arXiv:2506.16594v1 [cs.CL] 19 Jun 2025

A Scoping Review of Synthetic Data Generation for Biomedical Research and Applications

Hanshu Rao[1,+], Weisi Liu[1,+], Haohan Wang[2], I-Chan Huang[3], Zhe He[4], and Xiaolei Huang[1,*]

[1]University of Memphis, Dept. of Computer Science, Memphis, 38152, United States
[2]University of Illinois Urbana-Champaign, School of Information Sciences, Champaign, 61820, United States
[3]St Jude Children's Research Hospital, Dept. of Epidemiology and Cancer Control, Memphis, 38105, United States
[4]Florida State University, School of Information, Tallahassee, 32306, United States
[*]xiaolei.huang@memphis.edu
[+]these authors contributed equally to this work

ABSTRACT

Synthetic data generation—mitigating data scarcity, privacy concerns, and data quality challenges in biomedical fields—has been facilitated by rapid advances of large language models (LLMs). This scoping review follows PRISMA-ScR guidelines and synthesizes 59 studies, published between 2020 and 2025 and collected from PubMed, ACM, Web of Science, and Google Scholar. The review systematically examines biomedical research and application trends in synthetic data generation, emphasizing clinical applications, methodologies, and evaluations. Our analysis identifies data modalities of unstructured texts (78.0%), tabular data (13.6%), and multimodal sources (8.4%); generation methods of prompting (72.9%), fine-tuning (22.0%) LLMs and specialized model (5.1%); and heterogeneous evaluations of intrinsic metrics (27.1%), human-in-the-loop assessments (55.9%), and LLM-based evaluations (13.6%). The analysis addresses current limitations in what, where, and how health professionals can leverage synthetic data generation for biomedical domains. Our review also highlights challenges in adaption across clinical domains, resource and model accessibility, and evaluation standardizations.

Introduction

Data is the key to train reliable AI models for broad biomedical and clinical applications, such as medical diagnosis[1,2], therapeutic treatments[3], and drug discovery[4]. However, obtaining massive, privacy-preserving, and high-quality data faces critical challenges, such as data availability, noisy and missing data, and legal regulations. Increasing biomedical studies are turning attentions to synthetic data generation, a process of creating artificial datasets that accurately replicate the statistical and structural properties of real-world data. Nonetheless, creating high-quality synthetic data remains challenging due to the inherent heterogeneity, complexity, and variability characteristic of biomedical data. Large language models (LLMs), as a generative AI method, offer a promising solution by enhancing the accuracy and diversity of synthetic datasets through advanced learning algorithms.

Synthetic data generation in the biomedical and clinical fields has achieved significant advances by LLMs (e.g., GPT-4[5] and Llama 3[6]) in the recent years, as shown in Figure 2. For example, LLMs have been applied successfully to obtain clinical narratives and simulate real patient records of mental health for diagnosis and behavior analysis[7], such as predicting suicide[8] and depression patterns[9]. The strong generation capability not only relieve data privacy concerns[10] to facilitate model developments but also may promote data quality[11] to support clinical decision-making. For example, synthetic text corpora have augmented classification models for cardiovascular and Alzheimer's disease diagnosis[12]; A study develops a multi-agent dialogue generation framework (NoteChat) that creates synthetic patient-physician conversations to improve clinical documentation[13]. Nevertheless, questions and uncertainties remain regarding the best practices, validation approaches, and specific application areas for those LLM-generated synthetic data.

The goal of this scoping review is to comprehensively summarize and assess recent research publications on the biomedical and clinical applications leveraging LLMs for synthetic data generation. While multiple reviewing studies[14–16] have covered the LLM-generated synthetic data for general domains, studies on biomedical and clinical domains are still not available. For example, a close study[16] examines recent developments of generation algorithms and models on news and social media domains instead of biomedical fields. Specifically, through this study, we will identify and present the current state-of-the-art models, technical methods, evaluations or assessments of synthetic data quality, and gaps and opportunities for future research. We seek to answer a concrete yet unsolved question: *what biomedical and clinical applications can be effectively addressed using LLMs-generated synthetic data, and by how?*

Results

Our search and initial review resulted in 132 articles and finalized with 59 articles for this scoping review. XH, HW, and HR reviewed full texts of the articles and removed 73 articles due to ineligible article types (n=2), no large language model (n=7), unrelated to biomedical data generation (n=18), and not peer-reviewed and published in a journal or conference proceedings (n=46). Fifty-nine articles met our criteria and were included in our subsequent analyses. In terms of publication venues and formats, the included literature consists of 43 journal articles (72.9%) and 16 peer-reviewed conference papers (27.1%), leaving 59 studies to be included in this scoping review. Figure 1 shows that the study number of biomedical synthetic data generation climb steadily over years. Particularly, the substantial growth after 2022 reflects a particularly evolving trend of synthetic biomedical data generation via LLM, an emerging generation tool in recent years.

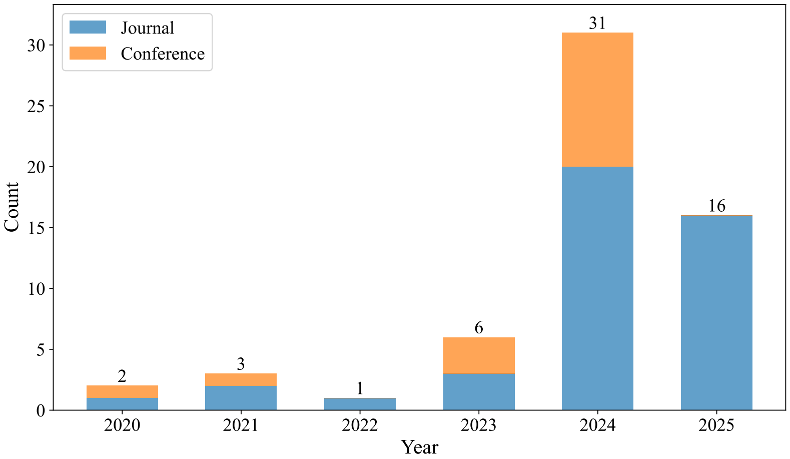

Figure 1. Publications between Jan 01, 2020 and April 05, 2025. Orange and blue colors refer to peer-reviewed conference and journal publications, respectively. We can observe a surging increase on the biomedical synthetic data related studies.

We present an overview of recent progress in synthetic biomedical data generation and application from three perspectives, aiming to provide readers with an insightful understanding of current practices and inform future research directions:

1. Synthetic data types and medical applications: Research has explored broad synthetic data types of clinical text, tabular data, and multimodal information (e.g., images, audio, and sensor signals). These data types cover a range of applications, including EHR analysis, medical imaging, telemedicine, mental health, and clinical trial matching, addressing challenges related to data scarcity and privacy restrictions.

2. Synthetic data generation methodology: Evolving methodologies encompass prompting-based generation, knowledge infusion, and multi-agent and multimodal approaches. These advancements enhance the semantic consistency, efficiency, and scalability of synthetic biomedical data generation.

3. Quality assessments and evaluation metrics: Ensuring data quality and utility remains a priority. Current studies employ evaluation strategies targeting fidelity, utility, and privacy protection. Such strategies integrate statistical metrics, human-in-the-loop review, automated model assessments, and privacy testing to establish a robust framework for verifying the reliability and compliance of synthetic data.

Synthetic Data Types and Medical Applications

Increasing model developments for biomedical applications lead to the surging needs of synthetic data[18,20]. In this section, we provide an overview of recent studies using synthetic data and address three critical questions: 1) What types of synthetic data are being generated? 2) What modalities and clinical tasks do these datasets target? 3) Are those data resources accessible for reusable research, such as benchmarking and model training? To answer those questions, we examine several aspects of the collected studies and summarize them in Table 1, including data modality, data language, data size, published year, medical application, and data or model accessibility. We aim to provide insights into current and future trends of data types and biomedical applications.

Biomedical synthetic data has shown diverse medical topics, languages, and modalities. Most of the studies (55.9%, $n = 33$) are for general topics, while others include specific medical areas, such as suicide[8,26], colorectal cancer[66], and radiology[57].

Table 1. A summary of data types, generation Methods, and medical applications for our collected studies. We use language codes for each study's data language, including English (EN), Dutch (NL), French (FR), Chinese (ZH), and Arabic (AR). The K (1,000) shortens spaces of the size column. We use acronyms for medical applications, including QA (question answering), IE (information extraction), and SDoH (social determinants of health). We use the "-" to indicate not specified information.

Study	Modality	Lang.	Architecture	Approach	Size (K)	Year	Accessible	Purpose	Medical Application
17	Text	EN	Transformer	Fine-tuning	500	2020	No	Training	Biomedical QA
18	Text	EN	Transformer	Specialized model	11	2020	Yes	Training	Phenotype inference
19	Text	EN	GPT	Fine-tuning	<1K	2021	Yes	Training	Clinical IE
20	Text	NL	GPT, RNN	Fine-tuning	355	2021	Yes	Training	De-identification
21	Text	EN	GPT	Prompting	45	2021	No	Training	Dialogue summarization
22	Text	EN	GPT	Fine-tuning	48	2022	No	Training	Readmission prediction
23	Text	EN	GPT	Fine-tuning	200	2023	Yes	Training	General
24	Text	FR	GPT	Fine-tuning	5	2023	Yes	Training	Clinical IE
25	Text	EN	GPT	Fine-tuning	5	2023	No	Training	Syndromic detection
26	Text	EN	GPT	Fine-tuning	<1K	2023	Yes	Training	Assisted suicide
27	Tabular	EN	Transformer	Specialized model	976	2023	Yes	Training	General
12	Text	EN	GPT	Prompting	32	2023	No	Training	Alzheimer's disease
28	Text	EN	GPT	Prompting	<1K	2024	Yes	Training	Discharge summary
29	Text	EN	GPT	Prompting	2	2024	Yes	Training	PTSD diagnosis
30	Image	ZH, EN	GPT	Prompting	4,000	2024	No	Training	Medical imaging
31	Text	EN	GPT	Prompting	9	2024	Yes	Training	Acute renal failure
10	Text	EN	GPT	Prompting	4	2024	No	Privacy	Phenotype inference
32	Text	EN	GPT	Prompting	2	2024	No	Training	Cohort selection
33	Text	EN	GPT	Prompting	<1K	2024	No	Supplement	Cohort selection
34	Text	DE	GPT	Prompting	4	2024	No	Training	Emergency medicine
35	Text	EN	GPT	Prompting	19	2024	Yes	Training	Psychological therapy
36	Text	EN	GPT	Fine-tuning	30	2024	Yes	Training	Medical QA
37	Text	EN	Transformer	Fine-tuning	2	2024	Yes	Training	Medical QA
11	Text	EN	GPT	Prompting	5	2024	Yes	Training	General
38	Text	EN	GPT	Prompting	2	2024	Yes	Supplement	SDoH extraction
39	Tabular	EN	GPT	Prompting	<1K	2024	No	Supplement	Medical education
13	Text	EN	GPT	Prompting	30	2024	Yes	Supplement	Dialogue generation
40	Text	EN	GPT	Prompting	14	2024	No	Training	SDoH extraction
41	Tabular	EN	GPT	Prompting	-	2024	No	Training	Gait rehabilitation
42	Text	EN	GPT	Prompting	158	2024	Yes	Training	General
43	Audio	EN	GPT	Prompting	87	2024	Yes	Supplement	EMS assistance
44	Text	EN	GPT	Prompting	37	2024	Yes	Training	Clinical IE
8	Text	EN	GPT, T5	Fine-tuning	148	2024	No	Supplement	Suicide prediction
45	Text	EN	GPT	Prompting	<1K	2024	Yes	Training	Cancer symptom extraction
46	Text	EN	GPT	Prompting	5	2024	Yes	Training	Clinical summary
47	Text	EN	GPT	Prompting	5	2024	Yes	Training	Medical QA
48	Tabular	-	GPT	Prompting	<1K	2024	No	Supplement	Nursing care
49	Text	EN	GPT	Prompting	<1K	2024	Yes	Training	Patient portal triage
50	Text	EN	GPT	Prompting	100	2024	No	Training	Report generation
51	Text	AR	GPT	Prompting	-	2024	Yes	Supplement	Health chatbot
52	Tabular	EN	GPT	Prompting	<1K	2024	No	Supplement	Clinical trial simulation
9	Text	EN	GPT	Prompting	3	2024	No	Supplement	Depression symptom
53	Text	EN	Transformer	Prompting	4	2024	No	Supplement	Autism diagnosis
54	Text	EN	GPT	Prompting	<1K	2025	No	Privacy	Stroke prediction
55	Audio	EN	GPT	Prompting	3	2025	No	Training	Dialogue analysis
56	Text	EN	GPT	Prompting	459	2025	No	Supplement	Phenotype ontology
57	Text	EN	GPT	Prompting	5	2025	No	Training	Radiology prediction
58	Text	ZH	GPT	Prompting	270	2025	No	Training	Anesthesia QA
59	Tabular	-	Transformer	Specialized model	-	2025	Yes	Supplement	Patient outcome prediction
60	Text	ZH	GPT	Prompting	6	2025	Yes	Training	Mental health IE
61	Tabular	EN	GPT	Prompting	6	2025	Yes	Privacy	Data scarcity
62	Text	EN	GPT	Prompting	88	2025	No	Privacy	Phenotype inference
63	Text	EN	GPT	Prompting	5	2025	Yes	Training	Medical reasoning
64	Text	ZH	GPT	Prompting	45	2025	No	Training	Database search
65	Text	EN	GPT	Prompting	74	2025	No	Training	Metastases detection
66	Image	EN	Transformer	Fine-tuning	2	2025	Yes	Supplement	Colonoscopy analysis
67	Image	EN	GPT	Prompting	-	2025	No	Supplement	Pathology analysis
68	Text	ET	GPT	Fine-tuning	4	2025	No	Training	Clinical IE
69	Tabular	EN	GPT	Prompting	10	2025	No	Training	Patient outcome prediction

For example, studies generate synthetic data for augmenting diagnosis accuracy for Post-traumatic stress disorder (PTSD)[29] and Autism Spectrum Disorders (ASD)[53]. English is the predominant language (84.7%, $n = 50$), though there are also studies in French[24], Chinese[30,58,60,64], German[34] and Arabic[51], reflecting some progress toward non-English-speaking patients. the study[34] deployed multilingual LLMs to generate German data to simulate realistic emergency-medical dialogues between ambulance staff and patients. Most datasets focus on unstructured text modality (e.g., clinical narratives[13,21,50] and discharge summaries[28]), while a smaller but growing subset includes tabular[27,39,41,48,52,59,61,69], image[30,66,67], and audio[55]. For example, Theodorou et al.[27] generate novel tabular datasets of high-dimensional longitudinal EHR records to provide realistic, privacy-preserving alternatives for machine learning and statistical analysis, while Ejiga et al.[66] generate synthetic colonoscopy image datasets via fine-tuned text-to-image synthesis to augment training data for robust colorectal cancer detection and precise polyp segmentation. Data sizes vary across health issues (e.g., cancer and mental health), yet the majority has relatively small sets with fewer than 10K samples, indicating the critical utility of synthetic data under low-resourced scenarios. For example, FairPlay[59] generates synthetic data from authentic EHR records to bolster under-represented patient subgroups, boosting mortality-prediction F1 scores by up to 21% and markedly narrowing performance gaps across different demographic groups. The various synthetic datasets highlight a growing community to deploy the biomedical synthetic data.

Synthetic data generation in biomedical research increasingly supports a wide array of clinical applications. Those tasks cover diverse medical needs, including phenotype classification[12,25,45,53,65], PSTD symptom extraction[29], de-identification (SynNote)[20], clinical note summarization [28,46], and mental health assessment [8,9,35]. For example, CALLM[29] constructs clinical interview datasets to aid PTSD diagnosis by generating synthetic transcripts; and Barabadi et al.[65] generated synthetic radiology reports to enable automatic detection of metastatic sites in cancer patients. Those applications broadening the utility of synthetic data for health challenges, such as data shortage[29] and privacy concerns[10,20,54,61,62]. However, accessibility remains a central concern for advancing synthetic data utility. Table 1 shows 50.8% ($n = 30$) of the studies are explicitly described as accessible, providing open or partially open resources for reproducible development, while others' accessibility details or licensing remain unclear. Open datasets such as Syn-HPI[19], ClinGen[11], and Syn-EMS-Audio[43] serve as synthetic data benchmarks for evaluating new models and methods, while others—particularly those derived from sensitive clinical domains—may be restricted or anonymized to protect patient privacy. Ensuring clear, well-documented access protocols and licenses will be essential for fostering broader collaboration and translational impacts of synthetic data in biomedical research.

Synthetic Data Generation Methodology

Emerging Transformer-based large language models (LLMs) have significantly advanced synthetic data generation methodologies in biomedical research, enabling diverse applications, as summarized in Table 1. These novel approaches address challenges associated with limited clinical data by augmenting datasets, thus enhancing generalizability and mitigating overfitting in downstream machine learning tasks across various healthcare domains. In this section, we systematically examine generation methodologies among 59 reviewed studies, summarizing recent trends in model architectures, generation techniques, and integration strategies for synthetic data generation. Specifically, we aim to answer two key questions: 1) what predominant methods (e.g., prompt-based vs. specialized architectures) are currently employed for biomedical synthetic data generation; and 2) what trends and variations do exist regarding model selection (open-source vs. closed-source) and prompting strategies (zero-shot, few-shot, knowledge-infused)?

Our review identifies prompt-based generation as the predominant synthetic data generation method, employed by approximately 72.9% ($n = 43$) of the analyzed studies. Prompt-based approaches[31,52,64] primarily rely on meticulously crafted textual instructions, leveraging fine-tuned LLMs, such as T5 and GPT variants (e.g., GPT-4) for diverse biomedical tasks. Few-shot prompting provides a small number of curated examples (typically 2-5) within the prompt to illustrate the task and guide LLMs for synthetic data generation, which count 20.3% ($n = 12$). The approach uses limited curated examples to guide LLMs in generating clinically relevant synthetic data across diverse applications, including suicide symptom extractions[8], phenotype classification[12,25,45,53,65], and therapeutic dialogue generation[13,35]. For example, Ghanadian et al.[8] use a patient's suicidal narrative and a medication counseling dialogue as few-shot prompts to generate clinical texts for suicide symptom extraction and therapeutic dialogue generation, while Nievas et al.[32] use clinical trial inclusion and exclusion criteria paired with patient summaries as few-shot prompts to generate a synthetic patient–trial matching dataset. Zero-shot prompting involves prompting LLMs without any illustrative examples and replies solely on explicit task instructions and the pretrained knowledge encoded in LLMs[61]. This approach is suitable for simpler tasks, such as dialogue summarization[21] or generating structured medical reports[50], where extensive contextual training is less critical. For example, Barr et al.[61] used zero-shot prompting with GPT-4o to synthesize perioperative clinical tables of patient demographics, surgical parameters and outcomes, showing that the model can produce realistic synthetic data and preserve key statistical patterns based solely on qualitative instructions. Additionally, ClinGen[11] studies utilize knowledge-infused prompting, which explicitly incorporates external biomedical resources, such as clinical ontology or structured knowledge graphs. The prompts integrate the biomedical resources to improve factual accuracy, clinical relevance, and contextual richness. For example, Chuang et al.[62] construct keyword-driven prompts by appropriate

ICD-9 (International Classification of Diseases, 9th Revision) codes per primary diagnosis and incorporate relevant clinical terms into prompts for GPT-3.5 and -4.

A notable trend is the model selection between open-source and closed-source LLMs, such as Llama 4 vs GPT-4. Approximately half of the reviewed studies (67.8%, $n = 40$) directly utilize powerful closed-source models such as GPT-3.5 and GPT-4 through proprietary APIs. This approach prioritizes rapid development and sophisticated output quality without extensive model customization. In contrast, 39.0% ($n = 23$) of prompt-based studies utilize fine-tuned open-source models (e.g., GPT-2[19,20,22,24,25,68], T5[8,37]), offering flexibility to specialize the model with biomedical domain-specific data. Fine-tuning[17,24,26] involves additional training of a pretrained model on a smaller, task-specific dataset, enabling the model to specialize its outputs according to specific biomedical contexts or data types. Fine-tuning open-source models demonstrates advantages in privacy-sensitive or -nuanced clinical contexts[20,22,26], such as assisted suicide detection[8] and cancer diagnosis[45]. For example, Lu et al.[22] fine-tuned GPT-2 on MIMIC-III[70] discharge summaries to generate synthetic clinical notes for data augmentation, effectively mitigating class imbalance, preserving patient privacy, and improving the accuracy of a downstream readmission-prediction model. However, while open-source models offer greater customization potential, they typically underperform closed-source models in general language tasks and clinical reasoning benchmarks[8,12,34,36,57,62]. For example, Jeong et al.[36] report that the open-source model LLaMA2 7B underperforms GPT-4-base on clinical reasoning benchmarks.

Approximately 27.2% ($n = 16$) of the reviewed studies explore specialized model architectures or fine-tuned pre-trained LLMs for synthetic data generation, including multimodal, multi-agent frameworks and custom-designed transformer architectures (e.g., HALO)[13,18,27,41,43,55,59,66,67]. Multimodal approaches integrate multiple data modalities, such as text, images, and audio, to create richer, more realistic synthetic datasets, such as colonoscopy image analysis[66], digital pathology interpretation[67], and emergency medical services (EMS) assistance[43]. For example, CognitiveEMS[43] prompts an LLM with emergency protocols to generate synthetic cardiac-arrest dialogues, converts them to speech, and synchronizes the audio with AR smart-glasses video frames annotated by a zero-shot vision classifier, producing a unified dataset for cardiac emergency analysis. Multi-agent frameworks[13,29,51] involve employing multiple interacting models or agents, each performing distinct roles, to collaboratively produce complex and realistic data, such as doctor-patient dialogues or clinical scenarios. For instance, NoteChat[13] employs a cooperative multi-agent framework with planning, role play and refinement modules to generate synthetic patient–physician dialogues that mirror real clinical encounters. Custom transformer architectures[18,27,59] extend the standard decoder by integrating domain-specific metadata and employing hierarchical modeling strategies, first capturing broader structural information and then refining finer details, to generate structured and realistic synthetic clinical datasets tailored to specific medical scenarios. For example, Theodorou et al.[27] demonstrate that their hierarchical autoregressive model generates privacy-preserving EHRs whose temporal and code distributions yield downstream prediction performance on par with real data. Collectively, these specialized frameworks enhance the fidelity and applicability of synthetic biomedical data for various complex health scenarios.

Quality Assessment and Evaluation Metrics

Ensuring quality of the synthetic data is the key to build precise models and medical applications. In this section, we assess our collected studies from the synthetic data assessment and evaluation perspective, as shown in Table 2. We aim to answer the following three questions: 1) What metrics and evaluation approaches are commonly used for assessing synthetic data quality? 2) What are major challenges in model evaluations using synthetic data across biomedical diseases and tasks; and 3) What are the emerging trends in synthetic data quality control for biomedical research? To answer those questions, we systematically examined the collected 59 studies by disease topics, types of evaluation metrics (automated and human approaches), downstream biomedical tasks, and emerging strategies (e.g., LLM-based and intrinsic evaluations). We aim to highlight current practices, identify persistent challenges, and discuss notable shifts toward more robust and clinically meaningful data and model assessments in the biomedical synthetic data generation.

We examined two major evaluation metrics across diverse biomedical applications, intrinsic and extrinsic metrics, which are the essential measurements to select synthetic data and computational models. Intrinsic metrics (e.g., perplexity[42] or similarity measures such as KL Divergence[57]) assess properties of the synthetic data independently from downstream tasks, while extrinsic metrics rely on downstream tasks, such as classification[18]. As shown in Table 2, we can observe that the intrinsic metrics are not widely adopted and only count 27.1% ($n = 16$) of the studies. For example, Study[11] and Study[46] employ intrinsic metrics to directly compare the statistical distribution of real and synthetic datasets, while most works, such as Studies[22,25], and[56], rely primarily on extrinsic evaluation through classification (CLS) or information extraction (IE) tasks. Commonly used metrics include accuracy[26], F1 score[25], BLEU[30], and other task-specific measurements, such as ROUGE[63] and AUCROC[22]. The studies cover heterogeneous downstream tasks, such as or de-identification for privacy protection[10], phenotype prediction[56], disease entity extraction[42], and clinical note summarization[21,46]. Privacy protection assessment was reported in 20.3% ($n = 12$) of the studies, primarily through membership and attribute inference attacks[10], adversary test[27], and memorization evaluation[18]. Those studies cover multiple medical domains, with most studies focusing on general biomedical applications (n=33) and others targeting 25 specific medical conditions and health-related topics, such as Alzheimer's[12], breast

Table 2. Evaluation characteristics of reviewed studies: disease, downstream task, number of automated metrics (Metrics #), inclusion of intrinsic evaluation (Intrinsic+), human involvement in generation or task evaluations (Human-in-the-Loop), and LLM-based evaluation (LLM Eval). CLS, NER, IE, RE, and QA refer to downstream tasks of classification, named entity, recognition, information extraction, relation extraction, and question answering, respectively.

Study	Disease	Task	Metrics #	Intrinsic+	Human-in-the-Loop	LLM Eval
17	General	QA	3	Yes	No	No
18	Mental health	CLS	6	Yes	Yes	No
19	General	NER	4	No	Yes	No
20	General	NER	5	No	Yes	No
21	General	Summarization	3	No	Yes	No
22	General	CLS	3	No	No	No
23	General	RE, QA	4	No	Yes	No
24	General	NER	6	Yes	Yes	No
25	Febrile convulsions	CLS	3	No	Yes	No
26	Suicide	CLS	2	No	Yes	No
27	General	CLS	8	No	Yes	No
12	Alzheimer	CLS	4	No	Yes	No
28	General	CLS	3	No	Yes	No
29	PTSD	CLS	3	No	Yes	No
30	Fundus Fluorescein Angiography	QA	5	No	Yes	No
31	Acute Renal Failure	CLS	7	No	Yes	No
10	General	CLS	5	No	No	No
32	General	CLS	8	No	Yes	No
33	General	NLI	5	No	No	No
34	General	-	1	Yes	No	No
35	Psychological Therapy	-	9	Yes	No	No
36	General	QA	3	No	No	No
37	General	QA	8	No	Yes	No
11	General	CLS, RE, NER	7	Yes	No	No
38	General	IE, CLS	3	No	Yes	No
9	Depression	Semantic Search	4	No	No	No
13	General	Dialogue Generation	7	Yes	Yes	Yes
40	General	CLS	7	No	No	No
41	Gait-based Disease	Pose Estimation	4	Yes	No	No
42	General	IE	1	Yes	Yes	Yes
43	General	ASR	2	No	No	No
44	General	NER, RE	1	No	No	No
8	Suicide	CLS	2	No	Yes	No
45	Cancer	CLS	2	No	No	No
46	General	Summarization	4	Yes	Yes	Yes
47	General	QA	1	No	No	No
48	Skeleton pose	CLS	2	Yes	No	No
49	General	-	2	Yes	Yes	No
50	General	Summarization	6	No	No	Yes
51	General	QA	4	No	Yes	Yes
52	Breast cancer	CLS	6	Yes	No	No
39	COVID-19	CLS	1	No	Yes	No
53	Autism	CLS	3	No	Yes	No
54	Stroke thrombolysis contraindications	CLS	6	No	Yes	No
55	General	CLS	4	No	Yes	No
56	Phenotype ontology	IE	3	No	No	No
57	Limb Fractures	CLS	3	Yes	No	Yes
58	Anesthesiology	QA	5	No	Yes	Yes
59	General	CLS	2	No	No	No
60	Mental health	IE	2	No	Yes	No
61	Perioperative clinical data	-	1	Yes	No	No
62	General	CLS	3	Yes	No	No
63	General	QA	5	No	Yes	Yes
64	General	CLS	2	No	No	No
65	Cancer	CLS	1	No	Yes	No
66	Colorectal cancer	CLS, Segmentation	7	No	No	No
67	Breast cancer	CLS, Segmentation	4	No	Yes	No
68	General	NER	3	No	Yes	No
69	Breast Cancer, Diabetes	CLS	3	No	No	No

cancer[52], and depression[9].

Beyond automated metrics, human evaluation has played a selective yet important role in the assessment of synthetic biomedical data or its utility in different tasks (e.g., phenotype inference or clinical note summary), particularly for tasks where clinical relevance or nuanced interpretation is required. Human evaluation[12,13,32] typically involves domain experts to determine if synthetic data are meaningful, such as clinicians or biomedical researchers, who assess the synthetic data for factors like clinical validity. Human evaluation was included in a minority of studies (44.06%). Those studies cover tasks such as clinical entity recognition[19], clinical note summarization[21,46], or question answering[37], where expert judgment adds valuable context beyond quantitative scores. More recently, a few studies[13,42,46,50] (13.55% of our studies) explore using large language models (LLMs) as automated evaluators and leverage their ability to perform nuanced judgments, which approximates human assessment. Common LLM-based evaluation metrics employed in these studies include factual consistency[50], clinical correctness[42], and error identification[51], often measured through prompt-based scoring or rubric-guided assessments conducted by the LLM itself. For example, a recent study[50] prompts GPT-4 if the context supports / contradicts the response, or if there's not enough information as a hallucination evaluation, whereas another study[42] uses a rubric-guided prompt, requesting GPT-4 to score each response on a four-level scale, Unacceptable, Poor, Satisfactory, and Excellent. However, LLM-based evaluations remain an emerging approach and have not yet replaced the need for statistical metrics or domain expert involvement, especially in clinically sensitive or complex scenarios[71].

Despite substantial progress, several challenges persist in the quality control of biomedical synthetic data. A key issue is the lack of standardized evaluation frameworks, resulting in wide variability in both the selection and reporting of metrics across studies and disease domains[11]. This heterogeneity complicates direct comparison between methods and limits the transferability of findings across medical applications and diseases. Additionally, Table 2 shows that studies still underutilize intrinsic evaluation, focusing instead on extrinsic, task-specific benchmarks that may not fully capture utility or limitations of synthetic data. Emerging trends suggest a gradual shift toward more comprehensive assessments, including extrinsic and intrinsic evaluation integrations, LLM-based metrics adoptions, and human professional engagements.

Discussion

Our scoping review has presented the current developments and applications of biomedical synthetic data generation, which is being facilitated by large language models (LLMs). This section, we highlight several insightful takeaways of emerging methodological and evaluative trends and identify promising areas for future biomedical research and applications using synthetic data generation.

This review shows a clear shift towards leveraging synthetic data to overcome biomedical research challenges, such as data scarcity, privacy restrictions, and diverse clinical conditions. While synthetic data has been used effectively across varied clinical applications (e.g., phenotype ontology extraction[56] and cancer diagnosis[45,66,67]), the scale still vary across data types and disease fields, reflecting the nuanced capabilities and limitations of current LLMs. Clinical narratives and unstructured texts remain the most common data source, due primarily to their widespread availability and the inherent LLMs strengths in natural language processing tasks. We can also find an emerging trend toward multimodal and structured synthetic EHR, which indicates increasing expansions across more clinical settings, such as imaging, audio, and wearable sensors[41]. However, the advances indicate potential concerns in methodological developments. Our review shows that current studies are significant depending on prompting closing-sourced LLMs, like GPT-4. While few-shot and zero-shot approaches may achieve some successful cases[29,31,44], prompting closing-sourced LLMs (e.g., GPT-4) can limit transparency, reproducibility, and the ability to customize models for specialized clinical contexts. Additionally, access to proprietary models may be restricted by cost, availability, or data governance policies. For example, MIMIC-IV (Medical Information Mart for Intensive Care)[72] does not allow for uploading data to those closing-sourced LLMs. Future work should address these concerns by promoting the development and benchmarking of open-source models and by exploring alternative methods that reduce dependence on proprietary LLMs.

Our review indicates future research in biomedical synthetic data generation may evolve along three dimensions. There is a clear trend towards generating multimodal synthetic datasets to simulate more real-world and complex clinical scenarios, such as radiology[57,65], pathology[67], and emergency care[34]. Future synthetic data generation may depend on more precise LLM-based systems by knowledge-guided generations[11], multi-agent frameworks[13], and human-LLM collaborations[39], which leverage clinical knowledge to enhance factuality and specificity of synthetic data qualities. In terms of evaluation, more standardized, multi-dimensional, and human-centered assessments are critical. Future studies on biomedical synthetic data generation may consider a mixture of intrinsic and extrinsic metrics, human-in-the-loop approaches, and task-agnostic validations to ensure fidelity, utility, and privacy of synthetic data. We envision the directions will essential to advance biomedical research and applications by the synthetic data generations and LLMs.

Limitations This scoping review has several limitations that should be acknowledged to appropriately interpret our findings and analysis. First, the literature search was limited to specific databases and publication venues, potentially missing relevant studies published in less prominent journals or literature sources. Second, our review primarily included studies written in English, potentially overlooking important findings published in other languages. Additionally, the rapid pace of research in LLM-driven synthetic data generation means that emerging studies and recent methodological advances might not be captured fully in this review. Lastly, given the inherent heterogeneity in study designs, evaluation metrics, and application domains across the reviewed literature, direct comparison and generalization of findings are challenging, underscoring the need for more standardized reporting and evaluation frameworks in future research.

Methods

Data Source and Search. We conduct searches between April and May 2025 following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines[73] and the PRISMA Extension for Scoping Reviews (PRISMA-ScR)[74] (Figure 2). Recognizing the rapid developments of Large Language Models (LLMs) in the biomedical field, our review encompassed peer-reviewed articles between January 1, 2020, to April 5, 2025 (inclusive) from multiple database sources, including ACM Digital Library, PubMed, Web of Science, and Google Scholar. Supplemental articles were gathered by reviewing article bibliographies and by soliciting suggestions from XH, HW, IH, and ZH. Our search strategy utilized a logical combination of keywords: "Language Model" AND "synthetic data" AND "medical OR health OR clinical". We primarily limited the search to titles and abstracts, expanding to full texts where this function was unavailable. For Google Scholar, we conducted a keyword search, sorted by relevance, and selected the first 100 studies. The search strategies for each database were initially formulated in the early stages of the study and refined through team discussions and preliminary analysis of the results.

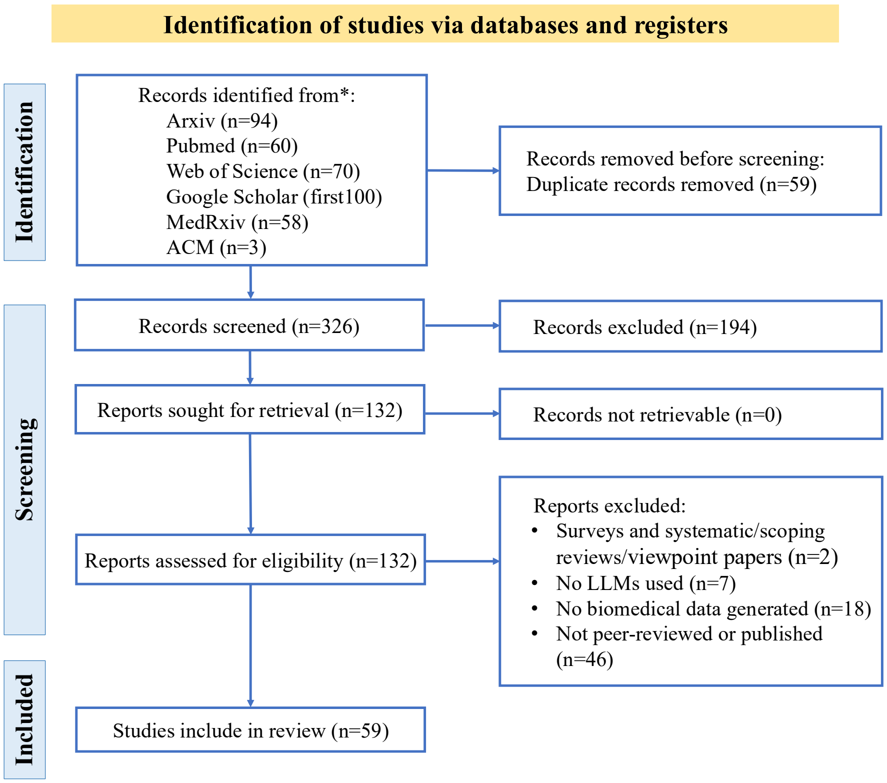
Identification of studies via databases and registers
Identification
Records identified from*:
Arxiv (n=94)
Pubmed (n=60)
Web of Science (n=70)
Google Scholar (first100)
MedRxiv (n=58)
ACM (n=3)
Records removed before screening:
Duplicate records removed (n=59)
Screening
Records screened (n=326)
Records excluded (n=194)
Reports sought for retrieval (n=132)
Records not retrievable (n=0)
Reports assessed for eligibility (n=132)
Reports excluded:
• Surveys and systematic/scoping reviews/viewpoint papers (n=2)
• No LLMs used (n=7)
• No biomedical data generated (n=18)
• Not peer-reviewed or published (n=46)
Included
Studies include in review (n=59)

Figure 2. PRISMA-ScR flow diagram.

Study Selection. We followed inclusion and exclusion criteria to narrow the candidate articles. Titles and abstracts were initially screened against predefined eligibility criteria by at least two independent reviewers, excluding records that clearly did not meet inclusion criteria and retaining ambiguous ones for full-text screening. Our inclusion criteria include: 1) an article conducts biomedical research; 2) an article uses Large Language Models (LLMs) for synthetic generation of biomedical data; and 3) an article has undergone peer review and been published in a journal or in conference proceedings within the range of

01/01/2020 and 04/05/2025. Our exclusion criteria include: 1) an article is a survey, literature review, news article, editorial, letter, opinion, study protocol, and comment; 2) an article is not in English; 3) articles do not study biomedical large language models; and 4) an article is not peer-reviewed and published in a journal or conference proceedings. Three undergraduate students from the XH lab assisted with review of candidate articles with the HS and XH, who divided the candidate articles into relatively equal numbers and assigned them to reviewers. Reviewers examined titles, abstracts, full-text, and other article metadata, recommending inclusion or exclusion based on the defined criteria. Each article was reviewed by at least two individuals, and disagreements were resolved after further review by XH and HW.

Data Extraction. Selected articles were mainly processed by two individuals (HR and XH), who are experts in natural language processing (NLP) and biomedical informatics. They independently reviewed the full text of each selected article and extracted key metadata. The extracted data focused on three main aspects: 1) Synthetic data types and applications, including data characteristics , such as language and domain; 2) Synthetic data generation methods, presenting generation pipelines (models used to generate synthetic data) and their data modalities (e.g., text, image, audio); 3) Quality assessment and evaluation metrics summarizing medical applications, evaluation approaches, and evaluation settings. Discrepancies in data extraction were resolved through discussion between HR and XH, with further arbitration by XH, HW, and IH when necessary. All extraction processes were documented for replication purposes.

References

1. Liu, W., He, Z. & Huang, X. Time matters: Examine temporal effects on biomedical language models. In *AMIA Annual Symposium Proceedings*, vol. 2024, 723–732 (San Francisco, CA, USA, 2025).
2. Jones, P., Liu, W., Huang, I.-C. & Huang, X. Examining imbalance effects on performance and demographic fairness of clinical language models (2025). 2412.17803.
3. Stade, E. C. *et al.* Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Heal. Res.* **3**, 12, DOI: 10.1038/s44184-024-00056-z (2024).
4. Blanco-Gonzalez, A. *et al.* The role of ai in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals* **16**, 891, DOI: 10.3390/ph16060891 (2023).
5. Achiam, J. *et al.* Gpt-4 technical report. Tech. Rep., OpenAI (2023). 2303.08774.
6. Grattafiori, A. *et al.* The llama 3 herd of models. Tech. Rep., Meta AI (2024). 2407.21783.
7. Han, G., Liu, W., Huang, X. & Borsari, B. Chain-of-interaction: Enhancing large language models for psychiatric behavior understanding by dyadic contexts. In *12th IEEE International Conference on Healthcare Informatics, ICHI 2024, Orlando, FL, USA, June 3-6, 2024*, 392–401, DOI: 10.1109/ICHI61247.2024.00057 (IEEE, 2024).
8. Ghanadian, H., Nejadgholi, I. & Osman, H. A. Socially aware synthetic data generation for suicidal ideation detection using large language models. *IEEE Access* **12**, 14350–14363, DOI: 10.1109/ACCESS.2024.3358206 (2024).
9. Bucur, A.-M. Leveraging llm-generated data for detecting depression symptoms on social media. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part I*, 193–204, DOI: 10.1007/978-3-031-71736-9_14 (Springer-Verlag, Berlin, Heidelberg, 2024).
10. Sarkar, A. R., Chuang, Y.-S., Mohammed, N. & Jiang, X. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Sci. Reports* **14**, 29669, DOI: 10.1038/s41598-024-81170-y (2024).
11. Xu, R. *et al.* Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. In Ku, L.-W., Martins, A. & Srikumar, V. (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*, 15496–15523, DOI: 10.18653/v1/2024.findings-acl.916 (Association for Computational Linguistics, Bangkok, Thailand, 2024).
12. Li, R., Wang, X. & Yu, H. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In Bouamor, H., Pino, J. & Bali, K. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*, 7129–7143, DOI: 10.18653/v1/2023.findings-emnlp.474 (Association for Computational Linguistics, Singapore, 2023).
13. Wang, J. *et al.* NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In Ku, L.-W., Martins, A. & Srikumar, V. (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*, 15183–15201, DOI: 10.18653/v1/2024.findings-acl.901 (Association for Computational Linguistics, Bangkok, Thailand, 2024).

14. Figueira, A. & Vaz, B. Survey on synthetic data generation, evaluation methods and gans. *Mathematics* **10**, 2733, DOI: https://doi.org/10.3390/math10152733 (2022).

15. Long, L. *et al.* On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In Ku, L.-W., Martins, A. & Srikumar, V. (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*, 11065–11082, DOI: 10.18653/v1/2024.findings-acl.658 (Association for Computational Linguistics, Bangkok, Thailand, 2024).

16. Li, Z., Zhu, H., Lu, Z. & Yin, M. Synthetic data generation with large language models for text classification: Potential and limitations. In Bouamor, H., Pino, J. & Bali, K. (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10443–10461, DOI: 10.18653/v1/2023.emnlp-main.647 (Association for Computational Linguistics, Singapore, 2023).

17. Shakeri, S. *et al.* End-to-end synthetic data generation for domain adaptation of question answering systems. In Webber, B., Cohn, T., He, Y. & Liu, Y. (eds.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5445–5460, DOI: 10.18653/v1/2020.emnlp-main.439 (Association for Computational Linguistics, Online, 2020).

18. Ive, J. *et al.* Generation and evaluation of artificial mental health records for natural language processing. *npj Digit. Medicine* **3**, DOI: 10.1038/s41746-020-0267-x (2020).

19. Li, J. *et al.* Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *J. Am. Med. Informatics Assoc.* **28**, 2193–2201, DOI: 10.1093/jamia/ocab112 (2021).

20. Libbi, C. A., Trienes, J., Trieschnigg, D. & Seifert, C. Generating synthetic training data for supervised de-identification of electronic health records. *Futur. Internet* **13**, DOI: 10.3390/fi13050136 (2021).

21. Chintagunta, B., Katariya, N., Amatriain, X. & Kannan, A. Medically aware GPT-3 as a data generator for medical dialogue summarization. In Shivade, C. *et al.* (eds.) *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, 66–76, DOI: 10.18653/v1/2021.nlpmc-1.9 (Association for Computational Linguistics, Online, 2021).

22. Lu, Q., Dou, D. & Nguyen, T. H. Textual data augmentation for patient outcomes prediction. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2817–2821, DOI: 10.1109/BIBM52615.2021.9669861 (2021).

23. Peng, C. *et al.* A study of generative large language model for medical research and healthcare. *npj Digit. Medicine* **6**, 210, DOI: 10.1038/s41746-023-00958-w (2023).

24. Hiebel, N., Ferret, O., Fort, K. & Névéol, A. Can synthetic text help clinical named entity recognition? a study of electronic health records in French. In Vlachos, A. & Augenstein, I. (eds.) *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2320–2338, DOI: 10.18653/v1/2023.eacl-main.170 (Association for Computational Linguistics, Dubrovnik, Croatia, 2023).

25. Khademi, S. *et al.* Data augmentation to improve syndromic detection from emergency department notes. In *ACM International Conference Proceeding Series*, 198–205, DOI: 10.1145/3579375.3579401 (Association for Computing Machinery, 2023).

26. Spitale, G., Schneider, G., Germani, F. & Biller-Andorno, N. Exploring the role of ai in classifying, analyzing, and generating case reports on assisted suicide cases: feasibility and ethical implications. *Front. Artif. Intell.* **6**, DOI: 10.3389/frai.2023.1328865 (2023).

27. Theodorou, B., Xiao, C. & Sun, J. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nat. Commun.* **14**, DOI: 10.1038/s41467-023-41093-0 (2023).

28. Sufi, F. Addressing data scarcity in the medical domain: A gpt-based approach for synthetic data generation and feature extraction. *Inf. (Switzerland)* **15**, DOI: 10.3390/info15050264 (2024).

29. Wu, Y., Mao, K., Zhang, Y. & Chen, J. Callm: Enhancing clinical interview analysis through data augmentation with large language models. *IEEE J. Biomed. Heal. Informatics* DOI: 10.1109/JBHI.2024.3435085 (2024).

30. Chen, X. *et al.* Chatffa: An ophthalmic chat system for unified vision-language understanding and question answering for fundus fluorescein angiography. *iScience* **27**, 110021, DOI: 10.1016/j.isci.2024.110021 (2024).

31. Litake, O., Park, B. H., Tully, J. L. & Gabriel, R. A. Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *J. Am. Med. Informatics Assoc.* **31**, 1404–1410, DOI: 10.1093/jamia/ocae081 (2024). https://academic.oup.com/jamia/article-pdf/31/6/1404/57769046/ocae081.pdf.

32. Nievas, M., Basu, A., Wang, Y. & Singh, H. Distilling large language models for matching patients to clinical trials. *J. Am. Med. Informatics Assoc.* **31**, 1953–1963, DOI: 10.1093/jamia/ocae073 (2024). https://academic.oup.com/jamia/article-pdf/31/9/1953/58868048/ocae073.pdf.

33. Wang, Y. *et al.* DKE-research at SemEval-2024 task 2: Incorporating data augmentation with generative models and biomedical knowledge to enhance inference robustness. In Ojha, A. K. *et al.* (eds.) *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, 88–94, DOI: 10.18653/v1/2024.semeval-1.15 (Association for Computational Linguistics, Mexico City, Mexico, 2024).

34. Moser, D., Bender, M. & Sariyar, M. Generating synthetic healthcare dialogues in emergency medicine using large language models. *Stud. health technology informatics* **321**, 235–239, DOI: 10.3233/SHTI241099 (2024).

35. Bird, J. J., Wright, D., Sumich, A. & Lotfi, A. Generative ai in psychological therapy: Perspectives on computational linguistics and large language models in written behaviour monitoring. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '24, 322–328, DOI: 10.1145/3652037.3663893 (Association for Computing Machinery, New York, NY, USA, 2024).

36. Jeong, M., Sohn, J., Sung, M. & Kang, J. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics* **40**, i119–i129, DOI: 10.1093/bioinformatics/btae238 (2024).

37. Zafar, A., Sahoo, S. K., Bhardawaj, H., Das, A. & Ekbal, A. Ki-mag: A knowledge-infused abstractive question answering system in medical domain. *Neurocomputing* **571**, DOI: 10.1016/j.neucom.2023.127141 (2024).

38. Guevara, M. *et al.* Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine* **7**, 6, DOI: 10.1038/s41746-023-00970-0 (2024).

39. Ehrett, C., Hegde, S., Andre, K., Liu, D. & Wilson, T. Leveraging open-source large language models for data augmentation in hospital staff surveys: Mixed methods study. *JMIR Med. Educ.* **10**, e51433–e51433, DOI: 10.2196/51433 (2024).

40. Gabriel, R. A. *et al.* On the development and validation of large language model-based classifiers for identifying social determinants of health. *Proc. Natl. Acad. Sci. United States Am.* **121**, DOI: 10.1073/pnas.2320716121 (2024).

41. Gao, Y. *et al.* Pressinpose: Integrating pressure and inertial sensors for full-body pose estimation in activities. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **8**, DOI: 10.1145/3699773 (2024).

42. Kweon, S. *et al.* Publicly shareable clinical large language model built on synthetic clinical notes. In Ku, L.-W., Martins, A. & Srikumar, V. (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*, 5148–5168, DOI: 10.18653/v1/2024.findings-acl.305 (Association for Computational Linguistics, Bangkok, Thailand, 2024).

43. Weerasinghe, K. *et al.* Real-time multimodal cognitive assistant for emergency medical services. In *2024 IEEE/ACM Ninth International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 85–96, DOI: 10.1109/IoTDI61053.2024.00012 (2024).

44. Delmas, M., Wysocka, M. & Freitas, A. Relation extraction in underexplored biomedical domains: A diversity-optimized sampling and synthetic data generation approach. *Comput. Linguist.* **50**, 953–1000, DOI: 10.1162/coli_a_00520 (2024).

45. Zeinali, N., Albashayreh, A., Fan, W. & White, S. G. Symptom-bert: Enhancing cancer symptom detection in ehr clinical notes. *J. Pain Symptom Manag.* **68**, 190–198.e1, DOI: 10.1016/j.jpainsymman.2024.05.015 (2024).

46. Mishra, P. *et al.* SYNFAC-EDIT: Synthetic imitation edit feedback for factual alignment in clinical summarization. In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20061–20083, DOI: 10.18653/v1/2024.emnlp-main.1120 (Association for Computational Linguistics, Miami, Florida, USA, 2024).

47. Zhang, J., Cui, W., Huang, Y., Das, K. & Kumar, S. Synthetic knowledge ingestion: Towards knowledge refinement and injection for enhancing large language models. In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 21456–21473, DOI: 10.18653/v1/2024.emnlp-main.1196 (Association for Computational Linguistics, Miami, Florida, USA, 2024).

48. Dobhal, U., Garcia, C. & Inoue, S. Synthetic skeleton data generation using large language model for nurse activity recognition. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '24, 493–499, DOI: 10.1145/3675094.3678445 (Association for Computing Machinery, New York, NY, USA, 2024).

49. Wang, N. *et al.* Taxonomy-based prompt engineering to generate synthetic drug-related patient portal messages. *J. Biomed. Informatics* **160**, DOI: 10.1016/j.jbi.2024.104752 (2024).

50. Jones, E. *et al.* Teaching language models to hallucinate less with synthetic tasks. In *12th International Conference on Learning Representations, ICLR 2024* (2024).

51. Alghamdi, H. M. & Mostafa, A. Towards reliable healthcare llm agents: A case study for pilgrims during hajj. *Inf. (Switzerland)* **15**, DOI: 10.3390/info15070371 (2024).

52. Wang, Y. *et al.* Twin-gpt: Digital twins for clinical trials via large language model. *ACM Trans. Multimed. Comput. Commun. Appl.* DOI: 10.1145/3674838 (2024).

53. Woolsey, C. R., Bisht, P., Rothman, J. & Leroy, G. Utilizing large language models to generate synthetic data to increase the performance of BERT-Based neural networks. *AMIA Jt Summits Transl Sci Proc* **2024**, 429–438 (2024).

54. Chen, B. Y. *et al.* Automated identification of stroke thrombolysis contraindications from synthetic clinical notes: A proof-of-concept study. *Cerebrovasc. diseases extra* **15**, 130–136, DOI: 10.1159/000545317 (2025).

55. Scroggins, J. K., Topaz, M., Song, J. & Zolnoori, M. Does synthetic data augmentation improve the performances of machine learning classifiers for identifying health problems in patient-nurse verbal communications in home healthcare settings? *J. Nurs. Scholarsh.* **57**, 47–58, DOI: 10.1111/jnu.13004 (2024).

56. Albayrak, A. *et al.* Enhancing human phenotype ontology term extraction through synthetic case reports and embedding-based retrieval: A novel approach for improved biomedical data annotation. *J. Pathol. Informatics* **16**, DOI: 10.1016/j.jpi.2024.100409 (2025).

57. Liu, J., Koopman, B., Brown, N. J., Chu, K. & Nguyen, A. Generating synthetic clinical text with local large language models to identify misdiagnosed limb fractures in radiology reports. *Artif. Intell. Medicine* **159**, DOI: 10.1016/j.artmed.2024.103027 (2025).

58. Wang, Z. *et al.* Hypnos: A domain-specific large language model for anesthesiology. *Neurocomputing* **624**, DOI: 10.1016/j.neucom.2025.129389 (2025).

59. Theodorou, B. *et al.* Improving medical machine learning models with generative balancing for equity and excellence. *npj Digit. Medicine* **8**, DOI: 10.1038/s41746-025-01438-z (2025).

60. Cai, Z. *et al.* Improving unified information extraction in chinese mental health domain with instruction-tuned llms and type-verification component. *Artif. Intell. Medicine* **162**, DOI: 10.1016/j.artmed.2025.103087 (2025).

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

62. Chuang, Y.-S., Sarkar, A. R., Hsu, Y.-C., Mohammed, N. & Jiang, X. Robust privacy amidst innovation with large language models through a critical assessment of the risks. *J. Am. Med. Informatics Assoc.* **32**, 885–892, DOI: 10.1093/jamia/ocaf037 (2025).

63. Kim, H. *et al.* Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine* **8**, 240, DOI: 10.1038/s41746-025-01653-8 (2025).

64. Li, J., Wang, Z., Yu, L., Liu, H. & Song, H. Synthetic data-driven approaches for chinese medical abstract sentence classification: Computational study. *JMIR Form. Res.* **9**, DOI: 10.2196/54803 (2025).

65. Barabadi, M. A., Zhu, X., Chan, W. Y., Simpson, A. L. & Do, R. K. Targeted generative data augmentation for automatic metastases detection from free-text radiology reports. *Front. Artif. Intell.* **8**, DOI: 10.3389/frai.2025.1513674 (2025).

66. Peter, O. O. E., Adeniran, O. T., John-Otumu, A. M. G., Khalifa, F. & Rahman, M. M. Text-guided synthesis in medical multimedia retrieval: A framework for enhanced colonoscopy image classification and segmentation. *Algorithms* **18**, DOI: 10.3390/a18030155 (2025).

67. Li, J. *et al.* Topofm: Topology-guided pathology foundation model for high-resolution pathology image synthesis with cellular-level control. *IEEE Transactions on Med. Imaging* DOI: 10.1109/TMI.2025.3548872 (2025).

68. Šuvalov, H. *et al.* Using synthetic health care data to leverage large language models for named entity recognition: Development and validation study. *J. Med. Internet Res.* **27**, DOI: 10.2196/66279 (2025).

69. Miletic, M. & Sariyar, M. Utility-based analysis of statistical approaches and deep learning models for synthetic data generation with focus on correlation structures: Algorithm development and validation. *JMIR AI* **4**, DOI: 10.2196/65729 (2025).

70. Johnson, A. E. *et al.* Mimic-iii, a freely accessible critical care database. *Sci. Data* **3**, 160035, DOI: 10.1038/sdata.2016.35 (2016).

71. Chen, X. *et al.* Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intell. Medicine* **5**, 151–163, DOI: https://doi.org/10.1016/j.imed.2025.03.002 (2025).

72. Johnson, A. E. W. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1, DOI: 10.1038/s41597-022-01899-x (2023).

73. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **169**, n71, DOI: 10.1136/bmj.n71 (2021).

74. Tricco, A. C. *et al.* Prisma extension for scoping reviews (prisma-scr): Checklist and explanation. *Annals Intern. Medicine* **169**, 467–473, DOI: 10.7326/M18-0850 (2018). PMID: 30178033, https://doi.org/10.7326/M18-0850.

Acknowledgments

WR, WL, and WH were partially supported by the National Science Foundation under Award IIS-2245920 and IIS-2440381, and the work was partially supported by National Cancer Institute Award R01CA258193. ZH was supported by the Agency for Healthcare Research and Quality Award R21HS029969, the National Library of Medicine Award R21LM013911, the National Institute of Mental Health Award R21MH137736, and the National Institute on Aging Award R01AG064529. The authors would like to thank three undergraduate students, Sampoorna Poria, Sashank Reddy Karri, and Precious T Jones, for assisting paper collection and filtering at the initial stage.

Author contributions statement

H.R. and W.L. led the literature search, data extraction, and formal analysis. H.R. and W.L. made equal contributions to this study. H.W. contributed to initial data collection, methodology development, and manuscript preparation. W.L. joined the process at a later stage. IC.H. and Z.H. provided domain expertise, supervised the review process, and revised the manuscript. X.H. conceived and designed the study, coordinated the project, and contributed to writing, editing, and finalizing the manuscript. All authors reviewed, edited, and approved the final manuscript.

Additional information

Competing interests The authors declare no competing interests.

Data availability Not applicable.