.00–199.00), weight (97.43 kg; 45.00–149.00), and BMI (32.62 kg/m^2; 11.40–66.20). All continuous parameters showed plausible means and ranges for a perioperative dataset. However, proportions among categorical and binary variables did not differ based on context. Proportions were evenly spread across categories for parameters which are likely to demonstrate uniform distributions (e.g., sex, operation type) as well as parameters that may have skewed distributions (e.g., ASA physical status classification, preoperative comorbidities, intraoperative transfusions). A percent stacked bar plot displaying proportional alignment of categorical and binary parameters can be seen in Figure 1. Overall, generated data in Phase 1 was realistic, displayed appropriate ranges, included correct calculations without the provision of descriptive statistics, formulas, or unit conversions (e.g., BMI), and maintained definitional boundaries of parameters without explicit instructions (e.g., ASA physical status classification).

The Phase 1 and 2 datasets were statistically compared to the VitalDB dataset. The results of the statistical testing (Tables 2, 3) revealed that 12/13 (92.31%) parameters from the Phase 2 dataset did not show statistically significant differences from VitalDB, including 6/6 (100.00%) of the categorical and binary parameters and 6/7 (85.71%) of the continuous parameters. The only continuous parameter which demonstrated statistically significant differences in Phase 2 was the BMI parameter, which was calculated based on each case's height and weight rather than generated based on descriptive statistics in the prompt. For the Phase 1 dataset, 2/13 (15.28%) parameters did not show statistically significant differences from VitalDB, one of which was the Case ID parameter. Overlap of 95% CI was observed in 6/7 (85.71%) of the Phase 2 continuous parameters. The measured 95% CI overlaps were as follows: case ID (100.0%), weight (85.93%), height (61.31%), age (43.12%), postoperative length of stay (34.84%), operation duration (15.17%), and BMI (0.0%). The Phase 1 dataset only showed 95% CI overlap in the case ID parameter (100.0%). Visualization of the 95% CI overlaps of each continuous parameter is displayed in Figure 2. Overall, 12/13 (92.31%) of the Phase 2 parameters met the predefined threshold for statistical



Generate a table with 13 columns with the following headers:

Column 1: "Case ID"
Column 2: "log(Operation Time)"
Column 3: "log(Post-Operative LOS)"
Column 4: "Age"
Column 5: "Height"
Column 6: "Weight"
Column 7: "BMI"
Column 8: "Sex"
Column 9: "ASA Classification"
Column 10: "Operation Type"
Column 11: "Preoperative HTN"
Column 12: "Preoperative DM"
Column 13: "Intraoperative Transfusion"

Where the table contains realistic patient data for 6166 patients, based on the following conditions and ensure that values for each patient make sense in the context of other values in that row:

Column 1: count up from 1
Column 2: mean = 0.533, standard deviation = 0.804, range = -3.746 to 2.767
Column 3: mean = 4.527, standard deviation = 1.13, range = -1.08 to 8.585
Column 4: mean = 57.73, standard deviation = 14.276, range = 18 to 89
Column 5: mean = 162.51, standard deviation = 8.608, range = 130.5 to 188.6
Column 6: mean = 61.747, standard deviation = 11.518, range = 24.4 to 139.7
Column 7: BMI calculated based on patient's corresponding height and weight
Column 8: 50.81% ("M"), 49.19% ("F")
Column 9: 28.77% (1), 59.58% (2), 10.88% (3), 0.57% (4), 0.19% (6)
Column 10: 12.86% ("biliary/pancreas"), 6.91% ("breast"), 21.38% ("colorectal"), 4.07% ("hepatic"), 9.28% ("major resection"), 8.73% ("minor resection"), 12.23% ("others"), 10.83% ("stomach"), 4.12% ("thyroid"), 5.77% ("transplantation"), 4.15% ("vascular")
Column 11: 31.25% (1), 68.75% (0)
Column 12: 10.43% (1), 89.57% (0)
Column 13: 5.29% (1), 94.71% (0)

All numbers in the table must be positive, except for in columns 2 and 3. Re-check the data to ensure that all conditions are met before displaying. Every condition must be met exactly. Re-iterate until exactly correct. Provide a downloadable excel file of the dataset

BOX 2
Prompt input to generate the Phase 2 synthetic dataset with GPT-4o.

similarity, demonstrating the parameter effectively replicated statistical properties of corresponding data from the VitalDB dataset.

The Excel files containing the Phase 1 and 2 synthetic datasets generated with GPT-4o are available in the Supplementary data.

4 Discussion

4.1 Overview

The present study evaluated the feasibility of generating tabular synthetic clinical data with GPT-4o using zero-shot prompting, and assessed the fidelity of the generated data by comparing it to a real clinical dataset, VitalDB. By examining two phases of data generation—one using qualitative prompts (Phase 1) and another incorporating descriptive statistics from VitalDB (Phase 2)—we explored GPT-4o's capacity to generate plausible data and replicate statistical properties, variable distributions, and contextual characteristics typical of clinical data. Generated data included various formats (numerical, text), variable types (continuous, categorical, binary), and distributions (normal, skewed), covering demographic, preoperative, intraoperative, and postoperative data. The results indicate: (1) GPT-4o can produce realistic synthetic data without descriptive statistics or reference data, and (2) GPT-4o can generate

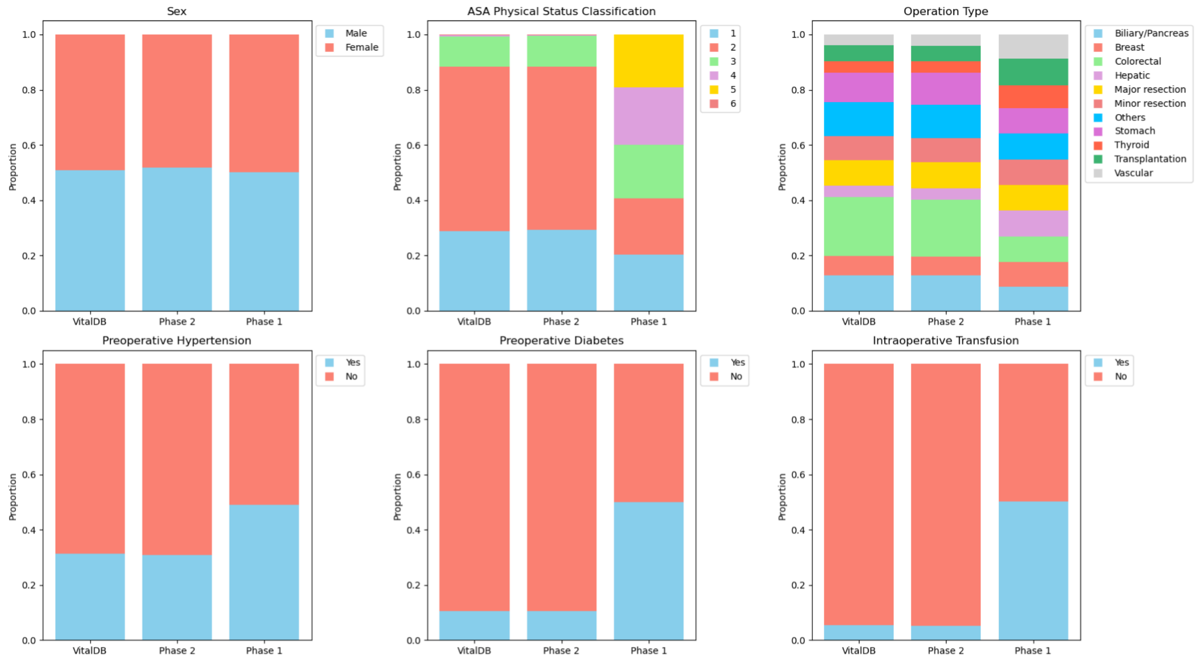


FIGURE 1
Percent stacked bar plot displaying proportional alignment of categorical and binary parameters between VitalDB, Phase 1, and Phase 2 datasets.

datasets that align closely with real clinical data, when provided with statistical guidance.

4.2 Principal findings and implications

The use of LLMs to generate structured tabular data using zero-shot prompting is a novel concept. Consistent with other modalities of synthetic data generation, LLM-generated data has the potential to address many of the challenges of accessing and using real clinical data (Beaulieu-Jones et al., 2019; El Emam et al., 2020; van Breugel and van der Schaar, 2023). While GPT-4o has yet to demonstrate equivalent fidelity and utility to GANs and VAEs, LLMs may offer solutions to some of their shortcomings. The generation of data using GANs and VAEs requires technical expertise and computational resources. However, data generation using LLMs is accessible to anyone with an internet connection, and can produce clean and ready-to-use datasets, outputted in a downloadable Excel file, through plain-language prompting. This holds substantial implications for democratizing data access in research, educational contexts, and ML model development (Rajotte et al., 2022).

Given that no reference data was required in Phase 1 or inputted for pre-training in Phase 2, LLMs may overcome privacy concerns associated with current approaches to synthetic data generation (Rajotte et al., 2022). The Phase 1 data, which generated realistic clinical data in the absence of guiding statistics, definitions, and formulas, further emphasizes the contextual relevance of outputs from GPT-4o (Brown et al., 2020; Nazir and Wang, 2023). This is particularly useful in educational contexts, whereby learning opportunities for students and trainees can be enhanced by practicing data analysis using synthetic data generated with desired statistical properties and without requiring a reference dataset. LLM-generated datasets can also include synthetic personal health information which would otherwise be removed or de-identified. Since synthetic data can be used without restriction, data can also be re-inputted into LLMs for analysis, which further broadens prospects and scope of future research.

Clean and structured synthetic datasets, with the statistical similarity Phase 2 data demonstrated to VitalDB, has vast implications for data-driven medicine and ML model development. Perioperative data, in particular, is inherently heterogeneous, encompassing a wide variety of sources, formats, and qualities (Maier-Hein et al., 2017). By generating synthetic data which can replicate real-world data distributions, researchers can bypass additional challenges associated with the use of raw clinical data. In this way, synthetic datasets can accelerate the development of predictive tools and surgical decision support systems, ultimately contributing to patient care and surgical outcomes.

4.3 Limitations

While this feasibility study demonstrated remarkable preservation of within-column statistical properties and simple relationships between variables, there are some limitations and challenges to consider. First, this study focused solely on GPT-4o, and it remains uncertain whether similar results would be achieved using other LLMs. Similarly, direct comparisons in performance between LLMs, GANs, and VAEs are necessary to assess their relative utility, fidelity, and privacy preservation, which may uncover additional limitations and strengths. At present, it is unclear whether bivariate and multivariate relationships were retained by the LLM-based approach, as this was not directly assessed. Demonstrating the preservation of correlations and other nuanced interdependencies, present in clinical data, is necessary before meaningful comparisons can be made between the performance of LLMs and other data generation methods.

This study also revealed the importance of prompt design in generating accurate and relevant synthetic data outputs. In Phase 1, where prompts lacked descriptive statistics, generated data deviated significantly from the reference dataset. While this underscores the notable results in Phase 2, it also suggests that without explicit guidance, LLMs may produce plausible but statistically unaligned data. Therefore, the quality of outputted data is reliant on effective prompting, and improper prompt design can introduce bias or errors into generated data. It is also unknown whether an iterative approach to prompting may result in greater fidelity.

Continued improvements in LLMs have been previously associated with greater accuracy in a variety of generative and clinically associated tasks (Meyer et al., 2024; Rosoł et al., 2023), and further iterations of GPT-4o may improve upon these results and current limitations.

TABLE 2 Means with associated 95% CI intervals, *p*-values from associated two-sample *t*-tests, and 95% CI overlap values comparing case ID, operation duration, postoperative length of stay, age, height, weight, and BMI between VitalDB and Phase 1 and 2 datasets.

Parameter	VitalDB mean (95% CI)	Phase 1 mean (95% CI)	VitalDB vs Phase 1 *p*-value	VitalDB vs Phase 1 95% CI overlap	Phase 2 mean (95% CI)	VitalDB vs Phase 2 *p*-value	VitalDB vs Phase 2 95% CI overlap
Case ID	3083.5 (3039.06–3127.94)	3083.5 (3039.06–3127.94)	1.000	100.0%	3083.5 (3039.06–3127.94)	1.000	100.0%
Operation duration (hours)	2.27 (2.22–2.31)	6.46 (6.38–6.54)	<0.001	0.0%	2.33 (2.28–2.39)	0.116*	15.17%
Postoperative length of stay (hours)	167.19 (160.48–173.90)	154.84 (152.78–156.90)	<0.001	0.0%	173.6 (166.99–180.31)	0.845*	34.84%
Age (years)	57.73 (57.37–58.09)	53.52 (53.00–54.04)	<0.001	0.0%	58.01 (57.66–58.37)	0.819	43.12%
Height (cm)	162.51 (162.30–162.73)	174.51 (174.16–174.87)	<0.001	0.0%	162.41 (162.20–162.62)	0.200	61.31%
Weight (kg)	61.75 (61.46–62.04)	97.43 (96.67–98.19)	<0.001	0.0%	61.70 (61.42–61.99)	0.327	85.93%
BMI (kg/m^2)	23.32 (23.23–23.41)	32.62 (32.33–32.91)	<0.001	0.0%	23.60 (23.47–23.73)	0.023	0.0%

**p*-values were calculated using log-transformed values.

4.4 Future directions

This study's use of an open-source dataset (VitalDB) as a comparator was intentional to facilitate reproducibility and encourage follow-up research. Future research should continue to investigate the capabilities of LLMs in generating tabular datasets, with particular focus on capturing complex interdependencies between parameters and further assessing reproducibility of results. This should involve using existing and robust frameworks to assess the fidelity and privacy preservation of LLM-generated synthetic datasets (El Emam, 2020; El Emam et al., 2022; Platzer and Reutterer, 2021; Vallevik et al., 2024). Direct comparisons should be made between the performance of GPT-4o and other prominent LLMs. Following further refinement of this zero-shot approach, direct comparisons in utility and privacy should also be conducted between LLMs, GANs, and VAEs, using a systematic benchmarking approach (Yan et al., 2022).

Future work should assess the potential of LLMs in data enhancement, including data amplification and augmentation (El Emam, 2023)—particularly in domains with missing data or limited data availability (e.g., rare diseases, underrepresented patient populations) (Rajotte et al., 2022). By supplementing existing datasets with synthetic data that preserves statistical properties, LLMs could mitigate data scarcity and enable more robust research in data-constrained fields. Applications of LLM-generated synthetic data toward ML model development and validation, predictive tools, and surgical decision support systems should also be explored.

Applications in educational contexts may be evaluated. Surveys, qualitative interviewing, or randomized trials involving students who have used LLM-generated datasets may reveal whether supplementing educational programs with synthetic data can enhance learning for students training toward careers in disciplines which analyze clinical data (e.g., statisticians, data scientists, epidemiologists). It is also recommended to assess whether LLMs can be used to effectively summarize and analyze outputted synthetic data.

5 Conclusion

This study demonstrates that zero-shot prompting with GPT-4o can generate realistic tabular synthetic datasets that replicate key

TABLE 3 Number and proportion of patients by sex, ASA physical status classification, operation type, preoperative hypertension status, preoperative diabetes mellitus status, and intraoperative transfusion status for VitalDB, Phase 1, and Phase 2 datasets with *p*-values from associated two sample proportion tests comparing distributions between VitalDB and synthetic datasets for each parameter.

Parameter	VitalDB *n*, (%)	Phase 1 *n*, (%)	VitalDB vs Phase 1 *p*-value	Phase 2 *n*, (%)	VitalDB vs Phase 2 *p*-value
Sex					
Male	3133 (50.81%)	3087 (50.06%)	0.407	3195 (51.82%)	0.263
Female	3033 (49.19%)	3079 (49.94%)		2971 (48.18%)	
ASA physical status classification					
1	1774 (28.77%)	1249 (20.26%)	<0.001	1810 (29.35%)	0.478
2	3674 (59.58%)	1254 (20.34%)	<0.001	3637 (58.98%)	0.497
3	671 (10.88%)	1198 (19.43%)	<0.001	683 (11.08%)	0.726
4	35 (0.57%)	1291 (20.94%)	<0.001	26 (0.42%)	0.246
5	0 (0.0%)	1174 (19.04%)	<0.001	0 (0.0%)	-
6	12 (0.19%)	0 (0.0%)	<0.001	10 (0.16%)	0.667
Operation type					
Biliary/Pancreas	793 (12.86%)	534 (8.66%)	<0.001	782 (12.68%)	0.764
Breast	426 (6.91%)	558 (9.05%)	<0.001	419 (6.80%)	0.803
Colorectal	1318 (21.38%)	556 (9.02%)	<0.001	1274 (20.66%)	0.332
Hepatic	251 (4.07%)	587 (9.52%)	<0.001	260 (4.22%)	0.682
Major resection	572 (9.28%)	569 (9.23%)	0.928	577 (9.36%)	0.881
Minor resection	538 (8.73%)	563 (9.13%)	0.430	536 (8.69%)	0.952
Others	754 (12.23%)	585 (9.49%)	<0.001	750 (12.16%)	0.912
Stomach	668 (10.83%)	571 (9.26%)	0.004	710 (11.51%)	0.230
Thyroid	254 (4.12%)	509 (8.25%)	<0.001	265 (4.30%)	0.624
Transplantation	356 (5.77%)	596 (9.67%)	<0.001	332 (5.38%)	0.347
Vascular	236 (4.15%)	538 (8.73%)	<0.001	261 (4.23%)	0.254
Preoperative hypertension					
Yes	1927 (31.25%)	3028 (49.11%)	<0.001	1899 (30.80%)	0.582
No	4239 (68.75%)	3138 (50.89%)		4267 (69.20%)	
Preoperative diabetes mellitus					
Yes	643 (10.43%)	3075 (49.87%)	<0.001	645 (10.46%)	0.952
No	5523 (89.57%)	3091 (50.13%)		5521 (89.54%)	
Intraoperative transfusion					
Yes	326 (5.29%)	3098 (50.24%)	<0.001	314 (5.09%)	0.624
No	5840 (94.71%)	3068 (49.76%)		5852 (94.91%)	

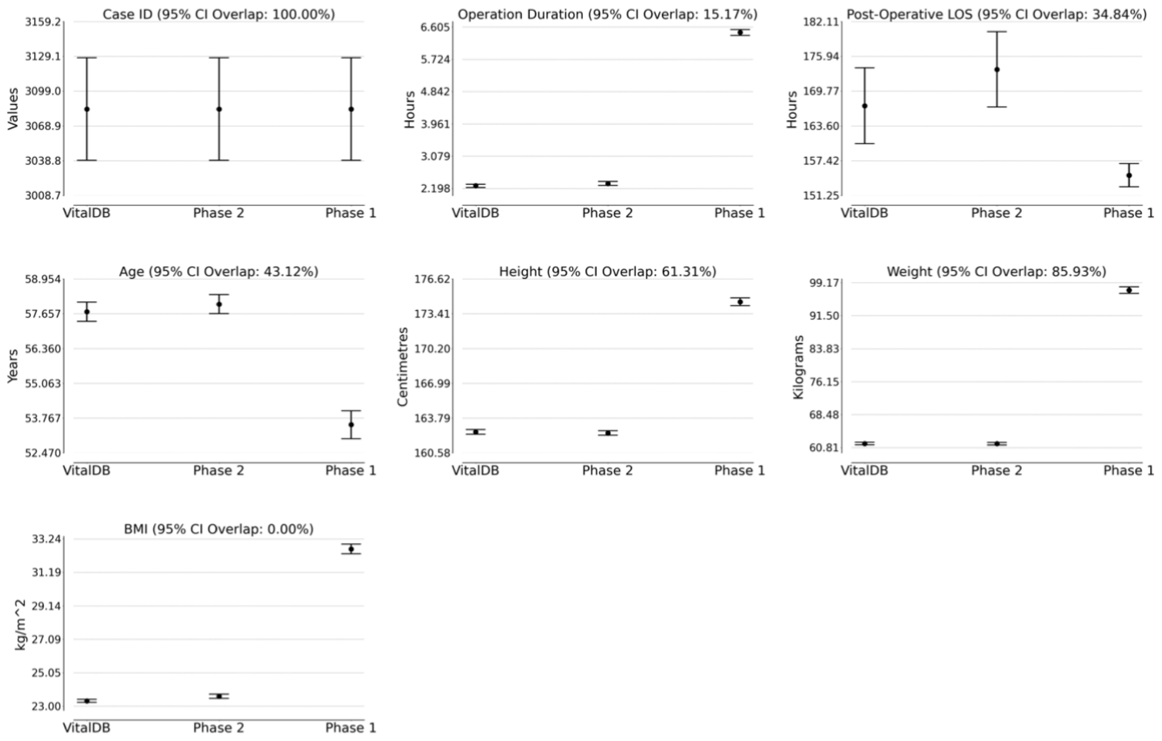


FIGURE 2
95% confidence intervals for VitalDB, Phase 1, and Phase 2 datasets with labeled percentage of 95% CI overlap between VitalDB and Phase 2 data. LOS, Length of Stay; BMI, body mass index.

statistical properties of real perioperative data. By eliminating the need for technical expertise, extensive computational resources, and pre-training in synthetic data generation, LLMs can offer an accessible modality to address critical barriers associated with clinical data access. Collectively, these findings highlight the broad implications of LLM-generated synthetic data in democratizing data access and enhancing educational opportunities. Future research should focus on enhancing fidelity and investigating the application of LLMs in data amplification and augmentation, replication of multivariate relationships, and ML model development.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

AB: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. JQ: Formal analysis, Investigation, Methodology, Writing – review & editing. EG: Methodology, Visualization, Writing – review & editing. ES: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. This project made use of GPT-4o, a language

model developed by OpenAI, to generate synthetic data. We acknowledge the contribution of GPT-4o in the present work.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1533508/full#supplementary-material

References

Beaulieu-Jones, B. K., Wu, Z. S., Williams, C., Lee, R., Bhavnani, S. P., Byrd, J. B., et al. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Cardiovasc. Qual. Outc.* 12:e005122. doi: 10.1161/CIRCOUTCOMES.118.005122

Bellovin, S. M., Dutta, P. K., and Reitinger, N. (2019). Privacy and synthetic datasets. *Stan. Tech. L Rev.* 22:1.

Bentzen, H. B., Castro, R., Fears, R., Griffin, G., ter Meulen, V., and Ursin, G. (2021). Remove obstacles to sharing health data with researchers outside of the European Union. *Nat. Med.* 27, 1329–1333. doi: 10.1038/s41591-021-01460-0

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Arxiv*. doi: 10.48550/arXiv.2005.14165

Calvo-Lorenzo, I., and Uriarte-Llano, I. (2024). Massive generation of synthetic medical records with ChatGPT: an example in hip fractures. *Med. Clín.* 162, 549–554. doi: 10.1016/j.medcle.2023.11.033

Chen, D., Yu, N., Zhang, Y., and Fritz, M. (2020). GAN-leaks: A taxonomy of membership inference attacks against generative models, in Proceedings of the 2020 ACM SIGSAC conference on computer and communications security, (New York, NY, USA: Association for Computing Machinery), 343–362.

El Emam, K. (2020). Seven ways to evaluate the utility of synthetic data. *IEEE Sec. Privacy* 18, 56–59. doi: 10.1109/MSEC.2020.2992821

El Emam, K. (2023). Status of synthetic data generation for structured health data. *JCO* e2300071:7. doi: 10.1200/CCI.23.00071

El Emam, K. E., Mosquera, L., and Bass, J. (2020). Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *J. Med. Internet Res.* 22:e23139. doi: 10.2196/23139

El Emam, K. E., Mosquera, L., Fang, X., and El-Hussuna, A. (2022). Utility metrics for evaluating synthetic health data generation methods: validation study. *JMIR Med. Inform.* 10:e35734. doi: 10.2196/35734

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., and Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* 20:108. doi: 10.1186/s12874-020-00977-1

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

Hämäläinen, P., Tavast, M., and Kunnari, A. (2023). Evaluating large language models in generating synthetic HCI research data: A case study., in Proceedings of the 2023 CHI conference on human factors in computing systems, (New York, NY, USA: Association for Computing Machinery), 1–19.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. (2018). LOGAN: Membership inference attacks against generative models. *Arxiv*. doi: 10.48550/arXiv.1705.07663

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., et al. (2024). GPT-4o system card. *Arxiv*. doi: 10.48550/arXiv.2410.21276

Jacobs, F., D'Amico, S., Benvenuti, C., Gaudio, M., Saltalamacchia, G., Miggiano, C., et al. (2023). Opportunities and challenges of synthetic data generation in oncology. *JCO* 7:45. doi: 10.1200/CCI.23.00045

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., et al. (2022). Synthetic data - what, why and how? *Arxiv*. doi: 10.48550/arXiv.2205.03257

Kingma, D. P., and Welling, M. (2022). Auto-encoding Variational Bayes. *Arxiv*. doi: 10.48550/arXiv.1312.6114

Lee, H. C., Park, Y., Yoon, S. B., Yang, S. M., Park, D., and Jung, C.-W. (2022). VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Sci. Data* 9:279. doi: 10.1038/s41597-022-01411-5

Li, J., Zhou, Y., Jiang, X., Natarajan, K., Pakhomov, S. V., Liu, H., et al. (2021). Are synthetic clinical notes useful for real natural language processing tasks: a case study on clinical entity recognition. *J. Am. Med. Inform. Assoc.* 28, 2193–2201. doi: 10.1093/jamia/ocab112

Maier-Hein, L., Vedula, S. S., Speidel, S., Navab, N., Kikinis, R., Park, A., et al. (2017). Surgical data science for next-generation interventions. *Nat. Biomed. Eng.* 1, 691–696. doi: 10.1038/s41551-017-0132-7

Meyer, A., Riese, J., and Streichert, T. (2024). Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German medical licensing examination: observational study. *JMIR* 10:e50965. doi: 10.2196/50965

Nazir, A., and Wang, Z. (2023). A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. *Meta Radiol.* 1:100022. doi: 10.1016/j.metrad.2023.100022

Ness, R. B. (2007). Influence of the HIPAA privacy rule on Health Research. *JAMA* 298, 2164–2170. doi: 10.1001/jama.298.18.2164

Newgard, C. D., and Lewis, R. J. (2015). Missing data: how to best account for what is not known. *JAMA* 314, 940–941. doi: 10.1001/jama.2015.10516

Pavlenko, E., Strech, D., and Langhof, H. (2020). Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. *BMC Med. Inform. Decis. Mak.* 20:157. doi: 10.1186/s12911-020-01177-z

Platzer, M., and Reutterer, T. (2021). Holdout-based empirical assessment of mixed-type synthetic data. *Front. Big Data* 4:679939. doi: 10.3389/fdata.2021.679939

Rajotte, J.-F., Bergen, R., Buckeridge, D. L., El Emam, K., Ng, R., and Strome, E. (2022). Synthetic data as an enabler for machine learning applications in medicine. *iScience* 25:105331. doi: 10.1016/j.isci.2022.105331

Rosoł, M., Gąsior, J. S., Łaba, J., Korzeniewski, K., and Młyńczak, M. (2023). Evaluation of the performance of GPT-3.5 and GPT-4 on the polish medical final examination. *Sci. Rep.* 13:20512. doi: 10.1038/s41598-023-46995-z

Thanh-Tung, H., and Tran, T. (2020). Catastrophic forgetting and mode collapse in GANs., In 2020 International joint conference on neural networks (IJCNN).

Tudur, C., Nevitt, S., Appelbe, D., Appleton, R., Dixon, P., Harrison, J., et al. (2017). Resource implications of preparing individual participant data from a clinical trial to share with external researchers. *Trials* 18:319. doi: 10.1186/s13063-017-2067-4

Vallevik, V. B., Babic, A., Marshall, S. E., Elvatun, S., Brøgger, H. M. B., Alagaratnam, S., et al. (2024). Can I trust my fake data – a comprehensive quality assessment framework for synthetic tabular data in healthcare. *Int. J. Med. Inform.* 185:105413. doi: 10.1016/j.ijmedinf.2024.105413

van Breugel, B., and van der Schaar, M. (2023). Beyond privacy: Navigating the opportunities and challenges of synthetic data. *Arxiv*. 10.48550/arXiv.2304.03722

Wartenberg, D., and Thompson, W. D. (2010). Privacy versus public health: the impact of current confidentiality rules. *Am. J. Public Health* 100, 407–412. doi: 10.2105/AJPH.2009.166249

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., et al. (2020). Preparing medical imaging data for machine learning. *Radiology* 295, 4–15. doi: 10.1148/radiol.2020192224

Yan, C., Yan, Y., Wan, Z., Zhang, Z., Omberg, L., Guinney, J., et al. (2022). A multifaceted benchmarking of synthetic electronic health record generation models. *Nat. Commun.* 13:7609. doi: 10.1038/s41467-022-35295-1