JMIR Diabetes

Emerging Technologies, Medical Devices, Apps, Sensors, and Informatics to Help People with Diabetes
Volume 11 (2026)   ISSN 2371-4379   Editors-in-Chief: Ricardo Correa, MD, EdD; Sheyu Li, MD

Contents

Original Papers

Research Letter

GraphRAG-Enabled Local Large Language Model for Gestational Diabetes Mellitus: Development of a Proof-of-Concept

Edmund Evangelista[1*], PhD; Fathima Ruba[2*], MSc; Salman Bukhari[3*], BEng; Amril Nazir[1*], PhD; Ravishankar Sharma[1*], PhD

[1]College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

[2]Department of Computer Science and Artificial Intelligence, Plaksha University, Chandigarh, India

[3]Department of Electrical Engineering, Capital University of Science and Technology, Islamabad, Pakistan

[*]all authors contributed equally

Corresponding Author:
Ravishankar Sharma, PhD
College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

Abstract

Background: Gestational diabetes mellitus (GDM) is a prevalent chronic condition that affects maternal and fetal health outcomes worldwide, increasingly in underserved populations. While generative artificial intelligence (AI) and large language models (LLMs) have shown promise in health care, their application in GDM management remains underexplored.

Objective: This study aimed to investigate whether retrieval-augmented generation techniques, when combined with knowledge graphs (KGs), could improve the contextual relevance and accuracy of AI-driven clinical decision support. For this, we developed and validated a graph-based retrieval-augmented generation (GraphRAG)–enabled local LLM as a clinical support tool for GDM management, assessing its performance against open-source LLM tools.

Methods: A prototype clinical AI assistant was developed using a GraphRAG constructed from 1212 peer-reviewed research articles on GDM interventions, retrieved from the Semantic Scholar API (2000 - 2024). The GraphRAG prototype integrated entity extraction, KG construction using Neo4j, and retrieval-augmented response generation. The performance was evaluated in a simulated environment using clinical and layperson prompts, comparing the outputs of the systems against ChatGPT (OpenAI), Claude (Anthropic), and BioMistral models across 5 common natural language generation metrics.

Results: The GraphRAG-enabled local LLM showed higher accuracy in generating clinically relevant responses. It achieved a bilingual evaluation understudy score of 0.99, Jaccard similarity of 0.98, and BERTScore of 0.98, outperforming the benchmark LLMs. The prototype also produced accurate, evidence-based recommendations for clinicians and patients, demonstrating its feasibility as a clinical support tool.

Conclusions: GraphRAG-enabled local LLMs show much potential for improving personalized GDM care by integrating domain-specific evidence and contextual retrieval. Our prototype proof-of-concept serves two purposes: (1) the local LLM architecture gives practitioners from underserved locations access to state-of-the-art medical research in the treatment of chronic conditions and (2) the KG schema may be feasibly built on peer-reviewed, indexed publications, devoid of hallucinations and contextualized with patient data. We conclude that advanced AI techniques such as KGs, retrieval-augmented generation, and local LLMs improve GDM management decisions and other similar conditions and advance equitable health care delivery in resource-constrained health care environments.

(JMIR Diabetes 2026;11:e76454) doi:10.2196/76454

KEYWORDS

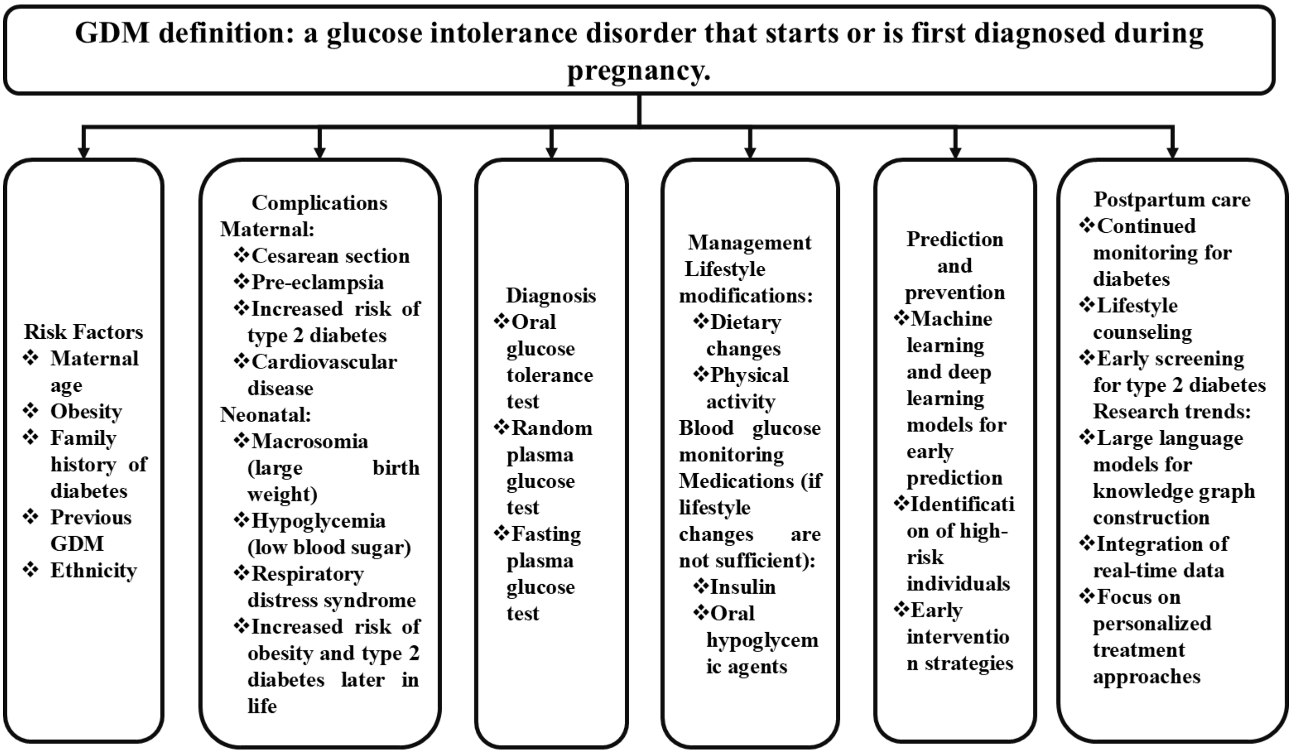
artificial intelligence for health care; generative AI; knowledge graph; retrieval augmented generation; large language model; gestational diabetes mellitus; explainable AI in medicine; GDM; artificial intelligence

Introduction

The growing use of electronic medical records linking diverse patient characteristics and prescription choices with positive treatment outcomes in large-scale use cases has resulted in platforms that guide optimal treatment options. For example, Sharma et al [1] presented an approach for delivering personalized health care as a means of effectively using scarce medical resources in underserved regions and populations, supporting the value of artificial intelligence (AI)–driven systems in such settings. While machine learning (ML) and data analytics have generated individualized treatment recommendations for improving outcomes, “these works focused on making broad [largely drug class level] treatment recommendations independently of specific drug and dose considerations... [whereas] guidelines and landmark trials highlight important drug- and dose-dependent variations in treatment efficacy, safety, and risk profiles" [2]. In short, personalized medicine should account for contextual variations in seeking more effective, cost-efficient treatments with better outcomes. This study presents an approach to clinical support to time- and resource-constrained practitioners using a generative artificial intelligence (GenAI) approach to treat a serious medical condition afflicting young mothers and their children with increasing alacrity. Such a need is particularly acute in the socioeconomically disadvantaged regions of the world.

Gestational diabetes mellitus (GDM) is a significant global health concern affecting many pregnancies [3]. Defined as glucose metabolism imbalance first detected during pregnancy, the International Association of Diabetes in Pregnancy Study Group reports that "GDM is not only related to perinatal morbidity but also to an increased risk of diabetes and cardiovascular disease in the mother in later life, and childhood obesity in the offspring" [4]. The pooled global prevalence was 14% in 2021, with the highest occurrence in the Middle East - North Africa (27.6%), Southeast Asia (20.8%), and among high-income countries (14.2%) [5]. There is considerable agreement among medical practitioners that the development of GDM could be influenced by various risk factors, including maternal age, obesity, family history of diabetes, previous occurrences of GDM, and specific ethnic backgrounds [6,7]. This is illustrated in Figure 1 (data sources: [3,8-10]) as the medical characterizations of GDM comprising factors such as diagnosis, risks, prediction, management, complications, and postpartum care.

Figure 1. Medical characterization of gestational diabetes mellitus. GDM: gestational diabetes mellitus.



Also, of concern to the WHO is that GDM leads to various complications for both affected mothers and their offspring, such as increased risks of cesarean delivery, pre-eclampsia, and type 2 diabetes (T2D) for mothers. Children are at higher risk of macrosomia, hypoglycemia, respiratory distress syndrome, and an increased likelihood of developing obesity and T2D later in life [11]. The long-term health risks include elevated chances of developing T2D and cardiovascular diseases for both mother and child [12]. In the Global South and developing countries [8,13], GDM presents significant challenges due to:

1. Higher prevalence rates in certain regions, particularly South Asia and the Middle East.
2. Limited health care resources for screening, diagnosis, and management.
3. Genetic factors in certain ethnic groups increase GDM risk.
4. Rapid urbanization and lifestyle changes leading to increased obesity rates.
5. Potential underdiagnoses due to lack of routine screening.

Effective GDM treatment requires multiple diagnostic tests, including oral glucose tolerance tests, random plasma glucose tests, and fasting plasma glucose tests. The treatment options include regular blood glucose monitoring, dietary modifications, lifestyle changes, and, when necessary, pharmacological interventions such as insulin or oral hypoglycemic agents [9]. The recent advancements in AI-driven tools, such as the AI Drug Mix and Dose Advisor developed for T2D [2], have shown potential in optimizing pharmacological interventions by customizing drug and dose recommendations to individual patient profiles. Similar approaches could be valuable in improving glycemic management in GDM cases, enhancing personalized care in postpartum treatment, drug discovery with therapy, and reducing long-term risks of developing chronic diseases in general.

Despite growing interest in AI-driven clinical support, current models often struggle to integrate diverse, multisource medical data into actionable insights, especially in conditions such as GDM, where missing information and diagnostic delays contribute to less desirable outcomes. These limitations are particularly prominent in resource-constrained settings, where systemic challenges, such as insufficient screening tools, lack of standardized care protocols, and limited provider training, complicate effective diagnosis and treatment [8,13]. As a result, the timely and effective treatment of GDM remains difficult, further endangering maternal and fetal health.

In such contexts, the unavailability of specialized professionals, economic constraints, and cultural challenges also influence treatment adherence and engagement [14,15]. The limited awareness between both the public and health care providers continues to contribute to improper management of GDM [16], reinforcing the urgent need for robust, context-sensitive clinical decision support [17,18].

To address these gaps, we propose a novel solution using specialized GenAI techniques for GDM management. Specifically, we develop a proof-of-concept (PoC) of a clinical support system that uses a knowledge graph (KG) supporting a local large language model (LLM). This system extracts and integrates intervention strategies from peer-reviewed research to support physicians in making contextually relevant treatment decisions.

Standalone local LLMs, however, face known limitations, including hallucinations and reduced reliability when handling domain-specific, complex queries [19]. To address these issues, we introduce a retrieval-augmented generation (RAG) mechanism that improves the accuracy and relevance of outputs by supplementing the LLM with contextual data [20,21]. This hybrid approach could elevate the clinical utility of GenAI for complex, low-resource health care scenarios such as GDM.

By generating structured, evidence-informed recommendations in real time, our system lays the foundation for scalable and explainable AI support tools customized to maternal health. The following section reviews previous ML and LLM-based approaches to GDM detection and prediction, positioning our work within this evolving research landscape. It is stated at this juncture that while the distinction between LLMs and local LLMs is clear, it is less so between local LLMs and small language models (SLMs). The prototype developed in this study assumed a local LLM architecture but could be repurposed as SLMs, particularly in resource-constrained locations of the Global South. A concise feature comparison of LLMs, local LLMs, and SLMs is provided in Textbox 1.

Textbox 1. Feature comparison of large language models, local large language models, and small language models.

Large language models

Large language models (LLMs) are typically based on deep learning, trained on massive amounts of text and increasingly multimedia data to understand, generate, and manipulate human language. LLMs work by learning to predict the next word in a sequence based on the context of the input prompt, using billions of parameters to refine these predictions. They excel at natural language processing tasks such as text completion, translation, summarization, question-answering, and content generation.

Local LLMs

Local LLMs run inside the private data center of an entity or organization. Local LLMs are fine-tuned with the organization's data (eg, patient records or standard rules) and can provide specific context to a query or prompt that general-purpose chatbots cannot or should be legally allowed to deliver. Particularly in the domains of sensitive and confidential data (such as a patient's medical conditions), such prompts may have to be subject to rigorous access, authentication, and accounting controls.

Small language model

A small language model is designed to understand and generate natural language, similar to LLMs, but on a much smaller scale, with fewer parameters and a simpler architecture. Small language models are optimized for efficiency and can be deployed on resource-constrained devices like smartphones or local servers, offering benefits such as faster training and execution, lower energy consumption, and improved privacy by allowing for on-device processing and less reliance on cloud connectivity. A use case could be first responders in emergency room situations.

Recent advances in ML have shown promise in improving the early diagnosis and personalized management of chronic conditions such as GDM. These models identify high-risk individuals during pregnancy, customize treatment plans, and ultimately enhance maternal and neonatal health outcomes. Several studies have developed ML algorithms that account for demographic variations, for example [22,23], present models customized to Asian women [10] used decision trees and ensemble learning for early GDM detection, reporting high sensitivity and specificity. However, these models often fail to capture the full complexity of GDM-related factors.

The efforts to improve model interpretability include research, such as meta-reviews of clinical studies on complications during pregnancy and their treatments [24], on clinically explainable ML approaches for blood glucose monitoring [25,26], and the use of extreme gradient boosting to identify key risk factors [27]. However, several studies [25,26,28,29] note limitations in integrating high-quality datasets, supporting real-time interventions, or embedding models within clinical systems. Table 1 presents these representative models, underscoring the trade-offs between accuracy, interpretability, and practical usability.

Table . Representative research deep learning or machine learning models for predicting gestational diabetes mellitus.

Study	Year	Model	Key contributions and limitations
Kokori et al [22] and Kumar et al [23]	2024	Demographic-specific ML[a] model	• KCs[b]: Accurate predictions for specific demographics (Asian women). • Limits: Limited integration into health care systems.
Kurt et al [10]	2023	Decision trees and ensemble	• KCs: High sensitivity and specificity. • Limits: Fails to capture all GDM[c]-related factors.
Wu et al [29]	2024	Clinically interpretable ML	• KCs: Emphasized interpretable models for GDM. • Limits: Limited real-time application.
Wu et al [25]	2022	ML-based models	• KCs: Importance of high-quality datasets. • Limits: Lacks interpretability and integration.

[a]ML: machine learning.

[b]KC: key contribution.

[c]GDM: gestational diabetes mellitus.

These limitations highlight the need for models that go beyond static risk prediction to support context-aware clinical decision-making. In this regard, LLMs offer transformative potential as they generate patient-specific recommendations by synthesizing heterogeneous clinical data. When augmented with retrieval techniques, such models become more effective.

Several recent studies have discussed the expanding role of LLMs across health care domains[30]. For example, an AI system developed for liver diseases [31] provided personalized treatment strategies that improved diagnostic outcomes. Graph-based retrieval-augmented generation (GraphRAG) integration has shown benefits in nephrology by increasing output precision and reliability [20], while LLMs have supported psychotherapy automation [32] and administrative workload reduction in personalized medicine [33]. Some of these use cases are captured in Table 2, reinforcing the applicability of RAG-augmented LLMs in clinical practice.

Table . Representative use cases of artificial intelligence in clinical health care.

Study	Year	Model	Key contributions
Ge et al [31]	2024	AI[a] model for liver diseases	Enhanced diagnostic accuracy and patient management tailored for liver diseases.
Ong et al [34]	2023	Clinical decision support system	Improved clinical decision-making with RAG[b]-enhanced LLMs[c], offering precise predictions and treatments.
Miao et al [20]	2024	LLM-RAG for nephrology	Improved accuracy and reliability in nephrology advice by integrating RAG with LLMs.
Stade et al [32]	2024	LLMs in psychotherapy	Explored the potential of LLMs to support and potentially automate aspects of psychotherapy.
Tripathi et al [33]	2024	Personalized medicine AI model	Demonstrated how LLMs can automate administrative tasks, reducing clinicians' workload from electronic medical records.

[a]AI: artificial intelligence.

[b]RAG: retrieval-augmented generation.

[c]LLM: large language model.

Noting the above, this paper proposes a novel architecture for GDM care that integrates (1) a *local LLM* for domain-specific control and privacy, (2) an *RAG* engine for contextual grounding, and (3) a *domain-specific KG* to capture interrelated medical evidence.

This combination enables real-time generation of explainable, evidence-informed treatment recommendations for GDM management, even in resource-constrained settings. As compared with previous studies, such as those by Nambiar et al [2] and Tripathi et al [33], which focused on general dosing automation or task simplification, this study addresses a critical gap: the need for adaptive, fine-grained, and explainable intervention support in the prenatal context.

From a technical standpoint, our contributions are (1) the construction of a GDM-specific KG derived from peer-reviewed literature; (2) the use of RAG-enhanced local-LLMs to retrieve, contextualize, and generate targeted care pathways; and (3) a PoC system architecture that is interpretable, domain-grounded, and designed for offline, privacy-preserving environments.

The PoC will support timely intervention and align with the practical realities of underserved clinical contexts; consider the plight of a rural doctor in the Global South, where internet connectivity, specialist clinician availability, and cutting-edge expertise may be limited. It represents a step toward deploying technically robust and clinically meaningful AI to applications of acute need.

Following this introduction, the remainder of this paper is organized as follows. The next section addresses the methods, and specifically, a description of developing design artifacts for a PoC. In the Results section, we put the system through simulated scenarios and test the responses for accuracy, bias, and performance benchmarking. In the Discussion section, we present the principal findings along with an analysis of key contributions of the research. The paper ends with a section on Conclusions, which also covers limitations and suggestions for further research.

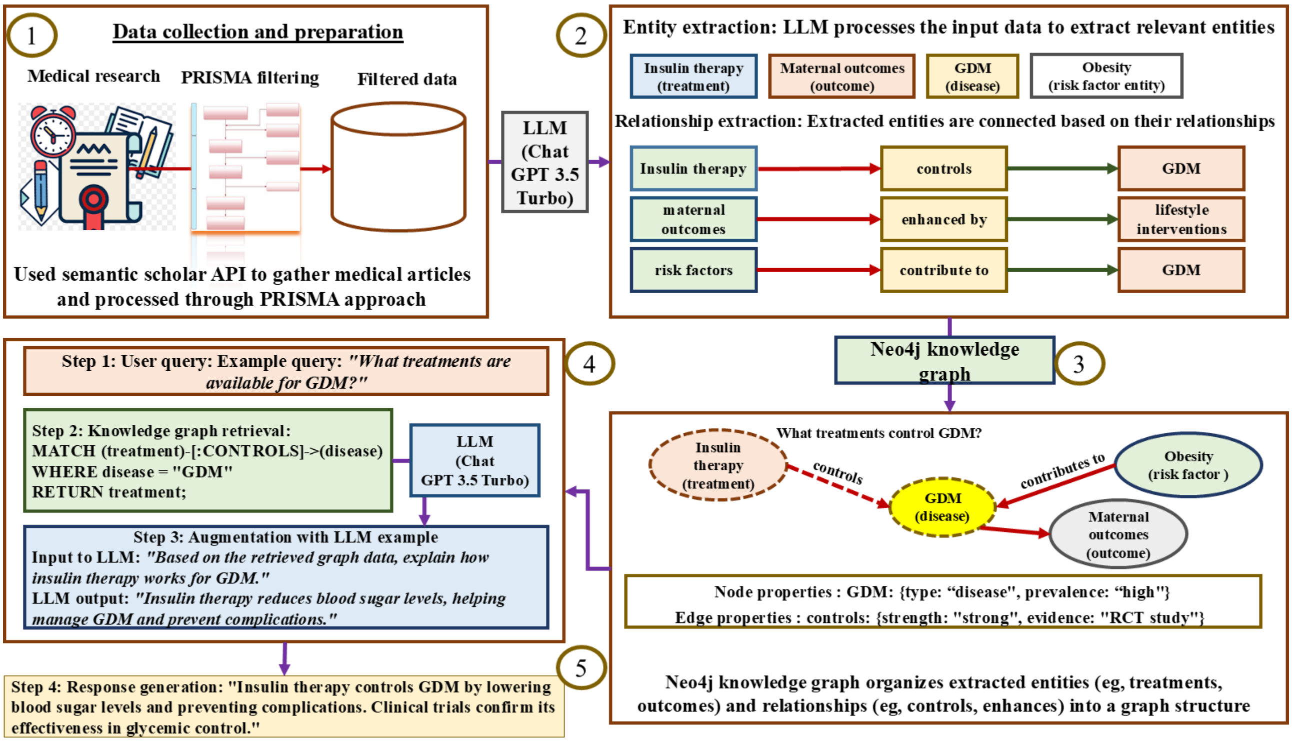
Methods

Prototyping a PoC

Health care professionals, particularly those in densely populated and resource-constrained regions of the Global South, often face significant challenges in accessing timely, evidence-based medical insights. Attending training sessions or reviewing vast volumes of literature under time pressure is impractical, especially in scenarios where specialist expertise or standardized guidelines are lacking. Our approach uses computational methods to extract, structure, and contextualize medical knowledge using GenAI and KG technologies to address this need.

Our primary objective was to develop a PoC of a clinical AI assistant that would support the management of GDM. This GraphRAG-based architecture combines entity extraction from published research, KG construction, and RAG to generate clinically grounded, context-aware responses. As illustrated in Figure 2, the PoC framework follows a 5-stage pipeline.

Figure 2. Process flow of the proposed graph-based retrieval-augmented generation approach, showing data collection, entity extraction, knowledge graph construction, and retrieval-augmented generation for AI-assisted clinical support for gestational diabetes mellitus. API: application programming interface; GDM: gestational diabetes mellitus; LLM: large language model; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



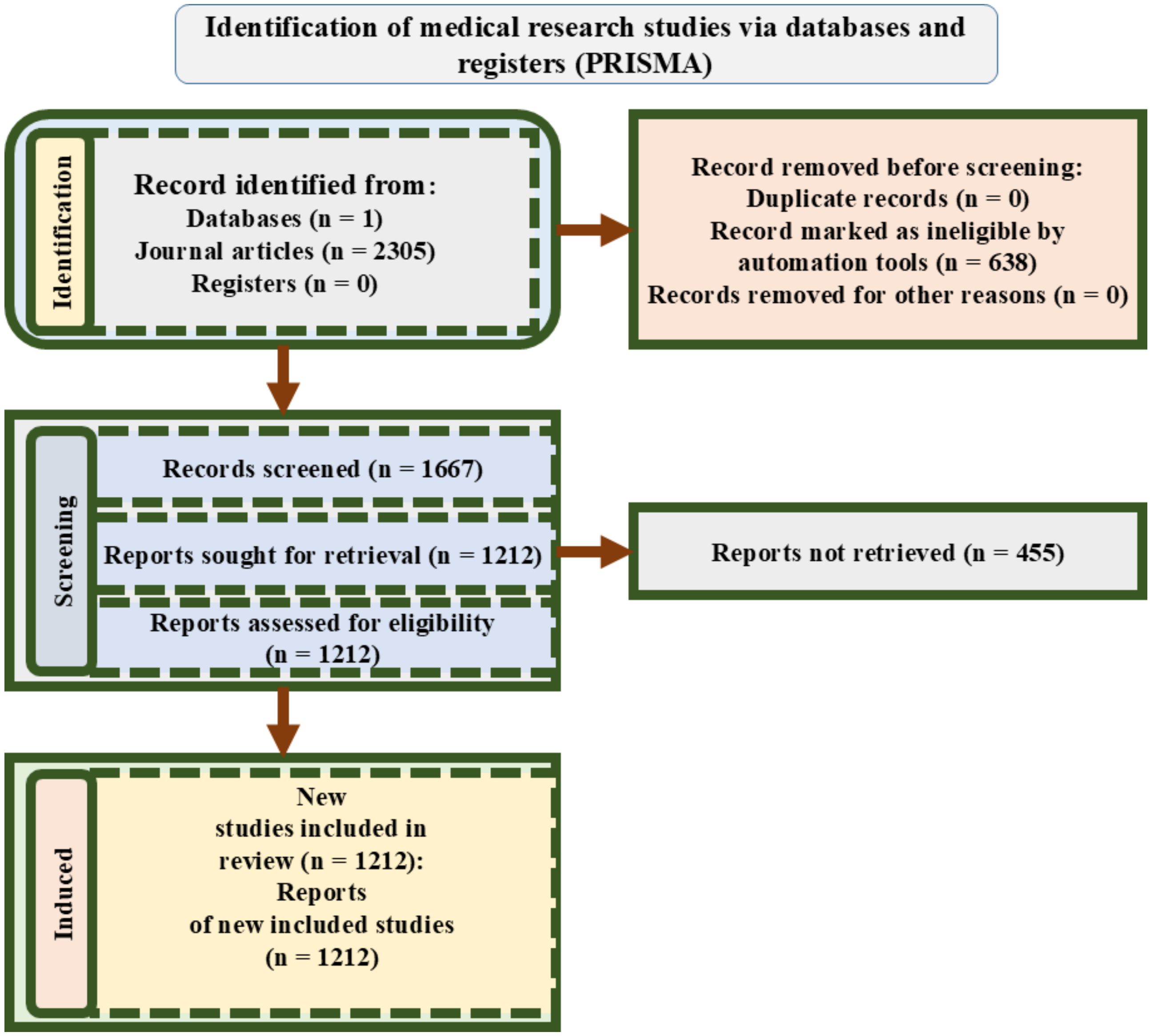
- Data collection and preparation: The Semantic Scholar API retrieved relevant research articles on GDM interventions. A PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)–guided filtering process was applied to ensure that inclusion criteria were met, resulting in a refined corpus of 1212 high-quality articles.
- Entity extraction: Using GPT-3.5 Turbo (OpenAI) and few-shot prompting, entities such as treatments, outcomes, risk factors, and disease indicators were extracted from full-text articles. Semantic consolidation (eg, grouping “low-carb diet” and “reduced carbohydrate intake”) ensured terminological consistency.
- KG construction: Extracted entities and their relationships were encoded into a Neo4j graph database. The graph allowed efficient traversal of clinical pathways, such as connecting interventions to outcomes and risk profiles. Each node and edge pair was annotated with medical metadata, such as intervention strength, evidence level, or prevalence.
- Query processing and graph retrieval: When a user query is submitted (eg, “What treatments control GDM?”), the system was designed to retrieve relevant subgraphs using Cypher queries. These results are then passed to the LLM for augmentation and contextual response generation by incorporating patient records.
- Response generation: The final output is a clinically coherent and relevant response integrating retrieved evidence and a generative explanation. For example, based on retrieved data, the model might respond: “Insulin therapy controls GDM by lowering blood sugar levels and preventing complications.” If asked why, the system might explain: “Insulin enables glucose uptake by cells throughout the body, particularly muscle and fat cells, by facilitating glucose transport across cell membranes. Without adequate insulin, glucose accumulates in the bloodstream while cells are starved of this essential energy source.”

This multistep process would allow the system to access reputable and current medical research to produce explainable, evidence-grounded outputs for clinical decision support. Each component of this workflow is further detailed in the following subsections.

Data Collection

To develop a high-quality domain-specific KG for GDM, we conducted a systematic search using the Semantic Scholar API [35], a widely used biomedical research platform. The query term “gestational diabetes interventions” was selected to target studies focused on treatment strategies and clinical outcomes. The search was restricted to articles published between January 2000 and May 2024, to cover both foundational and contemporary research. The data collection and filtering process adopted PRISMA guidelines, as illustrated in Figure 3.

Figure 3. PRISMA flow diagram showing the systematic data collection and filtering process, detailing identification, screening, eligibility assessment, and inclusion of research articles for knowledge graph construction.



- Identification: The initial search produced 2305 journal articles. No records were found from registers. Automated filters removed 638 ineligible records based on metadata mismatches or irrelevant domains. No duplicate entries were detected.
- Screening: The remaining 1667 articles were screened by 2 reviewers (FR and SB) based on titles and abstracts. This stage ensured that only articles related to GDM diagnosis, treatment, management, or intervention outcomes were retained.
- Eligibility: A total of 1212 full-text articles were deemed eligible based on the inclusion criteria. Articles were excluded at this stage (n=455) due to full-text unavailability, access limitations, or insufficient clinical relevance.
- Inclusion: The final corpus consisted of 1212 peer-reviewed studies, all of which were used to extract entities and construct the GDM-focused KG.

While Semantic Scholar provided comprehensive coverage and metadata-rich access, reliance on a single source introduces potential limitations, such as limited representation of non-English or region-specific research and sensitivity to keyword variations. Future work could explore multilingual database integration and broader query strategies to reduce potential selection bias.

Nonetheless, for developing our PoC, the selected dataset offered sufficient diversity and clinical validity to enable meaningful experimentation and system development.

Entity Extraction

Following the curation of the GDM research corpus, the next step involved extracting clinically relevant concepts, including treatments, risk factors, and outcomes, from the published research. This process was executed using OpenAI's GPT-3.5 Turbo 16K API [36], which supported advanced natural language processing for domain-specific knowledge extraction. Rather than relying on pretrained biomedical ontologies, we adopted a lightweight prompting-based approach aligned with our PoC's experimental and modular goals.

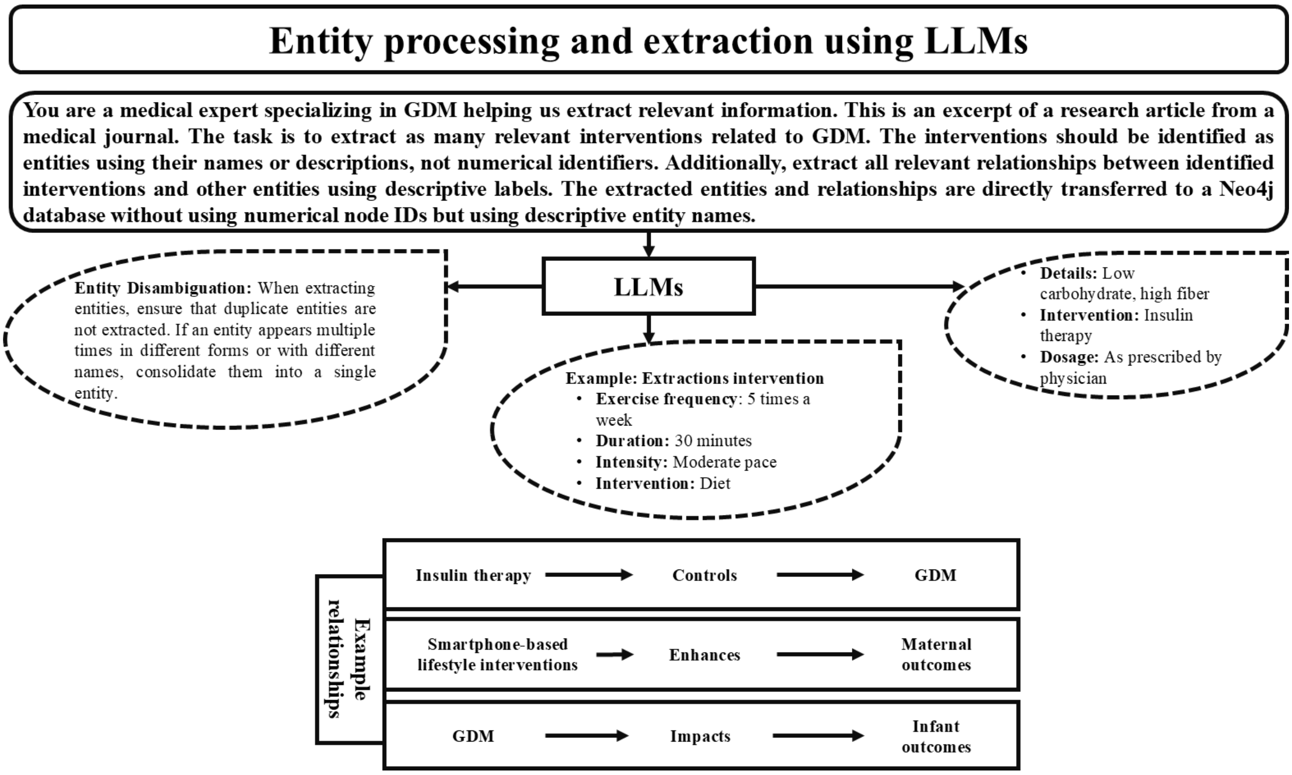
A few-shot prompting strategy was applied to guide the language model in identifying and structuring entities of interest in a usable format. Guided by 3 medical doctors, the prompts were manually engineered to show expected outputs, such as intervention types (eg, insulin therapy, diet, and physical activity), intervention parameters (eg, frequency, duration, and dosage), and associated maternal and infant outcomes. This enabled the model to consolidate synonymous or semantically related expressions (such as "low carbohydrate diet" and "reduced carb intake") into a unified entity representation. The same prompts also encouraged disambiguation of overlapping terms and discouraged the duplication of entities across articles.

The outputs were parsed into structured formats, which included both individual entities and the semantic relationships among them, for example, linking "insulin therapy" as a treatment that "controls" GDM, or connecting "smartphone-based lifestyle interventions" to enhanced "maternal outcomes." These entities and their connections were then directly integrated into the KG in the next stage of development.

This stage of entity extraction was led by the coauthor (FR), who specializes in bioinformatics and uses a technique we describe as "medical prompt engineering." The objective was to simulate how future clinical AI assistants might extract structured knowledge from unstructured medical literature autonomously. However, we acknowledge that such extractions would require validation by specialist health care professionals to ensure accuracy and reliability for clinical deployment.

The overall entity extraction workflow, including prompt design, model guidance, semantic structuring, and preparation for graph integration, is visualized in Figure 4.

Figure 4. Entity extraction workflow using large language models. The diagram is an example of the process for extracting interventions, risk factors, and relationships, which produces structured and context-aware knowledge representation for gestational diabetes mellitus management. GDM: gestational diabetes mellitus; LLM: large language model.



Construction of the KG

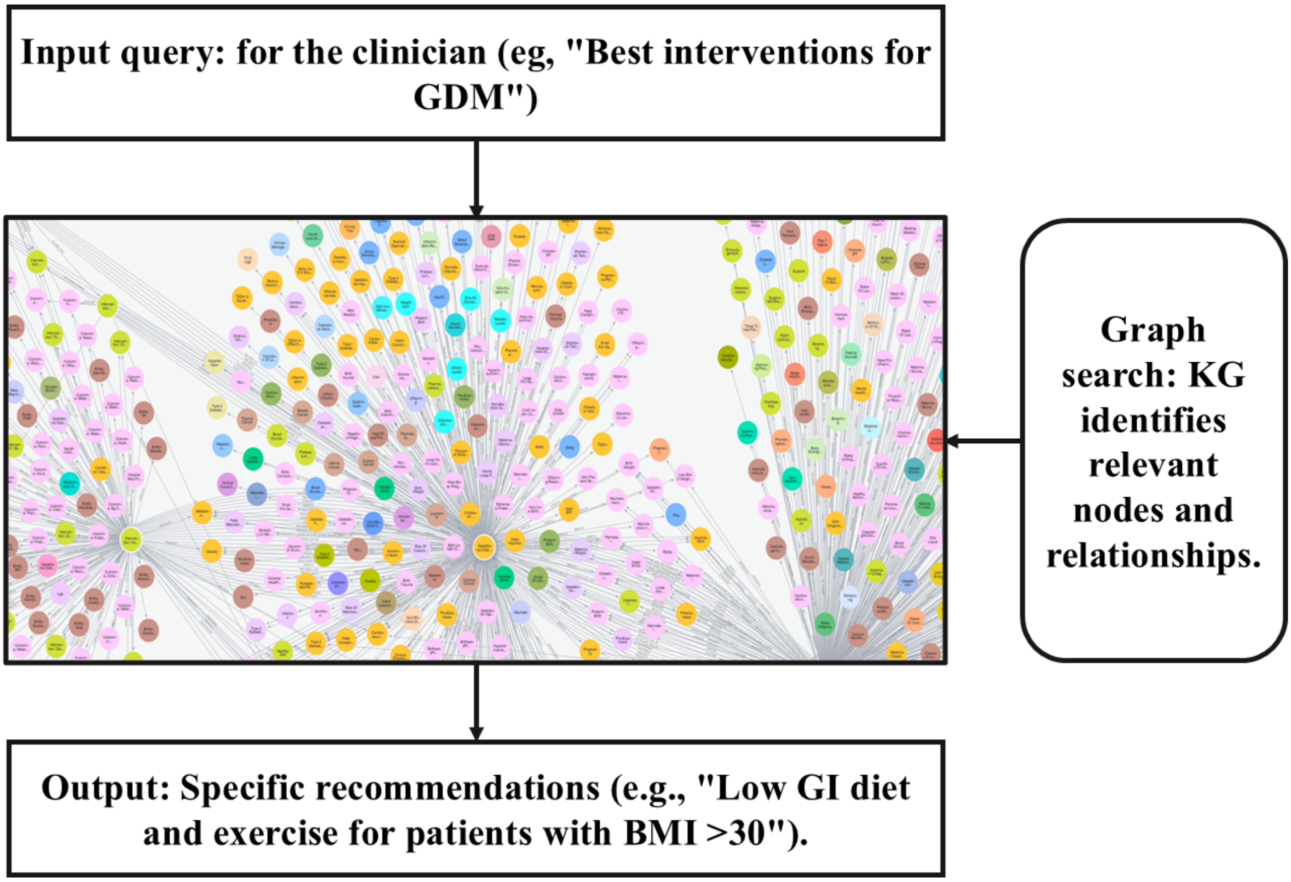
Upon completion of the entity and relationship extraction, the structured data were integrated into a KG using Neo4j, a widely used open-source graph database [37]. Neo4j is optimized for representing interconnected biomedical data, making it well-suited for capturing the multifactorial nature of GDM management, which involves dynamic relationships between interventions, risk factors, outcomes, and complications [38].

The KG construction process involved linking each extracted entity, such as insulin therapy, dietary strategies, or risk factors like obesity, to its semantically relevant mappings using directional edges labeled with relationship types (eg, "controls," "contributes to," and "enhances"). Each node was annotated with descriptive labels and properties derived from the literature, and relationships were encoded with metadata such as source references or study types, when available.

All nodes and edges were imported into Neo4j through a structured ingestion pipeline, enabling clinicians or researchers to query the KG using the Cypher query language. This functionality allowed for exploratory clinical queries, such as identifying interventions most frequently associated with improved maternal outcomes in high-risk GDM cases or tracing evidence paths for specific treatment combinations.

The resulting KG facilitated context-aware clinical decision support by surfacing specific evidence-informed insights. For example, a clinician's query, such as "What are the best interventions for GDM in patients with a BMI over 30?" could retrieve targeted graph segments linking relevant interventions (eg, low glycemic index diet and structured exercise regimens) to outcomes validated in the literature. This dynamic capability is depicted in Figure 5, which illustrates a representative graph traversal initiated by a clinician's question, leading to personalized treatment recommendations based on the structural relationships captured in the KG.

Figure 5. Knowledge graph–powered clinical support system for gestational diabetes mellitus. The graph-based search retrieves relevant interventions and relationships, giving treatment recommendations. GDM: gestational diabetes mellitus; KG: knowledge graph.



The KG serves as the core reasoning backbone of the prototype clinical assistant, consolidating distributed medical evidence into a queryable visual knowledge substrate that can be updated as new medical evidence emerges.

KG-Based RAG

To enhance the clinical utility of the constructed KG, we then implemented an RAG approach [39]. This hybrid architecture combines traditional retrieval mechanisms with generative LLMs to produce contextually grounded and medically sound responses. In clinical settings, where decision-making depends on subtle interpretation and evidence-based insights, this integration mitigates the limitations of standalone generative systems like SLMs.

While LLMs, such as ChatGPT (OpenAI), can produce fluent and context-aware responses, they are prone to hallucinations, outdated knowledge, and domain-specific inaccuracies [19]. Conversely, RAG addresses these gaps by coupling LLMs with reputable (peer-reviewed) external knowledge sources. For example, no medical claim, such as bleach being a valid treatment for COVID-19, would have gone into the KG. In our PoC, entity-aware retrieval from the Neo4j-based GDM KG provides factual context, which the LLM then uses to generate a tailored response. This integration significantly improves factual grounding and interpretability, essential in critical domains, such as maternal health [20,21].

Using the PoC follows a 5-stage pipeline, visualized in Figures 6 and 7. Beginning with an initial clinical query, the system encodes the user input and dynamically retrieves semantically matched information from the KG. This process accounts for risk factors, interventions, and patient-specific context, including medical records and socioeconomic profiles, thereby aligning output with real-world variability in treatment planning.

Figure 6. End-to-end process flow of the graph-based retrieval-augmented generation solution. The pipeline processes medical literature and patient data, integrating them into a structured knowledge graph for AI-driven clinical decision support. GDM: gestational diabetes mellitus; KG: knowledge graph; LLM: large language model.

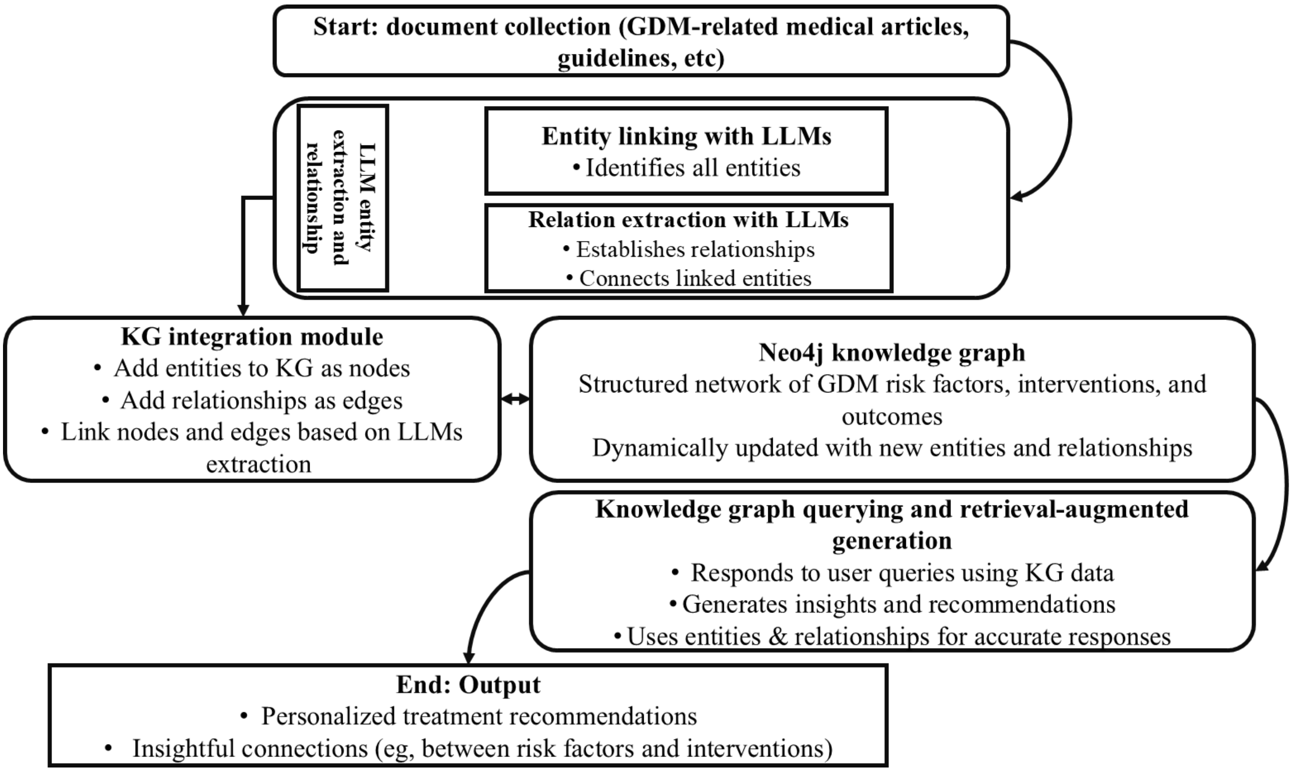


Figure 7. Structured retrieval and response generation process in graph-based retrieval-augmented generation. The diagram shows how clinician queries interact with medical knowledge sources, pattern matching, and graph-based retrieval to enhance artificial intelligence–generated responses. GDM: gestational diabetes mellitus; LLM: large language model.

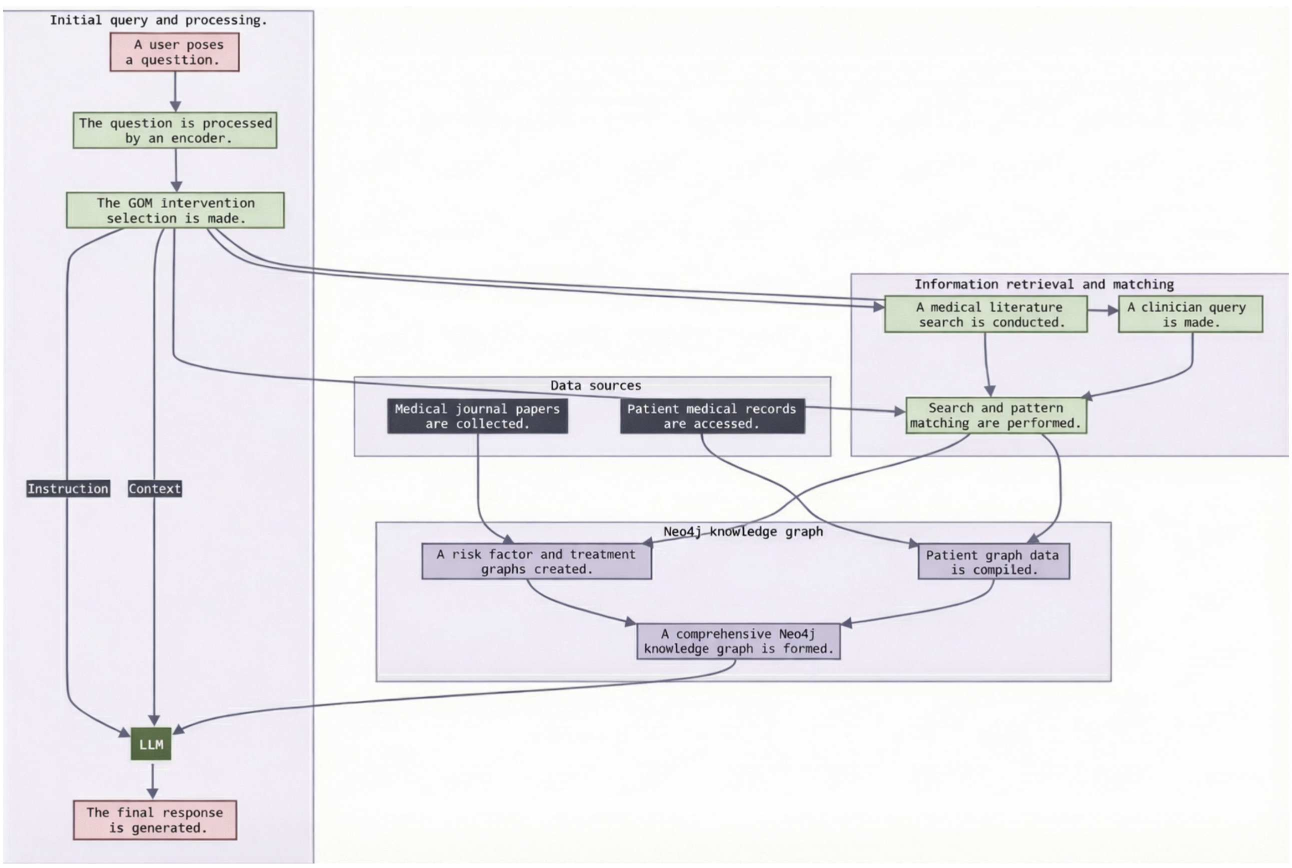


Figure 6 shows the underlying LLM-KG pipeline, including entity extraction, relationship linking, and graph query generation. Figure 7 offers a complementary perspective by emphasizing end-to-end data flow, from patient query and literature matching to LLM response generation, thus highlighting how both structured (graph-based) and unstructured (textual) data are integrated to yield context-aware, personalized responses.

Although Figures 6 and 7 present a simplified overview of system functionality, the development process required iterative prompt engineering, guided tuning, and manual validation to align LLM outputs with the domain-specific vocabulary and relationships obtained from GDM research literature [32,33]. This iterative refinement helped ensure that the GraphRAG PoC consistently produces clinically meaningful recommendations rooted in the KG, avoiding spurious correlations and unverified claims.

Evaluation Framework and Metrics

The evaluation of the GraphRAG-powered local LLM for GDM was conducted through a structured framework designed to assess both technical performance and clinical relevance. Applications of AI in health care require rigorous validation beyond prompt engineering. This study used a multidimensional evaluation process using a combination of quantitative metrics and clinician-generated prompts.

Evaluation Objectives

The primary objective of the evaluation was to measure the effectiveness of the proposed PoC in three "fit for purpose" criteria: (1) generating clinically relevant**,** context-aware responses to queries on GDM management; (2) comparing its performance against widely used open-source LLMs in terms of accuracy and interpretability; and (3) assessing whether the retrieval-augmented approach of GraphRAG significantly improves response quality in medical decision support. These criteria reflect the critical nature of clinical decision-making, where AI-generated content's clarity, accuracy, and contextual relevance directly affect patient safety and clinical outcomes.

Testing Environment

The evaluation was conducted in a simulated environment, without the involvement of live patients or human participants. The GraphRAG-powered local LLM was deployed on an offline computing environment, ensuring that no external API calls or third-party cloud services influenced the test outcomes. The KG was prepopulated with medical research articles, as described in the "Prototyping a PoC" section, and served as the contextual knowledge base for all retrieval-augmented queries.

Prompt Design and Benchmark Models

The prompts used in the evaluation were carefully crafted to simulate realistic clinical and layperson queries. These prompts were generated from two user groups: (1) *laypersons* represented by 5 contributors (the authors) simulating patient queries, verified for clarity and simplicity; and (2) *clinicians* comprising 2 general practitioners (GPs) and 1 specialist physician, who created queries based on typical clinical decision-making scenarios.

Furthermore, 2 independent medical practitioners reviewed all prompts to ensure clinical relevance (were the prompts aligned with real-world GDM management scenarios?) and content clarity (did the prompts avoid ambiguous phrasing or unrealistic edge cases?)

The GraphRAG system was then benchmarked against 3 open-source LLMs commonly used in medical AI research. The comparison is intended to analyze the performance of a domain-augmented local model (our PoC) against both general-purpose and specialized health care LLMs.

- ChatGPT [36]: A versatile, general-purpose LLM.
- Claude [40]: Known for generating coherent, contextually rich responses.
- BioMistral [41]: A domain-specific medical LLM optimized for health care contexts.

Our benchmarking compares the GraphRAG-enabled local LLM against the above 3 LLM models to assess clinical relevance, contextual accuracy, and terminological consistency. These models were selected based on availability, health care domain relevance, and ease of integration into our evaluation pipeline. While we acknowledge the increasing prevalence of open-source LLMs such as LLaMA 3 (Meta AI), due to hardware compatibility constraints and inference framework differences at the time of testing, we could not integrate LLaMA 3 within the test environment. LLaMA 3 and other emerging open-source models, such as Mistral 7B (Mistral AI) and Phi-3 (Microsoft), should be included in future benchmarking updates to expand our comparative analysis, which is suggested as future work.

Evaluation Metrics and Rationale

Following established practices in evaluating health care AI models [42,43], we used 5 complementary metrics, each addressing a distinct dimension of AI-generated response quality. These are presented in Table 3.

Table . Metrics and their clinical significance in evaluating artificial intelligence–generated responses.

Metric	Purpose	Significance
Relevance score	Measures alignment between response content and user query.	Critical for clinical decision support, where irrelevant or off-topic answers compromise safety.
BLEU[a] score	Evaluates syntactic similarity and phrase structure match against reference answers.	Ensures AI[b] responses replicate validated medical language without distortion.
Jaccard similarity	Quantifies overlap in key medical terms between model response and reference.	Captures preservation of clinical terminology essential in GDM[c] management.
BERTScore	Assesses semantic similarity using deep contextual embedding.	Evaluates whether model responses capture the intended clinical meaning beyond surface text.
METEOR	Evaluate fluency and coherence in response generation.	Ensures clarity and interpretability for both clinicians and patients.

[a]BLEU: bilingual evaluation understudy.

[b]AI: artificial intelligence.

[c]GDM: gestational diabetes mellitus.

Together, these metrics comprehensively address the precision, contextual relevance, and interpretability of an AI model's outputs, which are key requirements for clinical use cases.

Evaluation Process

The evaluation adopted the following steps:

First, each LLM, including GraphRAG, was presented with the same curated set of 20 prompts (10 from simulated layperson queries and 10 from clinicians), covering core aspects of GDM management, such as risk factors, diagnostics, treatment, and complications. The 5 coauthors (EE, FR, SB, AN, and RS) jointly drafted the layperson prompts, while clinical prompts were contributed by 2 practicing GPs and reviewed by a third medical specialist.

Second, the system's responses were compared against reference answers, curated from clinical guidelines and expert consensus statements.

Third, evaluation was conducted in a zero-shot retrieval-augmented setting. No supervised training or fine-tuning was performed. The local LLM operated on a preconstructed KG as the contextual grounding source.

Fourth, automated evaluation metrics (bilingual evaluation understudy [BLEU], Jaccard Similarity, BERTScore, and METEOR) were computed using standard natural language processing evaluation libraries. These scores reflect surface-level accuracy, overlap in medical terminology, and semantic similarity.

Fifth, manual relevance scores were assigned by 2 independent medical reviewers on a 1‐5 scale, based on clinical applicability, specificity, and usefulness of responses.

Finally, results were averaged across all prompts and models and reported for comparative analysis in the Results section. While performance scores are high (eg, BLEU=0.99 approximately), this reflects a small, curated test set and should not be considered generalizable. CIs and interrater agreement were not calculated in this phase of the research.

Benchmarking Scope and Qualifications

The evaluation was designed to show the technical feasibility and domain relevance of the GraphRAG framework, rather than to establish clinical deployment readiness for deployment. Consequently, the following qualifications would apply:

First, all responses were evaluated in a simulated, offline environment without involvement of human patients, real-time electronic health record data, or live clinical workflows.

Second, no supervised training or dataset splitting was involved, as the system uses RAG rather than end-to-end training. All prompts were presented statically to each LLM.

Third, as recorded in our research logs, the KG was constructed from a curated corpus of 1212 peer-reviewed, English-language articles on GDM interventions, extracted via Semantic Scholar API (2000‐2024). The KG contains approximately 2750 nodes, 5800 edges, and 18 entity types, including risk factors, therapies, dietary interventions, and outcomes.

Fourth, the evaluation prompt set, while medically validated, remains small and nonrandomized. No demographic stratification, multilingual testing, or subgroup fairness analysis was performed.

Fifth, performance metrics assessed linguistic and contextual quality only. There has been no empirical validation of clinical efficacy, patient safety, or decision-making utility.

Finally, future iterations should expand prompt diversity, compute interrater reliability scores, and explore prompt-based fairness auditing. Prospective clinical trials and feedback-integrated deployment pipelines are also planned.

Ethical Considerations

This study involved the development and technical validation of a PoC clinical AI assistant for GDM management. The research was conducted entirely in a simulated environment without involving human participants, personal health data, or clinical interventions. Accordingly, formal ethics board approval was not required for this PoC phase of the research study.

More specifically, this was in accordance with ethical research standards for early-stage AI system development in health care. We ensured that no human participants, no personal health data, and no real-time clinical interventions resulted from this PoC phase. While fairness across subpopulations was not evaluated in this version, future efforts shall explicitly address this dimension.

Data Source Transparency

The data used in this study were drawn exclusively from public-domain, reputable academic research, collected through the Semantic Scholar API. All articles retrieved were from peer-reviewed scientific publications, ensuring no private, sensitive, or patient-level data were accessed or processed. The use of publicly available literature aligns with ethical practices in computational biomedical research, where datasets are preferably in the public domain.

Simulated Testing Environment

The PoC was evaluated using simulated prompts designed by the research team and reviewed by independent clinicians. No real patient interactions, medical records, or clinical environments were involved in the testing. This approach was explicitly chosen to focus on the feasibility of the proposed GraphRAG-powered knowledge retrieval and response generation approach.

All comparisons against open-source LLMs (ChatGPT, Claude, and BioMistral) were also conducted offline, with no data sent to external servers during evaluation, ensuring data security and compliance with our concern that we do not train such models with our research data.

Responsible AI Development

The design and development of the GraphRAG framework adhered to ethical AI principles, emphasizing:

1. Transparency: Clear explanation of methods and evaluation.
2. Safety: Avoidance of deploying untested AI systems in live clinical environments.
3. Explainability: Use of a KG for contextual reasoning and improved interpretability.
4. Bias awareness: Although no patient data were used, future iterations will integrate fairness auditing to minimize algorithmic bias.

Fairness and Demographic Representation

The development of the PoC used a small set of curated prompts authored by the research team and clinicians. Hence, no demographic, linguistic, or regional diversity was represented in the evaluation. This limitation may impact the generalizability of the system’s recommendations across patient populations. Future prototyping iterations will integrate fairness-aware evaluations**,** including prompt diversity across age, gender, geography, and language, to improve equitable performance across clinical contexts.

Results

System Demonstration Scenarios

The PoC beta testing in a simulated environment highlighted the feasibility of the GraphRAG-powered clinical support system for GDM management. The PoC generated personalized, clinically relevant responses to GDM-related queries, simulating interactions between patients, health care professionals, and the system.

Figures 8 and 9 present an illustrative scenario displaying how the GraphRAG local LLM could support clinical consultations. In this example, a patient presents a question regarding the top risk factors for GDM. A health care professional, such as a GP or maternity nurse, uses the GraphRAG-enabled clinical support system to process the query into a prompt.

Figure 8. GraphRAG-based clinical support system for gestational diabetes mellitus - iconographic representation. AI: artificial intelligence; GDM: gestational diabetes mellitus; genAI: generative artificial intelligence.

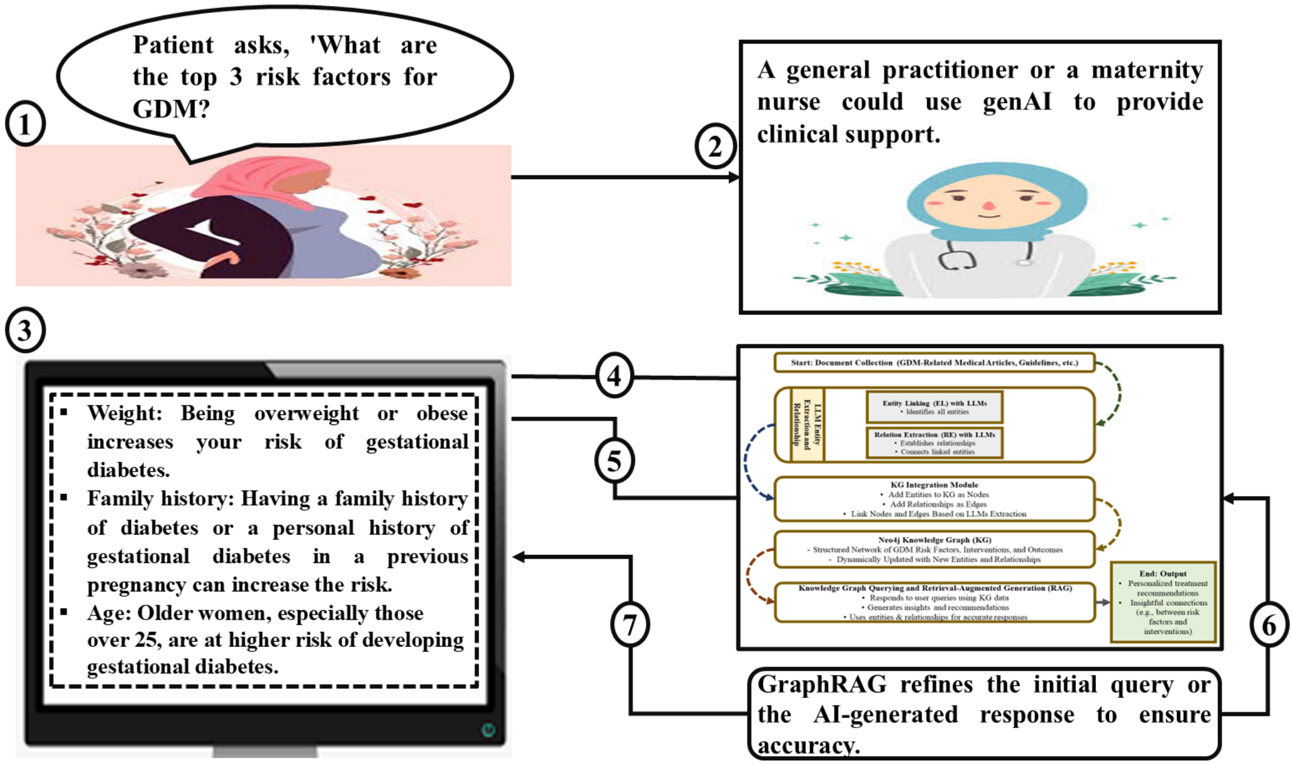
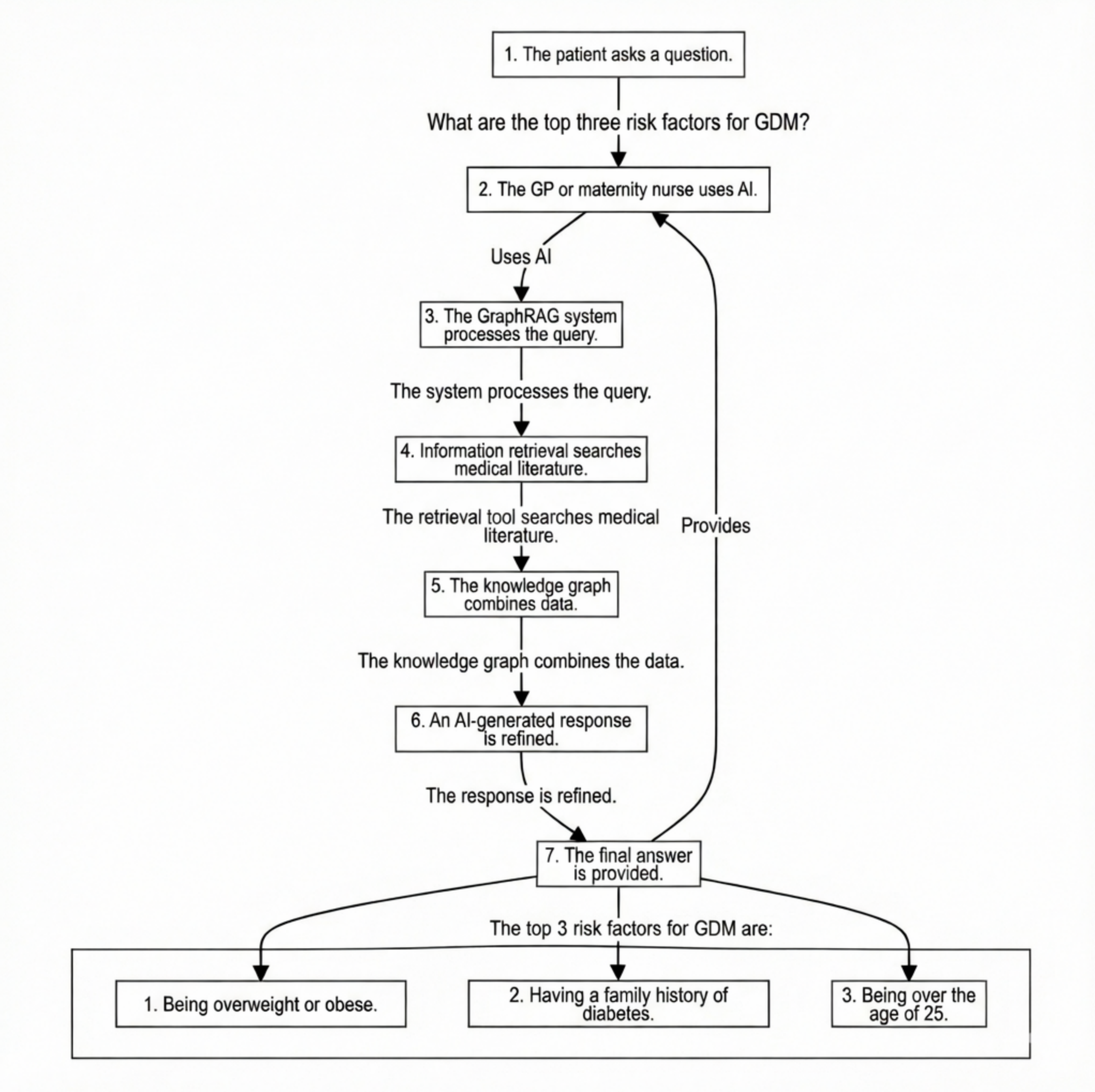


Figure 9. GraphRAG-based clinical support system for gestational diabetes mellitus - process flow diagram. AI: artificial intelligence; GDM: gestational diabetes mellitus; GP: general practitioner.



As illustrated in Figure 8, the system processes the initial query. It generates a concise, contextually relevant response (ie, with the benefit of the patient's medical records), listing key GDM risk factors such as maternal weight, family history of diabetes, and maternal age. The process flow highlights how the system integrates domain-specific medical literature and patient-related contextual data through its underlying Neo4j KG, enabling it to deliver evidence-based, patient-centered recommendations.

Figure 9 further elaborates on the underlying process. The patient's query initiates a series of steps where the system retrieves relevant interventions and relationships from the KG. The local LLM then generates a response with context-aware medical knowledge. The final advisory to the health care professional integrates the patient's specific context and up-to-date medical research, avoiding potential inaccuracies and hallucinations.

This demonstration shows how GraphRAG can streamline clinical consultations by providing up-to-date, evidence-backed responses. Beyond basic patient interactions, the system can address more advanced clinical questions, such as identifying appropriate screening tests for GDM or evaluating intervention effectiveness. With explanation features, these capabilities support informed decision-making for health care providers and patients, enhancing the overall quality of care for GDM.

The detailed results of the system's performance in various scenarios are presented in the following 3 subsections. The first examines the system's responses to layperson prompts. The one that follows analyzes its performance with clinically oriented queries from 2 GPs. The third presents the comparative analysis of model performance using standard evaluation metrics across GraphRAG and 3 other LLMs.

GraphRAG Local LLM Response to Prospective Patients' Prompts

The GraphRAG Local-LLM generated clear, clinically relevant responses to patients' queries related to GDM management. These prompts were designed to simulate questions that a general audience might ask during health consultations with their health care provider. The system's responses highlighted its ability to simplify complex medical concepts while maintaining factual accuracy and clinical relevance.

Table 4 presents example prompts and the corresponding GraphRAG-generated responses. For example, when asked about the *top 3 factors determining gestational diabetes risk*, the system correctly identified key risk factors such as *weight, family history of diabetes, and maternal age*.

Table . Responses to patient prompts.

Prompt	Response
What are the top 3 factors determining gestational diabetes risk?	• Weight: Being overweight or obese increases your risk of gestational diabetes. • Family history: Having a family history of diabetes or a personal history of gestational diabetes in a previous pregnancy can increase the risk. • Age: Older women, especially those over 25, are at higher risk of developing gestational diabetes.
How can we safeguard a baby's health and wellness for gestational diabetes patients?	• Monitor blood sugar levels, follow a healthy meal plan, engage in regular physical activity, take prescribed medications, attend prenatal check-ups, and follow healthcare guidance.
What nutrition is recommended to prevent gestational diabetes?	• Eat a balanced diet high in fiber and low in added sugars and unhealthy fats. Include fruits, vegetables, whole grains, lean proteins, and healthy fats.
Are gestational diabetes patients more likely to have a cesarean or a standard delivery?	• Women with gestational diabetes are more likely to have a cesarean delivery due to risks like macrosomia, shoulder dystocia, and birth trauma.
What neonatal screening should a couple with a family history of diabetes undergo before planning to have a baby? Are there precautions they could take?	• Consider genetic counseling and testing. Preconception counseling can help identify risks and precautions, with personalized plans developed alongside health care professionals.

These results show the system's capacity to convey medical information in a manner accessible to nonspecialist users. The responses accurately addressed the queries using medically appropriate language, ensuring clarity for a general audience.

Responses to GP Prompts

The GraphRAG system was evaluated using clinically relevant prompts provided by GPs. These prompts simulated real-world clinical queries related to the diagnosis, management, and long-term planning of GDM. The system generated medically valid responses as assessed by the participating GPs and diabetologists, addressing both basic and advanced aspects of GDM care.

Responses to Prompts From GP 1

Table 5 presents the responses generated by GraphRAG for the first set of prompts from GP 1. These prompts covered foundational aspects of GDM diagnosis and management. For example, when asked how to diagnose GDM, the system recommended the glucose challenge and tolerance tests, reflecting standard clinical guidelines. The responses addressed both diagnostic procedures and recommended clinical management steps.

The responses provided clear, evidence-aligned answers covering diagnostic tests, management strategies, fetal monitoring approaches, and long-term complications.

Table . Responses to prompts from general practitioner 1.

Prompt	Response
How do I diagnose gestational diabetes?	Gestational diabetes is typically diagnosed through a glucose challenge or glucose tolerance test. A sugary solution is consumed, and blood sugar levels are tested. If results are high, a more detailed glucose tolerance test is done. A health care provider must confirm the diagnosis and guide management.
What measures would you advise me to take once I am diagnosed with GDM[a]?	Monitor blood sugar regularly, follow a health care–guided meal plan, engage in physical activity, attend prenatal care visits, take prescribed medications, and stay in contact with your health care team.
What fetal changes should I monitor (and how) in gestational diabetes?	Monitor fetal growth via regular ultrasounds, track amniotic fluid and fetal movements (kick counts), and consult health care providers for ongoing prenatal assessments.
What are the cutting-edge medical and non-medical interventions I could take to control GDM in pregnant mothers?	Use technologies like continuous glucose monitors and personalized insulin therapy. Also, nonmedical options such as tailored diets, physical activity, and stress reduction should be applied, ideally with specialist consultation.
What are the long-term complications for GDM mothers and their offspring?	GDM increases the risk of type 2 diabetes in mothers and raises the chance of obesity, metabolic syndrome, and diabetes in children. Long-term follow-up is essential for both.

[a]GDM: gestational diabetes mellitus.

Prompts From GP 2 and GraphRAG Local LLM Responses

Further evaluation was conducted using prompts from GP 2, which explored broader clinical topics, including screening tests, genetic factors, and emerging therapies. Table 6 summarizes the system's responses.

According to the medical practitioners, the GraphRAG system generated contextually appropriate responses, addressing established clinical practices and emerging medical topics in GDM care.

Table . Responses to prompts from general practitioner 2.

Prompt	Response
What basic screening tests for GDM[a] should I perform and monitor regularly?	Standard tests for screening and monitoring GDM include: • Glucose challenge test (GCT) – a one-hour blood test after consuming a sugary drink. • Oral glucose tolerance test (OGTT) – a follow-up fasting test with multiple blood draws. • Glycated hemoglobin (A_{1C}) – reflects average glucose levels over the past months. • Regular blood glucose monitoring – daily checks using a glucometer. Consult a healthcare provider for tailored testing schedules.
On balance, does medical evidence suggest a greater genetic influence of the father or the mother for DM[b]?	Both parents can genetically influence the risk of DM. The extent varies by DM type and individual genetic profiles. Environmental factors also play a key role.
If the patient and her husband decide to go ahead with starting a family, what medical interventions can I prescribe to alleviate the risk of GDM?	Medical interventions include: • Healthy diet and lifestyle habits. • Regular blood glucose monitoring. • Medications such as metformin, if required. • Referral to a nutritionist or diabetes educator. • The provider should personalize recommendations.
Can proper oral and dental care alleviate GDM during pregnancy? What are some effective measures?	There is no direct link between oral care and GDM prevention. However, good dental hygiene supports overall pregnancy health. Recommendations include brushing twice daily, flossing, and routine dental visits.
Are there promising stem-cell or epigenetic treatments that could ease my patient's hesitance to start a family?	Stem-cell and epigenetic research show potential, but are still in the early stages. Patients should consult reproductive specialists for the latest updates and personalized advice.

[a]GDM: gestational diabetes mellitus.

[b]DM: diabetes mellitus.

Comparative Model Performance

Overview of Benchmarking Procedures

The GraphRAG system was benchmarked against 3 widely used LLMs, BioMistral, ChatGPT, and Claude, using a standardized set of clinical prompts focused on GDM management. The models' responses were evaluated using 5 quantitative metrics that assessed relevance, linguistic precision, terminology consistency, contextual understanding, and coherence.

Benchmarking Results

Figure 10 presents a comparative analysis of the models' average performance across 5 evaluation metrics. GraphRAG achieved the highest scores in BLEU, Jaccard Similarity, and BERTScore, indicating strong alignment with clinical phrasing, preservation of key medical terms, and deep contextual accuracy. Relevance Score and METEOR also reflect competitive performance across all models.

Figure 11 shows a radar chart (also known as a Kaviat diagram) of the same results, highlighting GraphRAG's balanced strengths across multiple evaluation dimensions.

Figure 10. Comparative performance of GraphRAG, BioMistral, ChatGPT, and Claude across evaluation metrics. BLEU: bilingual evaluation understudy.

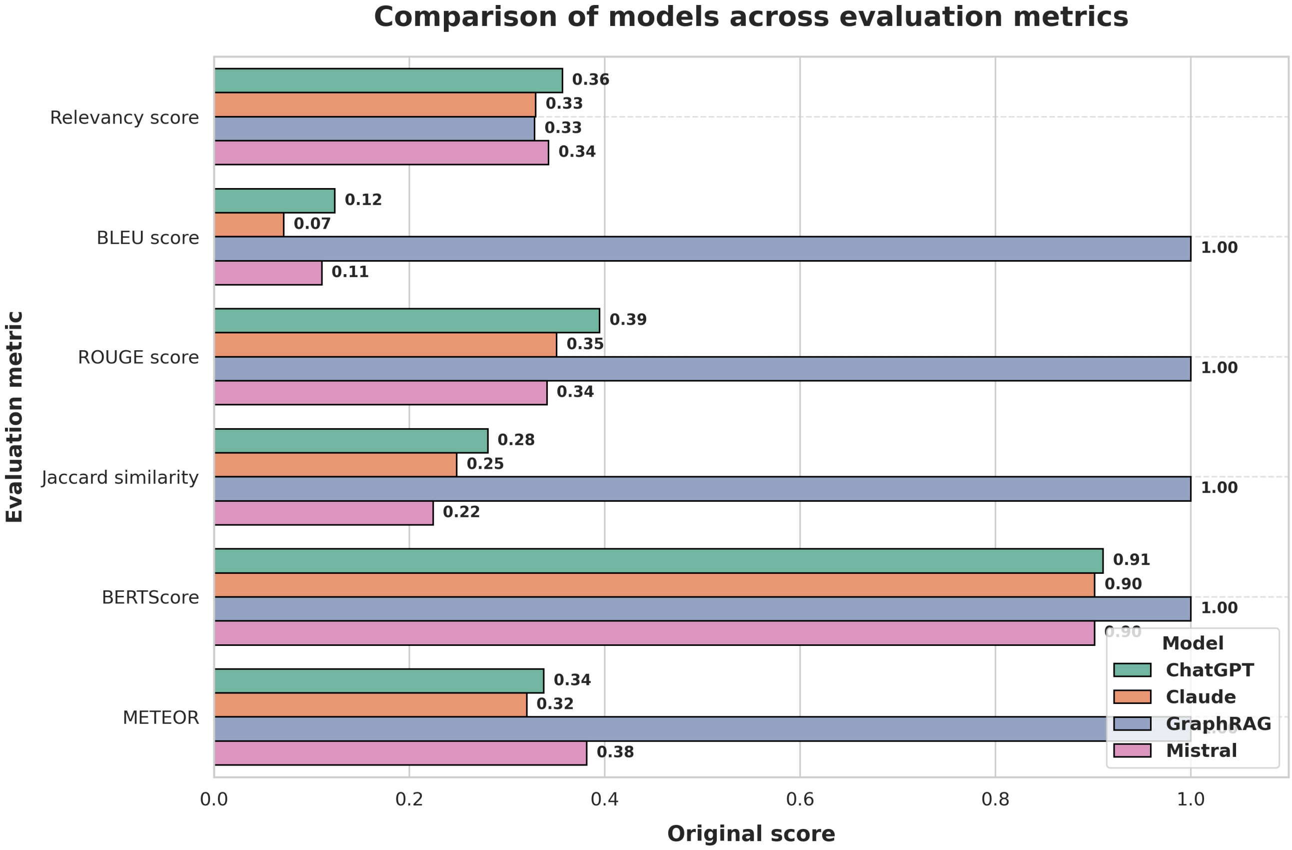


Figure 11. Radar chart visualizing model performance across key metrics. BLEU: bilingual evaluation understudy.

Key Observations

1. Relevance Score: GraphRAG and BioMistral showed comparable results, aligning well with the clinical intent of queries.
2. BLEU Score: GraphRAG outperformed all other models, reflecting precise replication of validated clinical expressions.
3. Jaccard Similarity: GraphRAG highlighted superior consistency in medical terminology usage across responses.
4. BERTScore: The model achieved the highest semantic similarity, indicating deep contextual understanding.
5. METEOR: GraphRAG generated coherent and fluent responses suitable for clinical communication, comparable with ChatGPT and Claude.

These findings demonstrate the technical feasibility of the proposed GraphRAG-enabled local LLM. However, we stress that as a PoC evaluated in a simulated environment, the prototype is not ready to be deployed in real-world clinical settings. Even so, these results show that the GraphRAG approach effectively balances linguistic precision, contextual depth, and clinical relevance in GDM decision support scenarios. Besides BioMistral, ChatGPT, and Claude, new open-source LLMs such as LLaMA 3, Mistral 7B, and Phi-3 are becoming prevalent in health care AI. Although hardware and framework limitations prevented their inclusion in this study, we recognize their importance as baselines. Future work will add these models to expand our comparative analysis.

Discussion

Principal Findings

This study demonstrates that the GraphRAG-enabled local LLM consistently produces clinically relevant, contextually grounded, and medically precise responses for managing GDM. Through a rigorous benchmarking process against established open-source models, BioMistral, ChatGPT, and Claude, the GraphRAG approach achieved top-tier performance across all key evaluation metrics: BLEU Score, Jaccard Similarity, and BERTScore. The radar chart (Figure 11) illustrates well-balanced strengths of the model across multiple evaluation dimensions. Notably, first, GraphRAG matched BioMistral in

Relevance Score, indicating strong alignment with the clinical intent behind queries of the health care providers. This score reflects how accurately the system understands and responds to the specific clinical context of the user, which is critical in decision support. Second, GraphRAG outperformed all models in BLEU Score and Jaccard Similarity, showcasing its ability to reproduce clinical phrasing with syntactic accuracy and maintain consistency in key medical terminologies, a vital factor for preserving the meaning of technical medical advice. Third, it achieved a superior BERTScore, showing deep semantic understanding. This reflects the capacity of the model to generate responses that not only match expected language structures but also accurately convey complex clinical relationships within GDM care. Finally, the superior performance across these diverse metrics stems from the architectural design of GraphRAG. By integrating domain-specific KGs with RAG, the system grounds its responses in verified clinical evidence rather than relying solely on probabilistic language patterns. This integration mitigates common challenges of general-purpose LLMs, such as hallucinations and domain irrelevance, ensuring that responses are both medically accurate and context-aware.

These findings validate the core hypothesis that combining structured medical knowledge and patient-specific information with GenAI significantly enhances the quality, safety, and reliability of AI-generated clinical recommendations.

Implications for AI-Driven Clinical Decision Support

The empirical results of this study highlight the transformative potential of GraphRAG-powered local LLMs in advancing AI-enabled clinical decision support systems. This is particularly the case within specialized domains, such as GDM management, where expertise is scarce and uneven. By combining structured retrieval from a domain-specific KG with contextual generation, the system addresses the longstanding gap between unstructured clinical queries and structured evidence-based medical knowledge.

The GraphRAG approach addresses a critical limitation of general-purpose LLMs, such as ChatGPT and Claude, which often prioritize linguistic fluency over clinical accuracy. While these models can generate coherent responses, they frequently lack the domain specificity needed for accurate clinical guidance. In comparison, the responses of GraphRAG consistently align with established clinical guidelines, reflecting a deep understanding of current medical standards and practices. For example, when prompted to hear about GDM diagnosis, GraphRAG accurately recommended the glucose challenge and tolerance tests, mirroring clinical best practices. This indicates that the system is not merely generating plausible text but retrieving and contextualizing domain-specific evidence to support clinical decision-making.

The clinical utility of such contextually enriched responses is profound. In healthcare, where treatment decisions directly affect patient safety and outcomes, factual accuracy and contextual relevance are not optional but essential. The ability of GraphRAG to consistently deliver these qualities positions it as a valuable tool for supporting health care providers, particularly in low-resource or high-pressure clinical environments where access to specialist knowledge may be limited.

The practical utility of GraphRAG is further illustrated in Textbox 2, which presents a representative response to a clinically relevant query about managing insulin resistance in patients with GDM. Unlike generic language models, GraphRAG provides structured, evidence-aligned recommendations grounded in recent clinical guidelines.

Textbox 2. GraphRAG response to a clinical query on managing GDM.

Query: What are the latest recommendations for managing insulin resistance in GDM patients?

GraphRAG response: Recent studies suggest a multi-faceted approach to managing insulin resistance in GDM patients, including:

1. Personalized low glycemic index diets, emphasizing complex carbohydrates and fiber-rich foods.

2. Structured exercise programs combining aerobic and resistance training for at least 150 minutes per week.

3. Early initiation of metformin where lifestyle modifications are insufficient, as per the 2023 guidelines.

4. Continuous glucose monitoring to optimize insulin dosing and adjust lifestyle recommendations.

5. Regular screening for comorbidities such as hypertension and dyslipidemia, which contribute to insulin resistance.

The above sections highlight the ability of GraphRAG to transform unstructured clinical questions into actionable, guideline-compliant insights. By synthesizing evidence from domain-specific KGs, the system avoids unsupported claims and produces responses aligned with best clinical practices, supporting its role as a trustworthy clinical decision support tool.

Contributions to AI in Health Care

This study advances the field of health care AI by presenting a scalable, contextually enriched clinical support system specifically designed for GDM management. We believe that our key contribution lies in the system's ability to empower GPs and nonspecialist clinicians, particularly in underserved and resource-limited health care environments with limited access to endocrinology specialists and up-to-date clinical knowledge. By using a KG-driven retrieval process, the system surfaces context-specific clinical insights without requiring clinicians to conduct exhaustive manual literature reviews or consult multiple sources. Here, a word of caution is in order. We reiterate that the PoC works best as a clinical assistant; that is, a health practitioner must be in the loop. This is important given the dangers of unsupervised AI agents, which may usurp the role of a human caregiver without human oversight [44]. It is concerning that a recent, peer-reviewed (and in our view, misguided) study actually normalizes a doctor versus machine "Turing-test of authenticity" [45].

Furthermore, this study shows domain-specific superiority over general-purpose LLMs. While models such as ChatGPT and Claude can produce coherent responses, they lack the fine-tuned contextual sensitivity and clinical precision essential for specialized health care domains. In comparison, the architecture of GraphRAG is optimized to capture the complex relationships inherent in GDM management, such as patient history with risk factors, availability of interventions, and outcome pathways for follow-up medical care, thereby enhancing both response accuracy and clinical applicability.

This study contributes to a novel retrieval-augmented GenAI architecture that translates domain-specific medical knowledge into clinically actionable insights. It serves a need; namely, access to the latest, credible medical research in time- and resource-constrained environments. In health care, timely and science-based interventions are crucial.

Technical Innovations Driving Performance Gains

The robust performance of the GraphRAG-enabled local LLM stems from the integration of 3 core technical innovations that address longstanding limitations in clinical AI systems.

First, the KG integration allows for the structured representation of complex clinical relationships between risk factors, interventions, symptoms, and outcomes. Unlike flat text embedding, the KG enables the system to reason over interconnected entities and contextual dependencies, ensuring that recommendations are grounded in the complete clinical scenario rather than isolated data points.

Second, the RAG framework of the system addresses the gap between static model knowledge and dynamic, evolving medical evidence. The system mitigates temporal gaps by integrating retrieval from an up-to-date domain-specific KG. It reduces the risk of hallucinated or outdated responses, a common flaw in general-purpose LLMs trained on static corpora.

Third, the domain-specific adaptation of the model through targeted prompting strategies and fine-tuning on GDM-related interventions enhances its ability to understand and accurately apply specialized clinical terminology in localized contexts. This adaptation ensures that the system's responses reflect the nuanced requirements of GDM management, capturing both the syntactic precision and semantic depth necessary for high-stakes clinical situations like emergency room triage.

We believe that these innovations enable the system to move beyond generic language generation, delivering interpretable, actionable, and clinically validated responses. This advancement represents a meaningful step toward reliable AI-assisted clinical decision-making, especially for chronic disease management scenarios where timely and context-aware recommendations are essential.

Conclusions

Limitations and Challenges for Clinical Deployment

While the initial results from this PoC study are promising, several critical limitations must be addressed before GraphRAG can be translated into clinical practice. Intended as a PoC, the system has not undergone field validation. Future studies involving real-world patient interactions, clinician feedback, and longitudinal follow-up are essential to establish the model's safety, reliability, and usability in live health care environments.

A second major consideration concerns data privacy and protection. Although this PoC did not involve patient-level data, real-world deployments would necessitate strict adherence to data protection frameworks. The integration of privacy-preserving learning paradigms, such as federated learning, would allow models to be trained on decentralized clinical data without exposing sensitive patient information. Complementary techniques, such as blockchain for differential privacy and secure multiparty computation, could further protect patient confidentiality.

The interpretability of AI-generated clinical responses remains a pressing challenge. While GraphRAG uses structured retrieval to enhance contextual grounding, clinicians must be able to trust and explain its outputs. Future iterations of the system should integrate explainability frameworks such as Shapley Additive Explanations or Local Interpretable Model-agnostic Explanations, enabling clinicians to trace and retrieve evidence on how specific KG pathways contribute to a given clinical recommendation.

In addition, seamless workflow integration will be critical for adoption. Clinical decision support systems must embed naturally within existing electronic health record platforms, minimizing disruption to physician workflows. Without such integration, even the most accurate systems risk being underused in clinical practice.

As with many multistage AI pipelines, GraphRAG is also subject to the risk of error propagation, where inaccuracies in earlier stages, such as entity extraction or graph construction, may be compounded in downstream response generation. While our current prompt engineering and domain-specific graph design reduce this risk, future versions will integrate intermediate validation checkpoints, feedback loops, and retrieval-failure auditing to ensure response fidelity and system transparency.

Another key limitation is the reliance on English-language peer-reviewed articles from a single aggregator (Semantic Scholar). This has excluded regional or non-English medical literature with culturally adapted GDM interventions. Future work should incorporate multilingual and regionally diverse corpora to improve the model's generalizability and contextual sensitivity, particularly in Global South health care settings.

Finally, the computational demands of GraphRAG's RAG architecture present scalability challenges. The latency and resource consumption must be optimized to support real-time inference in time-sensitive clinical settings, especially in environments where computational capacity may be limited. Addressing these challenges is essential for transitioning GraphRAG from an academic PoC to a clinically viable, ethically responsible AI system.

Broader Implications and Future Research Directions

Building on the demonstrated feasibility of our PoC, our future research agenda is designed to advance the GraphRAG framework along 2 primary axes: strategic domain expansion and core technical refinement. First, we propose to strategically adapt the framework for other data-intensive clinical areas, including cardiovascular disease, oncology, and mental health, where evidence-grounded decision support is crucial. Second, we will enhance the core retrieval engine by integrating advanced algorithms, such as contextual BM25 and embedding-based summarization, to improve precision. To improve robustness and transparency, we propose implementing new retrieval-specific metrics, such as recall and failure rates. We have established a roadmap and aim to pursue these enhancements in our next research cycle, solidifying the GraphRAG pipeline as a viable tool for real-world clinical decision support.

The legal, ethical, and intellectual property considerations will also shape future deployments. To ensure transparency and reduce legal risks, future iterations will prioritize training on open-access datasets such as PubMed Central, adhering to responsible AI development practices and open science principles.

To protect patient privacy and mitigate algorithmic bias will remain core ethical imperatives. The federated learning and anonymized blockchain solutions could support decentralized training across institutions without compromising patient confidentiality. Bias audits, fairness-aware modeling, and hallucination mitigation strategies, such as reranking retrieved evidence and diversifying training datasets, will improve the reliability and equity of the system's clinical recommendations. In such a trusted platform, integrating GraphRAG with real-time patient data could enable personalized clinical decision support, customizing recommendations to individual genetic profiles, lifestyle factors, and environmental exposures. This evolution toward precision medicine would represent a significant leap forward in AI-driven health care delivery.

To overcome the limitation of computational costs, the enhanced system will require architectural optimizations to enable scalability in resource-constrained clinical settings. Techniques such as prompt caching, adaptive chunking of graph queries, and hybrid retrieval strategies will reduce computational costs and response latency. This will support deployments in low-bandwidth environments, such as rural clinics and community health centers.

In the long term, retrieval-augmented LLMs, such as GraphRAG, are envisioned not as autonomous clinical agents but as clinical copilots, supporting, rather than replacing human clinicians. Their evaluation in live clinical workflows will be critical to determining their optimal role as decision-support systems. A reflective perspective on this motivation is presented in Textbox 3, showing the personal origins of our research question.

Textbox 3. Closing vignette on gestational diabetes.

"I do not wish to alarm you, Mrs. Sharma, but you have been diagnosed with gestational diabetes and your baby is 10 pounds at birth. Both of you need to be careful."
[Ward Nurse in Singapore's Kandang Kerbau Maternity Hospital to the mother of the last author, circa 1961]
In 2022, the mother passed away peacefully at the age of 88, her diabetes controlled with insulin injections for decades. The "baby" (the last author and principal investigator of this study [RS]) was diagnosed with type 2 diabetes at the age of 60, giving rise to our research question of whether a graph-based retrieval-augmented generation solution could change the outcome for both with timely, relevant best practices.

In closing, this paper sought to establish the feasibility of a GraphRAG-enabled local LLM architecture for generating clinically relevant, context-aware responses in the management of diseases, such as GDM [46]. By integrating domain-specific KGs with RAG, the system outperformed general-purpose LLMs across multiple evaluation metrics, offering evidence-grounded and terminologically precise clinical recommendations. While this work serves as a technical PoC, future research will need to focus on (1) prospective clinical validation involving real-time patient interaction, (2) multimodal agents to improve accessibility and cultural sensitivity, and (3) integration of explainable AI modules, such as Shapley Additive Explanations–based KG traceability, resulting in enhanced trust and transparency for the 2 key humans in the loop – the patient and her doctor. Ultimately, we believe the transformative potential of AI-powered decision support tools will personalize care and improve clinical outcomes, particularly in underserved societies.

Acknowledgments

The authors are grateful to the three medical practitioners, Dr Z (a diabetologist), Dr S (an emergency room physician), and Dr R (a general physician), for their thoughtful contribution to the prompts and for examining the GraphRAG responses for face validation. Many thanks are also due to the anonymous reviewers for guiding the revisions of the manuscript. No human subjects nor live patient data were used to develop the PoC.

Funding

This research was supported by the Research Incentive Fund Grant R22086 from Zayed University to AN, RS, and EE.

Data Availability

The datasets generated or analyzed during this study are available in the GITHUB repository [46].

Authors' Contributions

AN and RS conceived the research idea and were Principal Investigators. FR and SB conducted the empirical data collection and validation that produced the GraphRAG model. EE supervised the research and subsequent reporting as co-PI. All authors contributed equally to the research, analysis, and writing of this article.

No potential conflicts of interest are reported by the authors.

Conflicts of Interest

None declared.

References

1. Sharma R, Zhang C, Wingreen SC, Kshetri N, Zahid A. Design of blockchain-based precision health-care using soft systems methodology. IMDS 2019 Dec 31;120(3):608-632. [doi: 10.1108/IMDS-07-2019-0401]
2. Nambiar M, Bee YM, Chan YE, et al. A drug mix and dose decision algorithm for individualized type 2 diabetes management. NPJ Digit Med 2024 Sep 17;7(1):254. [doi: 10.1038/s41746-024-01230-5] [Medline: 39289474]
3. Kjos SL, Buchanan TA. Gestational diabetes mellitus. N Engl J Med 1999 Dec 2;341(23):1749-1756. [doi: 10.1056/NEJM199912023412307] [Medline: 10580075]
4. International Association of Diabetes and Pregnancy Study Groups Consensus Panel, Metzger BE, Gabbe SG, et al. International association of diabetes and pregnancy study groups recommendations on the diagnosis and classification of hyperglycemia in pregnancy. Diabetes Care 2010 Mar;33(3):676-682. [doi: 10.2337/dc09-1848] [Medline: 20190296]
5. Akinyemi OA, Omokhodion OV, Fasokun ME, Makanjuola D, Ade-Ojo IP, Adeniyi AA. Screening for gestational diabetes mellitus: is there a need for early screening for all women in developing countries? Cureus 2023;15(2):e35533. [doi: 10.7759/cureus.35533]
6. Ismail L, Materwala H. IDMPF: intelligent diabetes mellitus prediction framework using machine learning. ACI 2025 Jan 30;21(1/2):78-89. [doi: 10.1108/ACI-10-2020-0094]
7. Ejaz Z, Azhar Khan A, Sebghat Ullah S, Aamir Hayat M, Maqbool MA, Amin Baig A. The effects of gestational diabetes on fetus: a surveillance study. Cureus 2023 Feb;15(2):e35103. [doi: 10.7759/cureus.35103] [Medline: 36938248]
8. Macaulay S, Dunger DB, Norris SA. Gestational diabetes mellitus in Africa: a systematic review. PLoS ONE 2014;9(6):e97871. [doi: 10.1371/journal.pone.0097871] [Medline: 24892280]
9. Stewart-Field ZA. Gestational diabetes: screening, diagnosis, treatment and management. Obstetrics, Gynaecology & Reproductive Medicine 2023 Jul;33(7):185-189. [doi: 10.1016/j.ogrm.2023.04.001]
10. Kurt B, Gürlek B, Keskin S, et al. Prediction of gestational diabetes using deep learning and Bayesian optimization and traditional machine learning techniques. Med Biol Eng Comput 2023 Jul;61(7):1649-1660. [doi: 10.1007/s11517-023-02800-7] [Medline: 36848010]
11. Jani S, Parikh P, Patel K, Shah A, Patel B, Rangrej R. Fetomaternal outcome in patients with gestational diabetes mellitus. Natl J Physiol Pharm Pharmacol 2023;13(3):1. [doi: 10.5455/njppp.2023.13.01032202306022023]
12. Alum E, Ugwu OPC, Obeagu EI, Uti DE, Egba SI, Alum BN. Managing the dual burden: addressing mental health in diabetes care. Elite Journal of Medical Sciences 2024;2(6):1-9 [FREE Full text]
13. Hinneh T, Jahn A, Agbozo F. Barriers to screening, diagnosis and management of hyperglycaemia in pregnancy in Africa: a systematic review. Int Health 2022 May 2;14(3):211-221. [doi: 10.1093/inthealth/ihab054] [Medline: 34432050]
14. Karakoc Kumsar A, Taskin Yilmaz F, Demirel G. Preferences of participating in diabetes screening programs for postpartum women with gestational diabetes mellitus in a university hospital in Turkey. JHR 2022 Apr 27;36(3):524-532. [doi: 10.1108/JHR-08-2020-0372]
15. Sahu B, Babu GR, Gurav KS, et al. Health care professionals' perspectives on screening and management of gestational diabetes mellitus in public hospitals of South India - a qualitative study. BMC Health Serv Res 2021 Feb 12;21(1):133. [doi: 10.1186/s12913-021-06077-0] [Medline: 33579259]
16. Nikolaidis GF, Duarte A, Griffin S, Lomas J. "Beyond the mean" in biomarkers modelling for economic evaluations: a case study in gestational diabetes mellitus. In: Baltagi BH, Moscone F, editors. Recent Developments in Health Econometrics: A Volume in Honour of Andrew Jones: Emerald Publishing Limited; 2024, Vol. 297:1108. [doi: 10.1108/S0573-85552024297]
17. Morampudi S, Balasubramanian G, Gowda A, Zomorodi B, Patil AS. The challenges and recommendations for gestational diabetes mellitus care in India: a review. Front Endocrinol (Lausanne) 2017;8:56. [doi: 10.3389/fendo.2017.00056] [Medline: 28392778]
18. Mpondo BCT, Ernest A, Dee HE. Gestational diabetes mellitus: challenges in diagnosis and management. J Diabetes Metab Disord 2015;14(1):42. [doi: 10.1186/s40200-015-0169-7] [Medline: 25977899]

19. Pal A, Umapathi LK, Sankarasubbu M. Med-HALT: Medical Domain Hallucination Test for Large Language Models. Presented at: Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL); Dec 6-7, 2023; Singapore p. 314-334. [doi: 10.18653/v1/2023.conll-1.21]
20. Miao J, Thongprayoon C, Suppadungsuk S, Garcia Valencia OA, Cheungpasitporn W. Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications. Medicina (Kaunas) 2024 Mar 8;60(3):445. [doi: 10.3390/medicina60030445] [Medline: 38541171]
21. Oladimeji OO, Oladimeji A, Oladimeji O. Classification models for likelihood prediction of diabetes at early stage using feature selection. ACI 2024 Jun 17;20(3-4):279-286. [doi: 10.1108/ACI-01-2021-0022]
22. Kokori E, Olatunji G, Aderinto N, et al. The role of machine learning algorithms in detection of gestational diabetes; a narrative review of current evidence. Clin Diabetes Endocrinol 2024 Jun 25;10(1):18. [doi: 10.1186/s40842-024-00176-7] [Medline: 38915129]
23. Kumar M, Chen L, Tan K, et al. Population-centric risk prediction modeling for gestational diabetes mellitus: a machine learning approach. Diabetes Res Clin Pract 2022 Mar;185:109237. [doi: 10.1016/j.diabres.2022.109237] [Medline: 35124096]
24. Mennickent D, Rodríguez A, Opazo MC, et al. Machine learning applied in maternal and fetal health: a narrative review focused on pregnancy diseases and complications. Front Endocrinol (Lausanne) 2023;14:1130139. [doi: 10.3389/fendo.2023.1130139] [Medline: 37274341]
25. Wu YT, Zhang CJ, Mol BW, et al. Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning. The Journal of Clinical Endocrinology & Metabolism 2021 Mar 8;106(3):e1191-e1205. [doi: 10.1210/clinem/dgaa899]
26. Lu HY, Ding X, Hirst JE, et al. Digital health and machine learning technologies for blood glucose monitoring and management of gestational diabetes. IEEE Rev Biomed Eng 2023;17:98-117. [doi: 10.1109/RBME.2023.3242261]
27. Pinto Y, Frishman S, Turjeman S, et al. Gestational diabetes is driven by microbiota-induced inflammation months before diagnosis. Gut 2023 May;72(5):918-928. [doi: 10.1136/gutjnl-2022-328406] [Medline: 36627187]
28. Cubillos G, Monckeberg M, Plaza A, et al. Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy. BMC Pregnancy Childbirth 2023 Jun 23;23(1):469. [doi: 10.1186/s12884-023-05766-4] [Medline: 37353749]
29. Wu J, Zhu J, Qi Y, et al. Medical Graph RAG: evidence-based medical large language model via graph retrieval-augmented generation. : Association for Computational Linguistics Presented at: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1); Jul 27 to Aug 1, 2025; Vienna, Austria p. 28443-28467. [doi: 10.18653/v1/2025.acl-long.1381]
30. Nazi ZA, Peng W. Large language models in healthcare and medical domain: a review. Informatics (MDPI) 2024;11(3):57. [doi: 10.3390/informatics11030057]
31. Ge J, Sun S, Owens J, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. Hepatology 2024 Nov 1;80(5):1158-1168. [doi: 10.1097/HEP.0000000000000834] [Medline: 38451962]
32. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. NPJ Ment Health Res 2024 Apr 2;3(1):12. [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
33. Tripathi S, Sukumaran R, Cook TS. Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care. J Am Med Inform Assoc 2024 May 20;31(6):1436-1440. [doi: 10.1093/jamia/ocad258] [Medline: 38273739]
34. Ong JCL, Jin L, Elangovan K, et al. Development and testing of a novel large language model-based clinical decision support system for medication safety in 12 clinical specialties. arXiv. Preprint posted online on Jan 29, 2024. [doi: 10.48550/arXiv.2402.01741]
35. Semantic Scholar API. Semantic Scholar. 2023. URL: https://www.semanticscholar.org/product/api [accessed 2025-03-18]
36. ChatGPT: optimizing language models for dialogue. OpenAI. 2023. URL: https://openai.com/index/chatgpt/ [accessed 2025-12-14]
37. Trajanoska M, Stojanov R, Trajanov D. Enhancing knowledge graph construction using large language models. arXiv. Preprint posted online on May 8, 2023. [doi: 10.48550/arXiv.2305.04676]
38. Neo4j life sciences and healthcare network. Neo4j. URL: https://neo4j.com/developer/life-sciences-and-healthcare [accessed 2025-09-14]
39. Hu Y, Lei Z, Zhang Z, Pan B, Ling C, Zhao L. GRAG: Graph Retrieval-Augmented Generation. In: Findings of the Association for Computational Linguistics: NAACL 2025: Association for Computational Linguistics; 2025. URL: https://aclanthology.org/2025.findings-naacl.232/ [accessed 2025-09-14] [doi: 10.18653/v1/2025.findings-naacl.232]
40. Claude. 2023. URL: https://www.anthropic.com/claude [accessed 2025-12-14]
41. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: a collection of open-source pretrained large language models for medical domains. In: Findings of the Association for Computational Linguistics: ACL 2024: Association for Computational Linguistics; 2024. [doi: 10.18653/v1/2024.findings-acl.348]

42. Ngo NT, Van Nguyen C, Dernoncourt F, Nguyen TH. Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering. arXiv. Preprint posted online on Nov 14, 2024. [doi: 10.48550/arXiv.2411.09213]
43. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. In: Findings of the Association for Computational Linguistics: ACL 2024: Association for Computational Linguistics; 2024:6233-6251. [doi: 10.18653/v1/2024.findings-acl.372]
44. Dobkin R. A man asked ChatGPT how to remove sodium chloride from his diet. It landed him in the hospital. The Independent. 2025 Aug 7. URL: https://www.independent.co.uk/news/health/chatgpt-medical-advice-hospital-b2803992.html [accessed 2025-12-22]
45. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med 2024 Sep;30(9):2613-2622. [doi: 10.1038/s41591-024-03097-1] [Medline: 38965432]
46. Ruba F, Nazir A, Evangelista E, Bukhari S, Loft L, Sharma R. GraphRAG-architecture-of-a-local-LLM-for-gestational-diabetes-mellitus. GitHub. 2025. URL: https://github.com/fathima-ruba/GraphRAG-Architecture-of-a-local-LLM-for-Gestational-Diabetes-Mellitus [accessed 2025-12-08]

Abbreviations

AI: artificial intelligence
BLEU: bilingual evaluation understudy
GDM: gestational diabetes mellitus
GenAI: generative artificial intelligence
GP: general practitioner
GraphRAG: graph-based retrieval-augmented generation
KG: knowledge graph
LLM: large language model
ML: machine learning
PoC: proof-of-concept
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG: retrieval-augmented generation
SML: small language model
T2D: type 2 diabetes

Edited by R Zhao; submitted 23.Apr.2025; peer-reviewed by M Azam, Y Xu, Y Dong; revised version received 14.Oct.2025; accepted 14.Oct.2025; published 05.Jan.2026.

Please cite as:
Evangelista E, Ruba F, Bukhari S, Nazir A, Sharma R
GraphRAG-Enabled Local Large Language Model for Gestational Diabetes Mellitus: Development of a Proof-of-Concept
JMIR Diabetes 2026;11:e76454
URL: https://diabetes.jmir.org/2026/1/e76454
doi:10.2196/76454

Continuous Ketone Monitoring: Data From a Randomized Controlled Trial

Simon K Kjær[1,2], MD; Lukas Ochsner Reynaud Ridder[2,3], MD, PhD; Mads Svart[2,4], MD, PhD; Nikolaj Rittig[1,2], MD, PhD; Lise Nørkjær Bjerg[5], MSc, PhD; Birgitte Sandfeld-Paulsen[6], MD, PhD; Henrik Holm Thomsen[2,7], MD, PhD

[1]Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus, Denmark
[2]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
[3]Department of Endocrinology and Internal Medicine, Aarhus University Hospital, Aarhus, Denmark
[4]Department of Internal Medicine, Regional Hospital Horsens, Horsens, Denmark
[5]Department of Clinical Biochemistry, Regional Hospital Central Jutland, Viborg, Denmark
[6]Department of Clinical Biochemistry, Aalborg University Hospital, Aalborg, Denmark
[7]Medical Diagnostic Center, University Clinic for Innovative Patient Pathways, Regional Hospital Central Jutland, Heibergs Allé 4K, Viborg, Denmark

Corresponding Author:
Henrik Holm Thomsen, MD, PhD
Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

Abstract

In our study, a commercially available continuous ketone monitoring device captured β-Hydroxybutyrate (BHB) dynamics during exogenous ketosis but revealed a gradual decline day-to-day BHB concentrations over 14 days in both ketone ester and placebo groups, likely reflecting sensor drift.

(JMIR Diabetes 2026;11:e85548) doi:10.2196/85548

KEYWORDS

continuous ketone monitoring; exogenous ketosis; beta-hydroxybutyrate; ketone sensor technology; ketone ester

Introduction

Continuous measurement of ketone bodies is of scientific and clinical interest, providing insights into type 1 and type 2 diabetes, ketogenic diets, intermittent fasting, and exogenous ketone precursor supplementation. Current finger-prick point-of-care testing (POCT) devices are invasive, intermittent, and fail to capture dynamic fluctuations [1]. Continuous ketone monitoring (CKM), a small device measuring interstitial ketone (β-hydroxybutyrate, BHB) levels, offers a potential solution [2]. CKM research, however, remains in its early stages, with only a single commercially available device at present (SiBio KS1, Hong Kong), to the best of our knowledge. Exogenous ketone supplementations are currently studied for potential therapeutic applications, including weight loss, enhanced exercise performance, and the management of neurodegenerative, cardiovascular, and inflammatory conditions [3-5]. We hypothesized that CKM would accurately track BHB and evaluated its performance under sustained intermittent supraphysiological ketosis.

Methods

Study Design

This work is part of a larger study on exogenous ketosis and erythropoiesis (Thomsen et al, unpublished). CKM became available midway through the study and was therefore applied sequentially in the final 7 of the 16 healthy volunteers. Participants were randomized to receive either a ketone ester (KE) drink (500 mg/kg/d) or a placebo (PBO), matched for volume, taste, and viscosity. Over two weeks, drinks were consumed two to three times daily, with half the dose before sleep. Participants were blinded to CKM readings, while investigators were not blinded. We tested the effects of time, treatment, and their interaction on log-transformed BHB area under the curve (AUC) using a linear mixed-effects model and applied polynomial contrasts to assess linear trends.

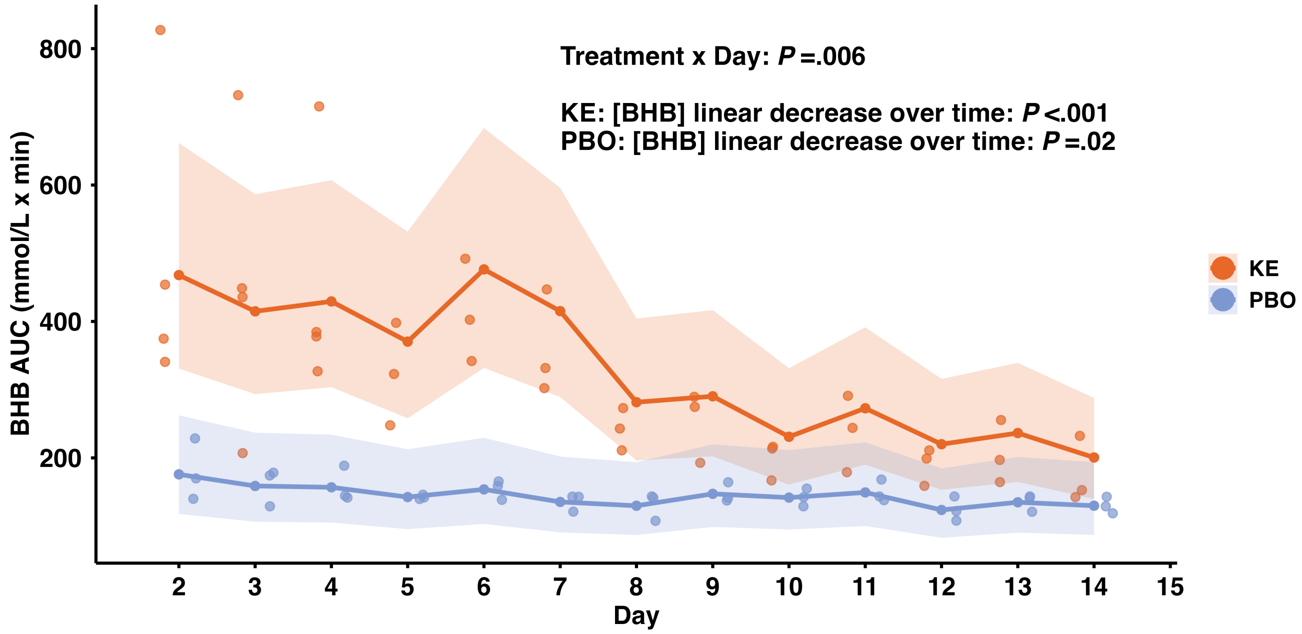
Ethical Considerations

The study was conducted in accordance with the Declaration of Helsinki II, approved by the regional ethics committee (#1-10-72-221-22), and registered with ClinicalTrials (NCT06053138). Oral and written informed consent was obtained from all participating patients. Participant data were pseudonymized to ensure confidentiality. Participants received financial compensation for their time and participation.

Results

A total of 7 participants wore CKM devices: 4 in the KE group (3 female, 1 male) and 3 in the PBO group (2 female, 1 male). Median age was 41 years (IQR 28–55). One KE participant's sensor detached on day 4 and was not replaced, but CKM readings until detachment were included in the analyses. BHB AUCs were significantly influenced by both day and treatment, with an interaction effect (*P*=.006). In the KE group, BHB showed a significant linear decrease over 14 days (*P*<.001), and a smaller but significant decline was also observed in the PBO group (*P*=.02). Consequently, group differences diminished, with KE and PBO becoming indistinguishable by the final day (Figure 1).

Figure 1. Day-by-day changes in total BHB area under the curve (AUC) for both the Ketone Ester (KE) group (n=4, orange) and placebo (PBO) group (n=3, blue). Scatter points represent individual AUC measurements for each participant across the 14 study days. Solid lines depict the back-transformed least-square means of BHB concentrations from a mixed-effects model, estimated separately for each day and treatment group, and the shaded regions represent the confidence intervals. BHB: beta-hydroxybutyrate.



Discussion

This study evaluated the performance of a commercially available CKM device during 14 days with intermittent exogenous ketone supplementation. Our findings demonstrate that the CKM detected increases in interstitial BHB concentrations following KE ingestion but revealed a progressive decline in BHB concentrations over the 14-day study period in the KE group, indistinguishable from the PBO group on the last study day. This contrasts with two prior studies in which participants received KE for 14 days before ingesting 25 g KE in a laboratory setting on day 15 [6,7]. In those studies, peaks reached ~2.3 mM at 1 hour and declined to ~0.5 mM at 4 hours, with no evidence of a declining peak BHB concentration following a comparable period of intermittent exogenous ketosis. Importantly, we observed a temporal decline in BHB concentrations also in the placebo group, highly suggesting a ketone-independent physiological or measurement-related drift. Therefore, this raises the possibility of sensor-related limitations. Potential explanations include sensor enzyme degradation, biofouling, temperature effects, compression, or interstitial variability [8]. The underlying sensor principle is not fully disclosed but thought to use a modified electrochemical method reacting selectively with BHB in interstitial fluid. In comparison, an in-development multianalyte sensor using a three-electrode system with NAD^+-dependent β-hydroxybutyrate dehydrogenase and osmium-based redox chemistry has shown stable 14-day performance in 12 healthy, low-carbohydrate-consuming participants [9,10]. A future study is anticipated with interest since it will assess the accuracy of the same device used in our study, SiBio KS1, in subjects following a 14-day ketogenic diet (NCT06420518). Limitations for our study include not comparing the CKM-derived ketone levels with gold standard blood BHB measurements (eg, finger-prick tests), making it difficult to definitively decide if our observations are due to sensor-specific limitations or not. Additionally, the small sample size and statistical power may impact the generalizability of our findings, and it is important to note that the study was not originally designed to evaluate CKM performance.

In conclusion, CKM captured BHB dynamics during exogenous ketosis but revealed a gradual decline in day-to-day BHB AUC over 14 days in both KE and PBO groups, likely reflecting sensor drift rather than physiological adaptation. Larger controlled studies with direct comparison of CKM and blood BHB measurements are needed to confirm accuracy and clinical utility, and must include more than a single batch of CKM devices.

Acknowledgments

The authors express their gratitude to laboratory technologist Trine Larsen at the Department of Clinical Biochemistry in Viborg Regional Hospital, Denmark, for her invaluable technical support.

Funding

The study received financial support from undisclosed research resources within the Medical Diagnostics Center at Regional Hospital Central Jutland, Viborg. Additionally, it was funded by research grants from the Research Funds Pool at Regional Hospital Central Jutland and the Clinical Academic Group for Multimorbidity at the same hospital.

Data Availability

The datasets generated and analyzed in this study are not publicly available because of participant confidentiality and institutional policy restrictions. However, access to the data may be granted upon reasonable request from the corresponding author, subject to the necessary approvals and agreements to ensure data security and adherence to ethical guidelines.

Conflicts of Interest

None declared.

References

1. Moonla C, Reynoso M, Casanova A, et al. Continuous ketone monitoring via wearable microneedle patch platform. ACS Sens 2024 Feb 23;9(2):1004-1013. [doi: 10.1021/acssensors.3c02677] [Medline: 38300831]
2. Kong YW, Morrison D, Lu JC, Lee MH, Jenkins AJ, O'Neal DN. Continuous ketone monitoring: exciting implications for clinical practice. Diabetes Obes Metab 2024 Dec;26 Suppl 7(Suppl 7):47-58. [doi: 10.1111/dom.15921] [Medline: 39314201]
3. Møller N. Ketone body, 3-hydroxybutyrate: minor metabolite - major medical manifestations. J Clin Endocrinol Metab 2020 Sep 1;105(9):dgaa370. [doi: 10.1210/clinem/dgaa370] [Medline: 32525972]
4. Nelson AB, Queathem ED, Puchalska P, Crawford PA. Metabolic messengers: ketone bodies. Nat Metab 2023 Dec;5(12):2062-2074. [doi: 10.1038/s42255-023-00935-3] [Medline: 38092961]
5. Robberechts R, Poffé C. Defining ketone supplementation: the evolving evidence for postexercise ketone supplementation to improve recovery and adaptation to exercise. Am J Physiol Cell Physiol 2024 Jan 1;326(1):C143-C160. [doi: 10.1152/ajpcell.00485.2023] [Medline: 37982172]
6. Berg-Hansen K, Gopalasingam N, Christensen KH, et al. Cardiovascular effects of oral ketone ester treatment in patients with heart failure with reduced ejection fraction: a randomized, controlled, double-blind trial. Circulation 2024 May 7;149(19):1474-1489. [doi: 10.1161/CIRCULATIONAHA.123.067971] [Medline: 38533643]
7. Gopalasingam N, Berg-Hansen K, Christensen KH, et al. Randomized crossover trial of 2-week ketone ester treatment in patients with type 2 diabetes and heart failure with preserved ejection fraction. Circulation 2024 Nov 12;150(20):1570-1583. [doi: 10.1161/CIRCULATIONAHA.124.069732] [Medline: 39162035]
8. Teymourian H, Barfidokht A, Wang J. Electrochemical glucose sensors in diabetes management: an updated review (2010-2020). Chem Soc Rev 2020 Nov 7;49(21):7671-7709. [doi: 10.1039/d0cs00304b] [Medline: 33020790]
9. Alva S, Castorino K, Cho H, Ou J. Feasibility of continuous ketone monitoring in subcutaneous tissue using a ketone sensor. J Diabetes Sci Technol 2021 Jul;15(4):768-774. [doi: 10.1177/19322968211008185] [Medline: 33832353]
10. Abbott's biowearable: one sensor for glucose, ketones. Abbott.com. 2022. URL: https://www.abbott.com/corpnewsroom/strategy-and-strength/abbotts-biowearable-one-sensor-for-glucose-ketones.html [accessed 2025-02-03]

Abbreviations

AUC: area under the curve
KE: ketone ester
BHB: β-hydroxybutyrate
CKM: Continuous ketone monitoring
PBO: placebo
POCT: point-of-care testing

Edited by S Li; submitted 09.Oct.2025; peer-reviewed by D O'Neal, Y Kong; revised version received 22.Dec.2025; accepted 24.Dec.2025; published 13.Jan.2026.

Please cite as:
Kjær SK, Ridder LOR, Svart M, Rittig N, Bjerg LN, Sandfeld-Paulsen B, Thomsen HH
Continuous Ketone Monitoring: Data From a Randomized Controlled Trial
JMIR Diabetes 2026;11:e85548
URL: https://diabetes.jmir.org/2026/1/e85548
doi:10.2196/85548

Cultural and Technological Barriers to Telehealth Adoption for Type 2 Diabetes Management Among Asian American Patients: Qualitative Case Study

Devi Gurung States[1,2], MS, MSW, MPH, DHSc, PhD

[1]Himalayan Family Healthcare Project (HFHCP), 4145 Manchester Ave, St. Louis, MO, United States

[2]Peabody College of Education and Human Development, Vanderbilt University, Nashville, TN, United States

Corresponding Author:
Devi Gurung States, MS, MSW, MPH, DHSc, PhD
Himalayan Family Healthcare Project (HFHCP), 4145 Manchester Ave, St. Louis, MO, United States

Abstract

Background: In the past decade, telehealth has transformed health care delivery by allowing patients more rapid and convenient access to necessary care without the cost and logistical challenges of traveling to a health care facility. Telehealth services can benefit patients with type 2 diabetes mellitus (T2DM) amid a growing epidemic of T2DM in the United States that affects people of all ages and races. In 2020, 33 million people were diagnosed with this chronic disease, with the number expected to rise by 50% by 2040. Telehealth facilitates regular contact between patients and their providers, especially when there are geographic barriers and time constraints prohibiting physical interaction, at little or no added cost to the patient and at their convenience.

Objective: This study examines cultural and technological barriers affecting telehealth adoption among Asian American people with T2DM.

Methods: A qualitative case study approach was employed, utilizing semistructured interviews with 30 Asian American individuals in Missouri. Thematic analysis was used to identify key barriers.

Results: Four major barriers emerged: (1) language and cultural barriers—limited availability of translated materials and interpreters; (2) limited digital literacy and access—older adults and individuals with low technological exposure struggled with telehealth platforms; (3) limited provider recommendations—health care providers did not actively endorse telehealth, reducing patient awareness of telehealth as an option; and (4) technology access and infrastructure disparities—low-income participants faced challenges with the costs of and access to broadband and telehealth-compatible devices.

Conclusions: Addressing cultural and technological barriers is crucial to increasing telehealth adoption among Asian American people with T2DM. Culturally tailored interventions, provider engagement, and digital literacy programs should be prioritized. Policy efforts must focus on expanding broadband access and providing multilingual telehealth resources.

(JMIR Diabetes 2026;11:e75689) doi:10.2196/75689

KEYWORDS

Asian American; type 2 diabetes; telehealth; digital literacy; cultural barriers; health disparities

Introduction

Telehealth refers to a tool, process, or system that can provide patients with a simpler appointment scheduling process, remote access to clinical services that lessen the need for travel, and increased interaction with health care providers to improve medical outcomes while reducing health care costs [1,2]. In the past decade, telehealth has transformed health care delivery by allowing patients more rapid and convenient access to necessary care without the cost and logistical challenges of traveling to a health care facility [3]. Telehealth services can benefit patients with type 2 diabetes mellitus (T2DM) amidst a growing epidemic of T2DM in the United States that affects people of all ages and races [4]. In 2020, 33 million people were diagnosed with this chronic disease, with the number expected to rise by 50% by 2040 [4].

Telehealth facilitates regular contact between patients and their providers, especially when geographic barriers and time constraints prohibit physical interaction, at little or no added cost to the patient and at their convenience [5]. Telehealth services have the potential to optimize T2DM management by empowering patients to engage in self-care to slow disease progression, prevent complications, and lessen the health care burden [4,6]. Self-management among patients with T2DM includes proper nutrition, adequate physical activity, regular blood glucose monitoring, medication compliance, disease knowledge, lifestyle modifications, and self-efficacy [7]. To ensure effective self-management among patients at high risk or with a diagnosis of T2DM, telehealth can facilitate digital health coaching for long-term management or prevention of T2DM across population subgroups [8]. Asian American people include multiple subgroups (eg, Chinese, Indian, Vietnamese, Filipino, Korean, and Nepali). Cultural differences among these groups are substantial, especially among first-generation individuals.

According to the US Department of Health and Human Services [9], Asian American people are 40% more likely to be diagnosed with diabetes than non-Hispanic White individuals. Despite this higher prevalence, telehealth utilization among Asian American people remains low, as this population is disproportionately underserved [6,10]. Several factors contribute to this disparity, including cultural norms, digital literacy, provider engagement, and technological access [11]. Many Asian American communities face language barriers, limiting their ability to navigate telehealth platforms effectively [12,13], and older adults may struggle with technological proficiency, creating challenges in virtual health care engagement.

Given the disproportionate burden of T2DM among Asian American people and their persistently low telehealth usage, this study sought to answer the following questions: what cultural and technological barriers limit adoption? and what strategies might address them? These questions are urgent given the stakes, including T2DM complications, loss of care access, and widening racial health disparities. By centering user experience and structural constraints, the study identifies critical leverage points for more equitable digital health policy and practice.

Understanding these barriers is critical for designing equitable telehealth interventions that improve access to diabetes care. Therefore, this study explored the cultural and technological challenges hindering telehealth adoption among Asian American people with T2DM, providing insights for health care providers and policymakers. The guiding theoretical framework was the unified theory of acceptance and use of technology (UTAUT), an integrated approach used to predict and understand how and why individuals or groups accept or reject various technologies [11]. In the context of this inductive study, UTAUT provided a foundation for examining telehealth adoption among Asian American people with T2DM and was used to scaffold the study. UTAUT was used to identify several characteristics that influence the adoption of health technologies, which included social influence, performance expectancy, effort expectancy, facilitating conditions, privacy risk, and the threat of the disease from which the patient is ailing. By investigating factors affecting telehealth use within this population, specifically for the management of T2DM, a greater understanding of barriers to adoption of telehealth services and associated effects was achieved. Depending on the applicability of this study's findings to specific Asian American communities, this knowledge can be applied to support the development of more effective methods of tailoring telehealth services to meet the unique needs of Asian American people, as well as those of other underserved populations. This study provided insights regarding ways to incorporate culturally relevant telehealth approaches to T2DM management into mainstream health care practice.

In addition to the UTAUT framework, this study was informed by a critical health equity lens that emphasizes how structural inequities—such as limited English proficiency, systemic underinvestment in minority-serving institutions, and digital exclusion—shape health care access. This critical perspective extends UTAUT by interrogating not only users' perceptions of telehealth but also the sociotechnical structures that enable or constrain its use among marginalized populations. By situating Asian American T2DM management within broader systems of racialized health care access, this study provides an intersectional understanding of telehealth adoption.

Methods

Study Design and Researcher Perspective

A qualitative case study approach was used to explore telehealth adoption barriers among Asian American people with T2DM in Missouri. The purpose of the study was to examine Asian American people's perspectives to determine the factors that influence their adoption of telehealth for T2DM management. Regarding positionality, my interest in this demographic is that I am an Asian American. Subjective perspectives, rather than objective facts, were elicited from participants; therefore, a qualitative methodology was appropriate, as it enabled in-depth examination [14].

The study focused on Asian American people with T2DM in Missouri due to the group's elevated diabetes risk and underrepresentation in telehealth research. Missouri, as a Midwestern state with rising Asian American populations [15], offered a novel geographic and demographic context. A qualitative case study design was used to capture the complexity of individual, cultural, and technological factors shaping telehealth use, particularly within a population that often faces intersecting language and access barriers. This approach aligns with case study methodology's strengths in revealing nuanced insights within specific, bounded systems.

The study population included Asian American people with T2DM via convenience sampling [16]. Convenience sampling may overrepresent individuals with similar socioeconomic status or community ties and introduce selection bias. Inclusion criteria specified that participants be Asian American, 21 years or older, and diagnosed with T2DM. Those with type 1 diabetes mellitus and gestational diabetes were excluded, along with anyone whose health condition was not under the management of a health care provider. Previous experience with telehealth was not necessary.

Participants were recruited through community health organizations. Patients receiving care from home care agencies in Missouri, being under the direct supervision of health care providers, were asked to volunteer for the study. Participants were categorized into 2 groups: adopters (those who had adopted telehealth for self-management) and nonadopters (those who had not adopted telehealth for self-management).

Ethical Considerations

This study was reviewed and approved by the Capitol Technology University Institutional Review Board (approval IRB05242023a, approved on June 8, 2023). All participants provided informed consent prior to participation. Privacy and confidentiality were protected by institutional review board approval; data were deidentified and securely stored. No compensation was provided.

Data Collection

A pilot study was conducted to establish the reliability of the semistructured interview, with 10 participants selected from the study population. These participants were then excluded from the main study. Data collected during the pilot study were scrutinized to determine the instrument's capacity to collect data relevant and applicable to the study aims. The findings of the pilot study were used to review the data collection instrument and processes and implement any necessary instrument modifications to enhance its reliability.

Data were collected via semistructured interviews [17] with an open-ended question format. The following are sample questions:

1. (AU) Have you used telehealth services for diabetes management before? If yes, could you describe your experience with using telehealth? If not, can you explain the reasons for your decision not to use it?
2. (PE) How do you think telehealth can be useful for managing type 2 diabetes? Please explain.
3. (PE) What are the benefits of using telehealth for diabetes management?

Using a flexible interview schedule, participants engaged in these interviews through phone and video calls, which were digitally audio-recorded. Probes, follow-up questions, and comments were used to encourage participants to clarify statements where further information was required for a comprehensive understanding of experiences they described [17]. Through phone conversations and video calls, participants shared their experiences, challenges, and perceptions regarding the use of telehealth, especially those associated with management of T2DM.

Data Demographics

Data gathered during the semistructured interviews included participants' demographic information: race, sex, household size, household income, occupation, education level, and age. Data were collected on performance expectancy, indicating perspectives on their expectations of how telehealth might be helpful for the management of T2DM and the realization of health goals. Furthermore, data on effort expectancy were gathered, describing the participants' perceptions of the ease of use of telehealth and factors they believed would affect the ease of use negatively and positively. Finally, data on the social and cultural influences on the use of telehealth for managing T2DM were collected, examining participants' social networks (eg, family members, friends, and health care providers) and cultural factors (eg, preference for in-person care and language barriers).

Data regarding participants' personal innovativeness, trust, behavioral intention, and actual use of telehealth services were also gathered. Personal innovativeness pertains to their comfort levels in using new technologies such as telehealth, while trust refers to their level of confidence in the security and privacy offered by telehealth applications and whether they had reservations about sharing their health information online. Behavioral intentions examined whether the participants would adopt telehealth for the management of T2DM and the factors that may influence those decisions. Finally, participants' actual use of telehealth in the past was explored, with data collected on their associated experiences. For participants who did not have previous experience using telehealth, data were collected on their willingness to use it in the future.

Data Analysis

Thematic analysis was employed to identify recurring patterns in participant responses. Qualitative data were categorized based on patterns that form specific themes; data excerpts were thereby organized within the concepts outlined in the UTAUT theoretical framework [18-20]. We first conducted line-by-line inductive coding of all transcripts to identify emergent concepts. These codes were grouped deductively under UTAUT constructs and refined through axial coding into subthemes. Finally, overlapping subthemes were consolidated into 4 primary barrier categories. Qualitative data from the interviews were examined using axial coding, enabling codes, subcategories, and categories contained in the participants' perspectives to be more easily identified [19,21].

Results

Thematic refinement yielded 4 consolidated barriers to telehealth adoption (language and cultural barriers, digital literacy and access, limited provider recommendations, and technology/infrastructure disparities), derived from 9 initial themes.

Data Demographics

A total of 30 participants were recruited for the study: 15 adopters and 15 nonadopters to telehealth services. The 15 adopters were between the ages of 31 and 57, with the average age being 35 years. All of the adopters had attended institutions of higher education, with 6 having bachelor's degrees, 4 having master's degrees, and 2 having doctorates. All of the adopters were Asian American individuals, with multiple ethnicities represented, including Chinese, Indian, Vietnamese, Filipino, Korean, and Nepali. Men were disproportionately represented, with 12 men compared to only 3 women. The highest household income was US $200,000, while the lowest was US $36,000, with most households consisting of 3 people. Participants who were telehealth adopters also worked in various professions, including information technology, academia, and engineering, as well as in miscellaneous jobs (Table 1).

Table . Demographics of the adopters.

Adopter No.	Household size	Age (y)	Ethnicity	Gender	Household income (US $)	Occupation	Education level
1	4	45	Nepali	Male	200,000	IT professional	Master's
2	4	39	Chinese	Male	90,000	Auto mechanic	Bachelor's
3	5	38	Chinese	Male	150,000	Researcher	Bachelor's
4	3	35	Indian	Male	80,000	IT professional	Bachelor's
5	4	51	Korean	Female	100,000	Computer science	Bachelor's
6	3	57	Korean	Male	150,000	IT networking	College
7	4	44	Vietnamese	Male	160,000	Professor	PhD
8	1	36	Filipino	Male	36,000	Sushi cook	College
9	4	46	Vietnamese	Female	120,000	Hair stylist	College
10	3	44	Chinese	Female	110,000	Academia	PhD
11	3	31	Vietnamese	Male	50,000	IT	Bachelor's
12	2	35	Filipino	Male	75,000	Retired army personnel	Bachelor's
13	4	49	Nepali	Male	175,000	Aircraft engineer	Master's
14	3	45	Indian	Male	150,000	Civil engineer	Master's
15	3	35	Indian	Male	108,000	Data engineer	Master's

The 15 nonadopters of telehealth services were between the ages of 32 and 80, with the average age being 55 years. Participants in this group were less educated in comparison to the adopter group, with only 1 having a graduate degree. Five nonadopters graduated high school, while 9 had no formal education. However, as with the adopter group, there were various Asian American ethnicities represented in the group of nonadopters, including Nepali, Indian, Chinese, Filipino, Vietnamese, and Korean. There was less gender disparity among nonadopters compared to adopters, with 9 females and 6 males. The nonadopters earned much less than the adopters in terms of household income, which was between US $9000 and US $36,000. Most households had 2 people. Finally, 10 of the nonadopters were unemployed, while only 3 worked part-time (see Table 2).

Table . Demographics of nonadopters.

Nonadopter No.	Household size	Age (y)	Ethnicity	Gender	Household income (US $)	Occupation	Education level
1	1	80	Korean	Female	9000	Unemployed	High school
2	2	45	Korean	Female	12,000	Casual labor	High school
3	2	75	Chinese	Female	9000	Casual labor	High school
4	1	67	Indian	Male	9000	Unemployed	High school
5	2	62	Filipino	Female	11,400	Unemployed	No education
6	3	55	Nepali	Female	10,000	Unemployed	No education
7	3	67	Nepali	Male	11,000	Unemployed	No education
8	2	53	Indian	Female	12,000	Unemployed	No education
9	2	61	Vietnamese	Male	10,000	Unemployed	No education
10	2	38	Nepali	Male	15,000	Casual labor	No education
11	2	55	Vietnamese	Male	11,000	Unemployed	No education
12	3	32	Chinese	Female	36,000	Student	Graduate
13	2	36	Filipino	Male	36,000	Cook	High school
14	2	58	Indian	Female	12,000	Unemployed	No education
15	2	52	Nepali	Female	12,000	Unemployed	No education

Though past researchers have found facilitators to incorporating telehealth, this study only yielded barriers. To review, facilitators of telehealth implementation were videoconferencing, caregiver engagement, and delivery via the favored language of patients and caregivers [22]. Approaches to enhance telehealth consultations included in-person meetings to establish a relationship before shifting to telehealth and using text and audio telemonitoring to ensure that patients understood advice and instructions [22].

The interpretation following thematic analysis enabled the identification of 4 barriers to telehealth adoption among Asian American individuals with T2DM: (1) language and cultural barriers; (2) digital literacy and access; (3) limited provider recommendations; and (4) technology and infrastructure disparities. Beyond barriers, participants highlighted facilitators that can inform targeted implementation strategies, including provider recommendations, interpreter support, device access, insurance coverage, and perceived convenience.

Language and Cultural Barriers

Nonadopters reported several problems and challenges with telehealth systems, with the most prevalent problem being a language barrier. Due to the lower level of education among nonadopters and most being first-generation immigrants to the United States, many could not speak English. The language barrier was an obstacle unless they found a doctor with whom they shared a common language. In fact, 10 of the 15 nonadopters spoke of the language barrier as the main challenge affecting their use of telehealth.

Digital Literacy and Access

Effort expectancy involves the convenience and usability levels that adopters of telehealth experience when using the system [23,24]. Participants' perceptions of the ease or difficulty of use of telehealth and any associated problems were explored, and a learning curve associated with initial use of telehealth services was identified.

Several adopters noted that there was a steep learning curve and they initially experienced significant difficulties. However, they quickly added that after using telehealth several times and/or having a doctor or a proficient family member explain its use, it eventually became easier to navigate. Some adopters, such as information technology professionals, did not have any problems using the telehealth system. Nonetheless, adopters mentioned some problems with the use of telehealth, such as not having access to the internet. Some telehealth systems are more complex than a simple call to the doctor and require the use of mobile apps accessible only through a smartphone connected to the internet. This issue affects accessibility, especially if the telehealth appointment is scheduled for a time when internet access is limited or not possible due to travel or other factors. Another problem associated with the use of telehealth relates to the availability of supporting technology, such as smartphones or other devices capable of complex operations (eg, camera-enabled desktop computers or laptops). Lack of these technologies represents a significant barrier to using telehealth.

Limited Provider Recommendations

Though provider recommendations were limited, some participants did receive recommendations. Some adopters were forced by circumstances to use telehealth to aid in managing their diabetes, as was the case for P15:

> *I was able to consult with a doctor over the phone to discuss my diabetes. I think useful; you can actually ask questions about your diabetes issues you have at any time.*

Some of those circumstances included constraints on health care access due to the COVID-19 pandemic, like P5, who said:

> *My doctor also asked me to use telehealth. My doctor always communicates with me via text about my health issues. I think my doctor's advice influenced me quite a bit. Because a doctor is a medical doctor, she knows what she's doing. And she was strongly recommended, especially during COVID-19 time.*

Although they were required to use telehealth, the positive experience of doing so encouraged them to continue after the pandemic.

Technology and Infrastructure Disparities

None of the nonadopters had ever used telehealth to assist them in their management of T2DM. There were various reasons given for nonuse, ranging from a lack of awareness of the existence of telehealth systems to not knowing how to use the system, like P2:

> *I think training about telehealth and how to use it.*

However, some nonadopters also mentioned facing language barriers, which prevented them from using telehealth, like P7:

> *No one said anything because of my language barrier.*

Performance Expectancy

Performance expectancy is the degree to which a user believes that using telehealth will help them make gains in managing their health, for this study, T2DM specifically. Performance expectancy was assessed through examining participants' thoughts on how using telehealth could help with their T2DM, as well as the perceived benefits of using telehealth.

Many adopters thought telehealth would introduce convenience and flexibility into the management of their T2DM because it eliminated travel to the doctor's office. In addition, cost-effectiveness and time saved were mentioned by adopters as expected benefits of using telehealth. For example, adopters felt its use was cost-effective, saving them a trip to the doctor—and the associated expenses—while its capacity to provide secure and rapid access to health services ultimately saved valuable time. Adopters reported that using telehealth enhanced their access to health care services.

For nonadopters, performance expectancy was significantly lower compared to adopters. Specifically, there was a significant number of nonadopters who were not aware that telehealth might enhance diabetes management, perhaps because some were not familiar with telehealth systems. Nonetheless, many believed that using telehealth would be beneficial in helping them to manage their T2DM more effectively. Furthermore, when asked about the benefits of using telehealth, several nonadopters highlighted the convenience of not having to travel to the doctor's office, reduced cost, and time efficiency as likely benefits. These benefits were similar to those mentioned by adopters. However, some nonadopters did not perceive that there would be any benefits associated with the use of telehealth, and their overall preference for in-person care was considerable. See Table 3 for adopters' views on performance expectancy.

Table . Selected quotes by adopters about performance expectancy.

Participant	Adopter/nonadopter	Selected quotes
1	Adopter	But you still have the benefit of using telehealth. For example, it's flexible because you do not have to travel to your doctor's office, saving time.
3	Adopter	It is beneficial because I can call my doctor from anywhere. It is a benefit because I cannot travel and receive timely care for my problems.
5	Adopter	It was pretty seamlessly easy, I thought. Because I asked for the appointment, they asked me to fill out the information online, which I did pretty quickly. And then, about an hour later, they said they would call me back with the doctor. And so I got connected with the doctor, and I was able to consult with a doctor over the phone to discuss my diabetes. I think useful; you can actually ask questions about your diabetes issues you have at any time. Even though you're not in the town, because sometimes I'm not in the United States. I am in Korea, visiting my parents or my friends in Korea I'm there for maybe three months or four months. I can just call or email my doctor. I can always connect with my doctor online or on the phone about my diabetes, and I don't have any problem with that.
2	Adopter	Like I said, it's convenient because I can call my doctor anywhere from a distance. It helps with remote monitoring and consultation of blood sugar levels. It is cost-saving because the gas price is going too high.
15	Adopter	I am very busy, and I do not have to drive to the doctor's office. So it saves money and time. It's a convenience.
13	Adopter	Like I said, using telehealth is helping me access the doctor faster. For example, it takes more than two months to get an appointment for the office visit, but it takes less than one week to make a telehealth appointment.
8	Adopter	Besides my diabetes problems and other health problems, telehealth increases access to my care in terms of fast service.
11	Adopter	It does help me to monitor my blood sugar. Using telemedicine increases my chances of improving my diabetes. I don't have to always wait for your appointment.
12	Adopter	Honestly, it has been a lot more accessible, especially during the pandemic time. I can get a telehealth appointment within a week or so, but office visits take more than one month for an appointment.

Data Analysis

The following 9 themes were identified using thematic analysis: actual use, performance expectancy, effort expectancy, social influence, facilitating conditions, cultural influence, personal innovativeness, trust, and behavioral intention.

Actual Use

Information was collected on whether participants had used telehealth to talk to their physicians about diabetes, their experiences, and, among those who reported not having used telehealth, the reasons for not using it. Findings showed that there was a high usage of telehealth among adopters, who felt that telehealth was beneficial. For example, P13 said:

> *Like I said, using telehealth is helping me access the doctor faster. For example, it takes more than two months to get an appointment for the office visit, but it takes less than one week to make a telehealth appointment.*

Similarly, P15 said:

> *I am very busy, and I do not have to drive to the doctor's office. So it saves money and time. It's a convenience.*

This group reported positive experiences that centered on the benefits of telehealth, including its convenience. Table 4 shows adopters' views on actual use.

See Table 5 for a summary of nonadopters' views on actual use.

Table . Selected quotes by adopters about actual use.

Participant	Adopter/nonadopter	Selected quotes
1	Adopter	I've used telehealth before for my diabetes management. I feel that telehealth is beneficial. I think it's more convenient, easy to use, and as you start using it more and more.
4	Adopter	I have talked to my doctor by phone about my diabetes care. One thing I like about that is I don't have to wait a month for the doctor. I can call them whenever I need to and discuss it with the doctor. It's convenient that I do not have to wait too long.
11	Adopter	I have talked to my doctor before. Usually, either be on the phone or a Zoom call. It was easy for me after COVID-19; everything was on lockdown. So there is no choice.
10	Adopter	I have been using telehealth since the COVID-19 pandemic. Because I still feel the risk of exposure to the virus

Table . Selected quotes by nonadopters about actual use.

Participant	Adopter/nonadopter	Selected quotes
1	Nonadopter	I don't know much about it. I never talked to my doctor using the phone or anything because I did not know there was a telehealth service available. I have no idea. My doctor did not say anything to me.
6	Nonadopter	I do not use it because of language barrier.
5	Nonadopter	I do not know how to use it. I never knew about it.

Social Influence

Social influence is the degree to which actors in a person's social circles, such as the primary physician, family members, and friends, influence the decision of an individual to use telehealth. Participants were asked whether their doctors, friends, or family members encouraged the use of telehealth; they were then asked to describe the nature of that input.

Among adopters, the social influence of doctors and family members impacted their decision to start using telehealth. As doctors and family members communicated the benefits of using telehealth services, participants were led to consider it. When doctors described the benefits of convenience, cost-effectiveness, and time efficiency, they were influential in the participants' decision. In addition to doctors, adopters indicated that family members also had an important role in encouraging them to adopt telehealth. Family members of participants indicated that there were tremendous benefits of using telehealth for managing diabetes and described it as an effective health care service that is gaining traction for treating this chronic illness.

In contrast, nonadopters did not receive advice from their doctors, friends, or family members regarding the use of telehealth. Some of the nonadopters believed that their nonproficiency in English was part of the reason their doctors never recommended telehealth. Others stated that their physicians did not suggest it.

Facilitating Conditions

Facilitating conditions concerns the user's belief that the technical infrastructure and other conditions necessary for the use of telehealth already exist [25]. Participants' perceptions regarding whether the necessary resources to facilitate telehealth use for diabetes care were in place as well as their perceptions regarding whether they had access to necessary devices were examined.

Regarding owning the necessary devices, all adopters reported that they owned a smartphone, allowing them to easily connect to the internet and run the mobile apps necessary to use telehealth effectively. In addition, insurance coverage was mentioned by adopters as a necessary support enabling the use of telehealth. Several adopters reported that their insurance did not cover the service and that they had to pay out-of-pocket. This finding highlighted the need for insurance coverage as one of the facilitating conditions to enable adopters to use telehealth services when needed for managing diabetes. Adopters also reported that they would like support in the form of training on effective use of telehealth in general, as well as for diabetes management, specifically. There was a perception that doctors and their staff could also benefit from such training.

Nonadopters faced conditions that did not facilitate their use of telehealth. For instance, a majority of nonadopters did not own the devices, such as smartphones or laptops, required for telehealth access. In addition, other nonadopters deemed the services of a translator to be an important support resource necessary for them to use telehealth services effectively. This finding was logical, considering that 10 of the 15 nonadopters identified language as a barrier limiting their use of telehealth. Furthermore, most nonadopters pointed out that financial health was an important aspect for them to consider in relation to telehealth services. Some reported social security as their sole source of income. Those who indicated that financial assistance was necessary to support their use of telehealth services reported they would use such aid to purchase a phone and pay for the services of an interpreter. In addition to financial aid, they reported a need for training on telehealth and how it applies to diabetes management. See Table 6 for a summary of nonadopters' views on facilitating conditions.

Table . Selected quotes by nonadopters about facilitating conditions.

Participant	Adopter/nonadopter	Selected quotes
5	Nonadopter	I have no smartphone or computer.
9	Nonadopter	I have no phone or computer.
7	Nonadopter	Need an interpreter like you and pay for an interpreter's services.
11	Nonadopter	Need help with an interpreter.
4	Nonadopter	I survive from my social security check.
8	Nonadopter	Financial help to by phone and an interpreter.
3	Nonadopter	Education about telehealth and training about how to use it.
2	Nonadopter	I think training about telehealth and how to use it.
13	Nonadopter	Training about telehealth in the community or clinic.

Cultural Influence

Although the study participants were from the Asian American community in general, there were multiple ethnicities represented. One cultural influence identified by adopters was a preference for in-person care. However, even though adopters preferred in-person care, a recommendation by a doctor to use telehealth overrode the cultural-based preference of some adopters. It was clear that the convenience of using telehealth had more influence on their decision regarding whether to adopt telehealth. In contrast, other adopters retained their preference for in-person care despite the benefits of telehealth. Language was not an issue for most adopters, especially among second-generation Asian American individuals. Likewise, language was not an issue for Asian American individuals who spoke the same non-English language as their doctor. Still, some adopters did experience language barriers that affected their effective use of telehealth services.

Most nonadopters were not proficient in English, and this adversely affected their ability to use telehealth systems. Therefore, among participants who chose not to use telehealth, the primary concern was the language barrier; several nonadopters specifically cited the language barrier as prohibitive. A preference for in-person care was also cited as part of the reason participants did not adopt telehealth services.

Personal Innovativeness

Personal innovativeness is related to participants' comfort in using new technologies and their willingness to learn how to use telehealth [26,27]. All adopters were comfortable with telehealth technology. Participant occupations were often related to comfort level with telehealth systems. As many adopters were professionals in fields where the use of different technologies was required, they were easily able to attain proficiency on different platforms. Other adopters decided they were comfortable with new technologies because they were already using telehealth.

Among adopters, there was an overwhelming willingness to learn how to use telehealth for diabetes care. As all adopters except one were already using telehealth, they were willing to commit themselves to learning about the functionality and use of the system in even more detail.

In contrast, many nonadopters were not comfortable with new technologies due to legacy challenges, such as the language barrier. Some nonadopters who could not read or write in English would undoubtedly experience challenges using telehealth. However, regardless of language concerns, nonadopters were willing to learn how to use the technology.

Trust

Adopters must trust in the security, privacy, and confidentiality of their health information stored in health care systems to continue using telehealth services. Most adopters reported that they were comfortable with the security and privacy of their health information, expressing their confidence as being attributable to Health Insurance Portability and Accountability Act policies. Nonetheless, some adopters had concerns regarding security, as they reported worrying about the safety of their identifying information. In addition, some adopters were concerned that others would see their sensitive health information, although there were a few who felt that they could trust their physicians to keep their information safe and protected.

Most nonadopters did not trust telehealth systems. Their concerns were related to sharing their health information online and the security and privacy of information disclosed while using the systems. However, several nonadopters did not have an opinion one way or the other, as they had never used telehealth systems.

Behavioral Intention

Behavioral intention relates to the decision to continue or discontinue use of telehealth and the reasons for this decision [28]. All adopters except one said they would continue to use telehealth systems. The reasons that the majority of participants decided to continue using telehealth services were attributed to the benefits it provided. The one adopter who was noncommittal mentioned safety and security concerns about health information stored in the telehealth systems. Despite such concerns, most adopters were willing to continue using telehealth as a modality for managing diabetes.

Even though nonadopters had never used telehealth systems and despite the significant challenges identified, they were positive in their intentions toward the use of telehealth. A majority of nonadopters were willing to learn more about telehealth if someone taught them. In fact, nonadopters attributed their behavioral intentions to the need to improve their diabetes care. Other nonadopters cited the benefits of telehealth such as convenience or saving time and money as the reasons for their desire to learn. However, a few nonadopters refused to use telehealth in the future, citing existing language barriers.

Discussion

Summary of Key Findings

The analysis of the emergent themes from the qualitative data yielded 9 themes. The themes were actual use, effort expectancy, performance expectancy, social influence, facilitating conditions, cultural influence, personal innovativeness, trust, and behavioral intentions. Several subthemes also emerged from the themes. They included the possession of access devices, language barrier, the convenience and ease of using telehealth systems, cost and time effectiveness resulting from using telehealth systems, the role of physicians in recommending the use of telehealth systems, the need for training and education on the use of telehealth systems, insurance coverage of telehealth services, and an intention to continue using or start using telehealth systems.

Regarding actual use, the findings showed that the adopters who used telehealth systems cited the convenience, cost and time efficiency, and flexibility as some of their reasons for using telehealth systems, corresponding to Hu et al’s [29] findings, showing that access to technology was a notable determinant in patients’ interest in mobile health interventions. The nonadopters attributed their lack of awareness of the existence of telehealth systems and how to use them as some of the reasons for not using telehealth systems for the management of T2DM. Regarding effort expectancy, the findings showed that the participants perceived the use of telehealth systems to be easy. Concerning performance expectancy, telehealth systems improved convenience, enhanced access to health care services, were cost-efficient, saved time, and enabled flexibility in access to services. Regarding social influence, the findings showed that physicians, friends, and family played a significant role in promoting the adoption of telehealth. Similarly, Mora and Golden [30] found that family strongly influences individuals’ diabetes management plans. Recommendations of telehealth from physicians, family, and friends were conspicuously absent among the nonadopters, while the adopters reported the influence of those recommendations on their decisions to adopt telehealth systems.

The facilitating conditions identified as being influential to the adoption of telehealth systems for the management of T2DM included access to devices, such as smartphones and computers. The coverage of telehealth services by insurance providers was another important facilitating condition. The other conditions were training on how to use telehealth systems and education on the use of telehealth systems for the management of T2DM. The findings also showed that cultural factors influenced the adoption of telehealth systems. The two most influential cultural factors were the language barrier and a preference for in-person care. Mora and Golden [30] also found that language barriers impacted diabetes management approaches.

Findings on personal innovativeness showed that there was an equal share of comfort and discomfort in using telehealth systems. The findings also showed that legacy challenges, such as the language barrier and lack of access to devices, affected the participants’ ability to use telehealth systems. Regarding trust, participants shared concerns regarding the privacy, security, safety, and confidentiality of the health information stored in telehealth systems. However, there was also trust in the professionalism of physicians and the safeguards provided by the Health Insurance Portability and Accountability Act policy. Concerning behavioral intentions, there was an overwhelming desire to continue using telehealth systems. Even those who had not used telehealth systems before were willing to use them, provided they were trained and offered financial aid through which to acquire access devices. The behavioral intentions of the participants were informed by the benefits of convenience, time savings, cost efficiency, and enhanced access to health care services.

The findings showed that cultural influences played a role in the participants' decision to adopt or not adopt telehealth systems for the management of T2DM. The most significant cultural factor influencing the decision not to use telehealth among the nonadopters was the language barrier. Research has shown that most of the telehealth platforms used in the United States use the English language [31]. The implications are that populations that are not fluent in English might find telehealth systems to be of limited benefit. This was certainly the case for many nonadopters. To address this challenge, the researcher recommends culturally-adapted telehealth systems that target underserved racial and ethnic minorities. Culturally adapted telehealth systems will not only address the language issue but also incorporate other cultural adaptations that would enhance the usability of telehealth systems for racial and ethnic minorities. Existing research supports this recommendation [32].

Barriers to Telehealth Adoption

The results of this qualitative case study indicate that language and cultural barriers significantly impact Asian American individuals' use of telehealth for managing T2DM. Language and cultural barriers have been recognized as key obstacles in health care access, often limiting patient engagement. For example, there is a limited availability of translated telehealth materials, and the scarcity of bilingual health care providers and interpreters makes it even more difficult for many groups to use this technology. When considering the culture of Asian American people, it also seems that there is a preference for in-person consultations with health care providers.

Another barrier to using telehealth is digital literacy and access, as many Asian American people do not have access to either technical support or patient education resources. Older adults especially seem to struggle with navigating telehealth platforms. Without proper training and support, these individuals remain excluded from telehealth-driven diabetes management.

The findings also highlight the limited recommendations of telehealth services by providers, which could be better promoted as a way for patients to improve their management of T2DM. Health care provider recommendations play a crucial role in shaping patient perceptions of telehealth. As the study uncovered, health care providers often do not actively recommend telehealth services, so many patients may not be aware of this option. Consequently, patients are also unaware of the many benefits of using telehealth services, which represents a significant barrier to its adoption. This highlights the need for provider engagement strategies to integrate telehealth into routine diabetes management.

Finally, technology and infrastructure disparities exacerbate other barriers to the use of telehealth services. Low-income individuals struggle with the cost of high-speed internet and smart devices, widening the gap of health care inequity [9]. For example, many patients, especially those with low income, are not able to obtain the devices (eg, smartphones and laptops) needed to access telehealth. Add internet connectivity issues and it is no surprise that many patients do not use telehealth. Addressing these disparities requires policy intervention that expands broadband access and subsidizes telehealth technology for underserved communities. Overall, the findings from this study align with existing literature on telehealth disparities among minority populations [11].

Limitations

Adopters were generally younger, more educated, and higher-income than nonadopters. These socioeconomic differences likely confound the observed adoption patterns. While our sample size precluded stratified or adjusted analyses, future studies should employ matched sampling or multivariable adjustment to disentangle cultural influences from socioeconomic status. Participants were recruited through two community-based health organizations in Missouri: A Federally Qualified Health Center and a local Asian-serving nonprofit clinic. Both provided limited interpreter support, which shaped recruitment feasibility and participant diversity.

The language barrier was a significant limitation during the collection of data. The participants were drawn from various ethnicities. Therefore, they had diverse native languages. Many were not proficient in English and did not share a common language with the researcher. The researcher relied on the services of an interpreter to translate the question to the participant and the response from the participant back to the researcher. The translation is prone to loss of meaning because the interpreter must interpret and decode the participant's words to derive their meaning. Context, whether personal or cultural, is important to the meaning of the participants' words. Furthermore, facial expressions, gestures, tone of voice, and pauses are also central to meaning. While the interpreter may translate the words said by the participant, an accurate translation of the meaning, semantics, and nuances behind the words may not always be possible. Therefore, some or the entire meaning of the communication may be lost during the translation.

The knowledge level of the nonadopters about telehealth systems is a limitation to the value of the data gathered from the cohort. Most of the participants were unaware of the existence of telehealth systems. Therefore, it is possible that they did not actually decide not to adopt telehealth for the management of their T2DM. The implications of this lack of awareness are that the information they provided may not have reflected the influence of the unique characteristics of telehealth on their nonadoption but rather an influence of their lack of awareness of the existence of telehealth systems and the value they provide. Lack of awareness of telehealth systems may explain responses to several prompts, most of which revolved around "I do not know." The value of this information in answering the research questions was limited.

Conclusions

This qualitative case study identified unique characteristics, supported by the UTAUT model, that influenced the adoption intent and adoption of telehealth among Asian American people with T2DM in Missouri. Overall, there were many valuable insights into the cultural and technological barriers facing Asian American people when using telehealth.

Conflicts of Interest

None declared.

References

1. Gordon HS, Solanki P, Bokhour BG, Gopal RK. I'm not feeling like I'm part of the conversation: patients' perspectives on communicating in clinical video telehealth visits. J Gen Intern Med 2020 Jun;35(6):1751-1758. [doi: 10.1007/s11606-020-05673-w] [Medline: 32016705]
2. Sim R, Lee SWH. Patient preference and satisfaction with the use of telemedicine for glycemic control in patients with type 2 diabetes: a review. Patient Prefer Adherence 2021;15:283-298. [doi: 10.2147/PPA.S271449] [Medline: 33603347]
3. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep 1;27(3):425-478. [doi: 10.2307/30036540]
4. De Groot J, Wu D, Flynn D, Robertson D, Grant G, Sun J. Efficacy of telemedicine on glycaemic control in patients with type 2 diabetes: a meta-analysis. World J Diabetes 2021 Feb 15;12(2):170-197. [doi: 10.4239/wjd.v12.i2.170] [Medline: 33594336]
5. de Kreutzenberg SV. Telemedicine for the clinical management of diabetes; implications and considerations after COVID-19 experience. High Blood Press Cardiovasc Prev 2022 Jul;29(4):319-326. [doi: 10.1007/s40292-022-00524-7] [Medline: 35579849]
6. Gal RL, Cohen NJ, Kruger D, et al. Diabetes telehealth solutions: improving self-management through remote initiation of continuous glucose monitoring. J Endocr Soc 2020 Sep 1;4(9):bvaa076. [doi: 10.1210/jendso/bvaa076] [Medline: 32864542]
7. Ernawati U, Wihastuti TA, Utami YW. Effectiveness of diabetes self-management education (DSME) in type 2 diabetes mellitus (T2DM) patients: systematic literature review. J Public Health Res 2021 Apr 14;10(2):2240. [doi: 10.4081/jphr.2021.2240] [Medline: 33855427]
8. Gershkowitz BD, Hillert CJ, Crotty BH. Digital coaching strategies to facilitate behavioral change in type 2 diabetes: a systematic review. J Clin Endocrinol Metab 2021 Mar 25;106(4):e1513-e1520. [doi: 10.1210/clinem/dgaa850] [Medline: 33206975]
9. Diabetes and Asian Americans. US Department of Health and Human Services: Office of Minority Health. 2021. URL: https://minorityhealth.hhs.gov/diabetes-and-asian-americans [accessed 2026-01-13]
10. Anjana RM, Pradeepa R, Deepa M, et al. Acceptability and utilization of newer technologies and effects on glycemic control in type 2 diabetes: lessons learned from lockdown. Diabetes Technol Ther 2020 Jul;22(7):527-534. [doi: 10.1089/dia.2020.0240] [Medline: 32522031]
11. Hoque R, Sorwar G. Understanding factors influencing the adoption of mHealth by the elderly: an extension of the UTAUT model. Int J Med Inform 2017 May;101:75-84. [doi: 10.1016/j.ijmedinf.2017.02.002] [Medline: 28347450]
12. Goh KLS, Lee CS, Koh CHG, et al. Evaluating the effectiveness and utility of a novel culturally-adapted telemonitoring system in improving the glycaemic control of Asians with type-2 diabetes mellitus: a mixed method study protocol. Trials 2021 Apr 26;22(1):305. [doi: 10.1186/s13063-021-05240-6] [Medline: 33902656]
13. Khalid A, Dong Q, Chuluunbaatar E, Haldane V, Durrani H, Wei X. Implementation science perspectives on implementing telemedicine interventions for hypertension or diabetes management: scoping review. J Med Internet Res 2023 Mar 14;25:e42134. [doi: 10.2196/42134] [Medline: 36917174]
14. Patton MQ. Qualitative Research & Evaluation Methods: Integrating Theory and Practice, 4th edition: Sage Publications; 2015. URL: https://discovered.ed.ac.uk/discovery/fulldisplay?vid=44UOE_INST:44UOE_VU2&search_scope=UoE&tab=Everything&docid=alma992387673302466&lang=en&context=L&isFrbr=true [accessed 2026-01-13]
15. American Community Survey data. US Census Bureau. 2022. URL: https://www.census.gov/programs-surveys/acs/data.html [accessed 2026-01-21]
16. Etikan I, Musa SA, Alkassim RS. Comparison of convenience sampling and purposive sampling. Am J Theor Appl Stat 2016;5(1):1. [doi: 10.11648/j.ajtas.20160501.11]
17. DeJonckheere M, Vaughn LM. Semistructured interviewing in primary care research: a balance of relationship and rigour. Fam Med Community Health 2019;7(2):e000057. [doi: 10.1136/fmch-2018-000057] [Medline: 32148704]
18. Guest G, MacQueen K, Namey E. Applied Thematic Analysis: Sage Publications; 2012. [doi: 10.4135/9781483384436]

19. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods 2017 Dec 1;16(1). [doi: 10.1177/1609406917733847]
20. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manage Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
21. Suki NM, Suki NM. Determining students' behavioural intention to use animation and storytelling applying the UTAUT model: the moderating roles of gender and experience level. Int J Manag Ed 2017 Nov;15(3):528-538. [doi: 10.1016/j.ijme.2017.10.002]
22. Truong M, Yeganeh L, Cook O, Crawford K, Wong P, Allen J. Using telehealth consultations for healthcare provision to patients from non-Indigenous racial/ethnic minorities: a systematic review. J Am Med Inform Assoc 2022 Apr 13;29(5):970-982. [doi: 10.1093/jamia/ocac015] [Medline: 35150266]
23. Batucan GB, Gonzales GG, Balbuena MG, Pasaol KRB, Seno DN, Gonzales RR. An extended UTAUT model to explain factors affecting online learning system amidst COVID-19 pandemic: the case of a developing economy. Front Artif Intell 2022;5:768831. [doi: 10.3389/frai.2022.768831] [Medline: 35573898]
24. Chao CM. Factors determining the behavioral intention to use mobile learning: an application and extension of the UTAUT model. Front Psychol 2019;10:1652. [doi: 10.3389/fpsyg.2019.01652] [Medline: 31379679]
25. Venkatesh V, Sykes TA, Zhang X. "Just what the doctor ordered": a revised UTAUT for EMR system adoption and use by doctors. Presented at: 44th Hawaii International Conference on System Sciences (HICSS 2011); Jan 4-7, 2011. [doi: 10.1109/HICSS.2011.1]
26. Bandyopadhyay K, Fraccastoro KA. The effect of culture on user acceptance of information technology. Commun Assoc Inf Syst 2007;19(19):23-29. [doi: 10.17705/1CAIS.01923]
27. Oh JC, Yoon SJ. Predicting the use of online information services based on a modified UTAUT model. Behav Inf Technol 2014 Jul 3;33(7):716-729. [doi: 10.1080/0144929X.2013.872187]
28. Twum KK, Ofori D, Keney G, Korang-Yeboah B. Using the UTAUT, personal innovativeness and perceived financial cost to examine student's intention to use e-learning. J Sci Technol Policy Manag 2022 Aug 3;13(3):713-737. [doi: 10.1108/JSTPM-12-2020-0168]
29. Hu L, Wyatt LC, Mohsin F, et al. Characterizing technology use and preferences for health communication in South Asian immigrants with prediabetes or diabetes: cross-sectional descriptive study. JMIR Form Res 2024 Apr 26;8:e52687. [doi: 10.2196/52687] [Medline: 38669062]
30. Mora N, Golden SH. Understanding cultural influences on dietary habits in Asian, Middle Eastern, and Latino patients with type 2 diabetes: a review of current literature and future directions. Curr Diab Rep 2017 Oct 23;17(12):126. [doi: 10.1007/s11892-017-0952-6] [Medline: 29063419]
31. Payán DD, Frehn JL, Garcia L, Tierney AA, Rodriguez HP. Telemedicine implementation and use in community health centers during COVID-19: clinic personnel and patient perspectives. SSM Qual Res Health 2022 Dec;2:100054. [doi: 10.1016/j.ssmqr.2022.100054] [Medline: 35169769]
32. Litchman ML, Kwan BM, Zittleman L, et al. A telehealth diabetes intervention for rural populations: protocol for a randomized controlled trial. JMIR Res Protoc 2022 Jun 14;11(6):e34255. [doi: 10.2196/34255] [Medline: 35700026]

Abbreviations

T2DM: type 2 diabetes mellitus
UTAUT: unified theory of acceptance and use of technology

Edited by S Li; submitted 08.Apr.2025; peer-reviewed by J Lin, Y Song; revised version received 04.Dec.2025; accepted 17.Dec.2025; published 03.Feb.2026.

Please cite as:
States DG
Cultural and Technological Barriers to Telehealth Adoption for Type 2 Diabetes Management Among Asian American Patients: Qualitative Case Study
JMIR Diabetes 2026;11:e75689
URL: https://diabetes.jmir.org/2026/1/e75689
doi:10.2196/75689

Content Validation of an Electronic Health Record–Based Diabetes Self-Management Support Tool for Older Adults With Type 2 Diabetes: Qualitative Study

Ploypun Narindrarangkura[1], MS, MD, PhD; Siroj Dejhansathit[2,3], MD; Uzma Khan[2,3], MD; Margaret Day[4], MSPH, MD; Suzanne A Boren[5,6], MHA, PhD; Eduardo J Simoes[5,7], MS, MPH, MD; Min Soon Kim[5,6], PhD

[1]Phramongkutklao College of Medicine, Bangkok, Thailand
[2]Cosmopolitan International Diabetes and Endocrinology Center, University of Missouri, Columbia, MO, United States
[3]Department of Medicine, University of Missouri, Columbia, MO, United States
[4]Department of Family and Community Medicine, University of Missouri, Columbia, MO, United States
[5]University of Missouri Institute for Data Science and Informatics, Columbia, MO, United States
[6]Department of Health Sciences, College of Health Sciences, University of Missouri, 310 Clark Hall, Columbia, MO, United States
[7]Department of Biomedical Informatics, Biostatistics and Medical Epidemiology, University of Missouri, Columbia, MO, United States

Corresponding Author:
Min Soon Kim, PhD
University of Missouri Institute for Data Science and Informatics, Columbia, MO, United States

Abstract

Background: Older adults with diabetes frequently access their electronic health record (EHR) notes but often report difficulty understanding medical jargon and nonspecific self-care instructions. To address this communication gap, we developed Support-Engage-Empower-Diabetes (SEE-Diabetes), a patient-centered, EHR-integrated diabetes self-management support tool designed to embed tailored educational statements within the assessment and plan section of clinical notes.

Objective: This study aimed to validate the clarity, relevance, and alignment of SEE-Diabetes content with the Association of Diabetes Care & Education Specialists 7 Self-Care Behaviors framework from the perspectives of older adults and clinicians.

Methods: An interdisciplinary team conducted expert reviews and qualitative interviews with 11 older adults with diabetes and 8 clinicians practicing in primary care (family medicine) and specialty diabetes care settings at a Midwestern academic health center. Patients evaluated the readability and relevance of the content, while clinicians assessed clarity, sufficiency, and potential clinical utility. Interview data were analyzed using inductive thematic analysis, and descriptive statistics were used to summarize participant characteristics.

Results: Patients (mean age 72, SD 4.9 y; mean diabetes duration 26, SD 15 y) reported that the SEE-Diabetes statements were clear, relevant, and written in plain language that supported understanding of self-care recommendations. Clinicians (mean 13, SD 9.5 y of diabetes care experience) viewed the content as concise, clinically appropriate, and well aligned with patient self-management goals and the Association of Diabetes Care & Education Specialists 7 Self-Care Behaviors framework. Both groups identified the tool's potential to enhance patient engagement and patient-clinician communication, while noting opportunities to improve the specificity of language, particularly within medication-related content.

Conclusions: SEE-Diabetes demonstrated content validity as a practical, patient-centered digital health tool for supporting diabetes self-management communication within EHR clinical notes. The findings support its use as a complementary approach to reinforce self-care communication in routine clinical practice and highlight areas for refinement to enhance personalization.

(JMIR Diabetes 2026;11:e83448) doi:10.2196/83448

KEYWORDS

diabetes mellitus, type 2; electronic health records; self-management; patient education; older adults; digital health; health literacy

Introduction

Background

Diabetes is highly prevalent among older adults in the United States, with an estimated 29.2% of adults aged 65 years or older having been diagnosed with or undiagnosed diabetes during 2017 - 2020, and approximately 48.8% of adults in this age group had prediabetes according to the most recent National Diabetes Statistics Report [1]. As the aging population grows, primary care clinicians face increasing pressure to deliver effective, individualized diabetes self-management education within routine visits. Diabetes self-management education and support (DSMES) has been shown to improve glycemic control, reduce complications, and enhance self-efficacy [2-5]. However, the delivery of DSMES in outpatient settings is frequently constrained by limited visit time, complex documentation requirements, challenges in referral and access, and poor integration with routine clinical workflows [6,7].

National DSMES standards outline 4 critical times when individuals with diabetes should receive structured education and support [2]; however, referrals and access to formal DSMES services remain inconsistent. As a result, self-management guidance is often delivered informally during routine visits, underscoring the need for tools that reinforce evidence-based messaging within existing clinical workflows.

To address these challenges, our team developed Support-Engage-Empower-Diabetes (SEE-Diabetes), a patient-centered educational aid designed to support clinicians in delivering tailored diabetes education to older adults during clinic visits. SEE-Diabetes integrates directly into the electronic health record (EHR) by embedding brief, personalized education statements—drawn from a curated content library—into the assessment and plan section of the clinician's note. The content is organized according to the 7 core domains of the Association of Diabetes Care & Education Specialists 7 Self-Care Behaviors (ADCES7), including healthy coping, healthy eating, being active, taking medication, monitoring, reducing risk, and problem solving [8].

Placement of SEE-Diabetes in the Assessment and Plan section was intentional. Prior formative research with older adults with diabetes from our group found that the majority (80%) accessed and read their clinic notes through patient portals, yet many found these notes difficult to understand due to medical jargon and vague or nonactionable self-care guidance [6,7]. Embedding clear, relevant, and actionable statements in a section that patients already read may therefore address an important communication gap while also integrating seamlessly into clinician documentation.

SEE-Diabetes was developed using a user-centered design (UCD) approach to ensure alignment with real-world clinical needs [9,10]. The first stage of development involved an analysis of EHR documentation patterns related to diabetes care [11], followed by a second stage comprising focus groups with older adults with type 2 diabetes and clinicians involved in diabetes management to identify gaps in the clarity, readability, and consistency of self-management information [6,7]. This study represents the third stage of the UCD process and focuses on content validation of the SEE-Diabetes educational statements to ensure their accuracy, relevance, and practical utility for both patients and clinicians [12].

Objective

Our objective was to assess the clarity, helpfulness, and perceived value of SEE-Diabetes education content by conducting in-depth interviews with older adults and clinicians practicing in primary care (family medicine) and specialty diabetes care settings. This validation step is essential before the broader implementation of SEE-Diabetes in primary care settings. By embedding actionable, comprehensible diabetes education into clinical notes, SEE-Diabetes may enhance patient understanding, improve continuity of care, and support more effective chronic disease management among older adults.

Methods

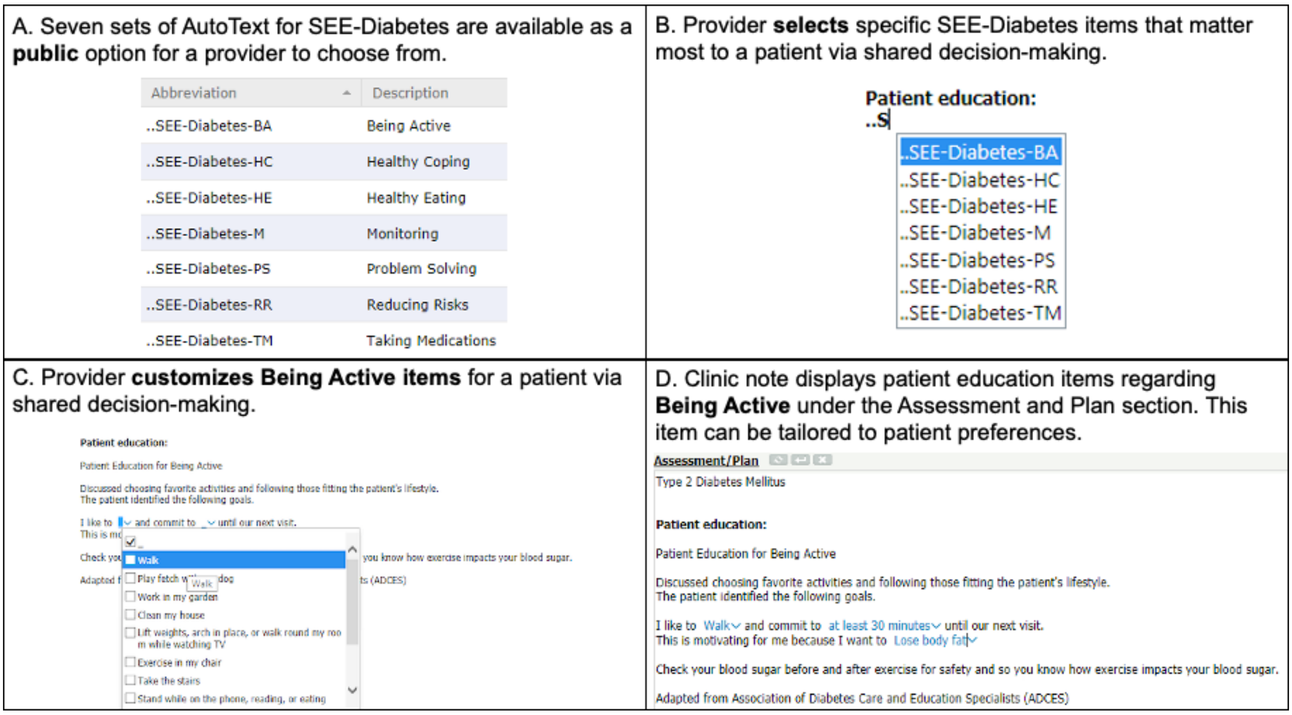
Study Design

We conducted a qualitative content validation study to assess the clarity, readability, and clinical relevance of SEE-Diabetes, an EHR-integrated education tool for older adults with diabetes. This phase represented the third stage of a UCD process. The content validation process included (1) expert reviews by clinicians and certified diabetes care and education specialists, and (2) user feedback through semistructured interviews with older adults with diabetes and with primary care or endocrinology clinicians. The interdisciplinary research team included experts in informatics, endocrinology, primary care, and diabetes education.

Description of SEE-Diabetes

SEE-Diabetes content was implemented within the EHR as "auto-text" templates in Oracle Cerner's PowerChart. During documentation, the clinician first selects the SEE-Diabetes category most relevant to the patient's needs, informed by shared decision-making during the visit. Within the category chosen, the clinician can review and select multiple educational statements addressing specific self-care behaviors. Each statement can be further customized to reflect the patient's individual preferences, goals, literacy level, and clinical circumstances. Examples of customization include changing the activity type (eg, walking and gardening) or specifying behavior targets (eg, number of minutes per day) (Figure 1).

Figure 1. Overview of the Support-Engage-Empower-Diabetes framework illustrating integration of tailored patient education statements into electronic health records, aligned with the Association of Diabetes Care & Education Specialists 7 Self-Care Behaviors. (A) Seven publicly available autotext sets (being active, healthy coping, healthy eating, monitoring, problem solving, reducing risks, and taking medications) are mapped to Association of Diabetes Care & Education Specialists 7 Self-Care Behavior domains. (B) The clinician selects the relevant Support-Engage-Empower-Diabetes category in the Patient education field. (C) Within the chosen category, statements are customized collaboratively (eg, activity type, frequency, or targets) during shared decision-making. (D) The finalized, tailored patient education statements are inserted into the Assessment and Plan section of the clinic note and become available to patients via the portal. ADCES7: Association of Diabetes Care & Education Specialists 7 Self-Care Behavior; SEE-Diabetes: Support-Engage-Empower-Diabetes.



The finalized patient education text was embedded within the assessment and plan section of the clinic note. Embedding SEE-Diabetes in the assessment and plan positions the guidance where patients already expect to find follow-up instructions, while requiring minimal change to clinician workflow. Insertion and customization generally take less than 1 minute, minimizing any disruption to the visit flow.

Study Setting

The study was conducted at the University of Missouri Health Care, an academic medical center serving 114 counties in Missouri [13]. The center uses the Oracle Cerner PowerChart EHR system to consolidate patient data across facilities. Patients can access their medical records, including clinic notes, through the HEALTHConnect portal. Clinic notes were retrieved from PowerChart, and patient recruitment was facilitated using PowerInsight, Oracle Cerner's operational reporting platform.

Interview Development

A total of 3 representative clinical scenarios were developed based on de-identified data from older adults with type 2 diabetes. For each case, SEE-Diabetes was applied to generate tailored patient education statements aligned with ADCES7 domains. The scenarios were (1) a 69-year-old woman with uncontrolled diabetes (monitoring and healthy eating), (2) a 72-year-old man with stable diabetes and obesity (medication adherence and risk reduction), and (3) a 67-year-old woman with type 2 diabetes (physical activity and healthy coping). An endocrinologist drafted the clinic notes (history of present illness and assessment and plan), and the multidisciplinary team reviewed all content for clinical accuracy and guideline concordance. The history of present illness sections are shown in Textbox 1.

Textbox 1. History of present illness section of three clinic notes. These are in-model screenshots of three example clinic notes designed for patients with diabetes aged 65 years and older attending follow-up visits at the Cosmopolitan International Diabetes and Endocrinology Center. Participants were asked to review the history of present illness section along with the assessment and plan.

Clinic Note 1: History of Present Illness

- 69-year-old lady presents for discussion regarding long-term management of diabetes mellitus.
- Initially diagnosed in 2009,
- started on metformin 1000 mg BID
- glimepiride 4 mg BID started in 2010
- pioglitazone 45 mg QD in the morning started in 2019
- has never been on insulin
- She has not had any diabetic education since her diagnosis. Denies numbness/tingling in her extremities. She has tried a keto diet in the past, but this led to frequent hypoglycemia. Since July 2022, she has been eating <1700 calories daily, which resulted in weight gain. Eye exam was done in April 2022 but not sure if her eyes were dilated. She has noted that her vision changes as her BG fluctuates, and sometimes her vision is blurry despite wearing her bifocals. There is no family history of T2DM, T1DM, or osteoporosis, and has never had a DEXA scan. She sees gynecologist yearly but does not have a regular well-woman exam.

Clinic Note 2: History of Present Illness

- This is a 72-year-old gentleman who presents for follow-up of his diabetes.
- He was diagnosed with diabetes around the age of 50 and has been on metformin since that time.
- Blood glucose is slightly worse; he checks every day and has been running above 150 mg/dl
- He is on Metformin 1000 mg twice a day. He has noted increased blood glucose since he got Covid in 1/2022.
- He has had fatigue, feels nauseous, so has not been taking his metformin daily, only taking it "on good days."
- He was supposed to meet with a dietitian, but has had so many doctors' appointments that did not make it.
- His family doctor added a "small pill every day," but he is not sure what medication it is.
- He does not want to use any medications that are injections at this time and feels he can control his diabetes once he is feeling better.
- His HbA1c$_{1c}$ was 7.5%; it has increased now to 8.8%.
- He plans on focusing on lifestyle and has been having increased burning in feet, so he was not walking much.

Clinic Note 3: History of present illness

- 67-year-old patient who presents to discuss diabetes mellitus type 2 management
- DMT2 diagnosed age 55 years. There is no retinopathy, neuropathy; she has microalbuminuria. She also has hyperlipidemia and hypertension
- Current regimen includes glipizide 10 mg daily and metformin extended release 500 mg, takes 1 tablet twice a day
- Her last diabetes class was before 2014
- Checks FSG once a day, ranges from 122-140s, no hypoglycemia
- She was walking, had to quit because of arthritis, now spends most of her time at home, and feels discouraged about her diabetes
- She likes to bake but has no motivation to do it anymore. Three friends have passed away in the last four years, and she has no family near home. She tries to eat healthy, mainly frozen meals.
- She is a smoker, has been trying to quit but feels she cannot do it.
- Blood pressure had been controlled on triamterene/HCTZ, 37.5/25 mg, and losartan 100 mg daily, but has increased and now also amlodipine 10 mg daily. For hyperlipidemia, takes pravastatin 10 mg daily.
- Takes ASA 81 mg daily.
- Last eye exam was on May 5, 2022, and showed no retinopathy. She has had cataract surgery also.
- Last urine microalbumin on October 14, 2022, showed microalbuminuria (high: 93)
- Denies numbness in feet or tingling, no foot ulcers
- No chest pain, palpitations, no nausea, vomiting, diarrhea, no abdominal pain, no cough or fever

Parallel semistructured interview guides were developed for patients and clinicians. Patient interviews consisted of 4 open-ended questions assessing readability, helpfulness, relevance, and anticipated future use of the SEE-Diabetes statements. Clinicians reviewed the same notes and answered 4 corresponding questions addressing clarity, completeness, clinical applicability, and suggestions for improvement. This mirrored design enabled direct comparison of perspectives across patient and clinician groups.

Data Collection and Analysis

Participants were recruited from Family and Community Medicine clinics and the Cosmopolitan International Diabetes and Endocrinology Center in October-November 2022. Participants were asked to evaluate the Patient Education section generated via SEE-Diabetes, which was included under the assessment and plan section of the 3 clinic notes for patients (Textbox 2).

Textbox 2. Assessment and plan section of three clinic notes. These are in-model screenshots of the Assessment and plan sections from three example clinic notes for patients with diabetes aged 65 years and older. The patient education sections were generated using Support-Engage-Empower-Diabetes, based on reviews of each patient, and then customized by an endocrinologist. Subsequently, they were reviewed by other team members. Participants were asked to review the Patient Education section and answer open-ended questions to assess the readability, helpfulness, and values of Support-Engage-Empower-Diabetes.

Clinic Note 1 : Assessment and Plan

1. Uncontrolled type 2 diabetes mellitus with hyperglycemia

- Reviewed lab results with patient emphasizing the importance of optimizing HbA_{1c}, with target below 8%
- Advised to check FSG regularly and record, bring records for review next visit
- Reviewed risks of hypoglycemia, prevention, and management of hypoglycemic episodes
- Reviewed foot care, call me if notice an open area on foot
- She will schedule an eye exam

Patient Education for Monitoring

- Monitoring is an important aspect of self-care. It helps you know if you are meeting recommended treatment goals to keep you healthy.
- My goal is to learn how to use my monitor, learn how to interpret my blood sugar levels
- I want to use this information to learn how different foods affect my blood sugar
- I commit to checking my blood sugar at the following times: 1 time a day and plan to bring in my readings to my next visit

Adapted from Association of Diabetes Care and Education Specialists (ADCES)

2. Obesity

- The patient is motivated to use weight control, which will improve metabolic health, including diabetes mellitus type 2, hypertension, and hyperlipidemia.

Patient Education for Healthy Eating

Discussed the meal plan today and the patient set the following goals:

- I will read the Nutrition Facts Label.
- I will add 2 servings of vegetables to my diet.
- I will cut down added sugar in my drinks from my diet to help to control my blood sugar.
- I plan to learn more about considering different healthy eating options by meeting with a diabetes specialist by the time of our next visit.

Adapted from Association of Diabetes Care and Education Specialists (ADCES)

Clinic Note 2: Assessment and Plan

1. Type 2 diabetes mellitus without complications

- Reviewed lab results with patient emphasizing the importance of optimizing HbA_{1c}, with target HbA_{1c} below 8 %
- Advised to check FSG regularly and record, bring records for review next visit,
- Reviewed risks of hypoglycemia, prevention, and management of hypoglycemic episodes
- Reviewed foot care, call me if notice an open area on foot

Patient Education for Taking Medications

- Taking medications helps lower your risk for heart attack, stroke, and kidney damage by managing blood glucose, blood pressure, and cholesterol levels in your body. The longer you have diabetes, the more help you will need from medications to keep you and your heart, eyes, and kidneys healthy.
- I plan to take my medications on time by bringing in all my medications to my next appointment between now and my next visit.

Adapted from Association of Diabetes Care and Education Specialists (ADCES)

2. Body mass index 40+ - severely obese (finding)

- Patient has started to feel somewhat better after his COVID infection and is motivated to increase activity and control his weight to improve management of his diabetes, hyperlipidemia.

Patient Education for Reducing Risks

- Reducing risks means doing behaviors that minimize or prevent complications and negative outcomes of prediabetes and diabetes. Risks mean doing behaviors that minimize or prevent complications and negative outcomes of prediabetes and diabetes.
- I plan to make positive lifestyle changes, participate in diabetes self-management education.
- I will do this by scheduling an appointment by the time of our next visit.

Adapted from the Association of Diabetes Care and Education Specialists (ADCES)

- Follow-up in clinic in 3 months with labs before the appointment
- Referral placed for diabetes education again

Clinic Note 3: Assessment and Plan

1. Diabetes Mellitus

- Detailed discussion with the patient, reviewed HbA_{1c} of 7.2%, her target is below 8% so she is doing well. HbA_{1c} of 7.2%, her target is below 8% so she is doing well.
- However, she has gained weight and is not feeling well.
- We discussed medications that might make her mood better; however, the patient wants to focus on positive thinking first.

Patient Education for Being Active

- Discussed choosing favorite activities and following those fitting the patient's lifestyle. The patient identified the following goals.
- I like to walk, park farther away from the door and commit to 10 minutes daily until our next visit. This is motivating for me because I want to improve mood
- Check your blood sugar before and after exercise for safety and so you know how exercise impacts your blood sugar.

Patient Education for Healthy Coping

- Discussed with patient that it is important to find healthy ways to cope and not to turn to harmful habits such as smoking, overeating, drinking or alcohol. This is especially true if you have diabetes. Having a lot of stress can increase blood glucose (sugar) levels, make you feel more negative and may lead to less healthy choices.
- I plan to cope with stress by make a list of people I can turn to for support and report back at my next visit to share how that went. I will observe/record my mood daily, I will seek help if I feel challenged.

Adapted from Association of Diabetes Care and Education Specialists (ADCES)

- She will continue her current medications, focus on lifestyle and I will see her back in 3 months with labs before the appointment. She will call if she needs to make an earlier appointment.

In-depth interviews were conducted in private settings and lasted approximately 30 minutes. Sessions were audio recorded, transcribed verbatim, and de-identified. Descriptive statistics summarized participant demographics. Thematic analysis [14] was conducted using an inductive approach to identify key themes, and transcripts were coded independently by 2 researchers (PN and SD) before being reviewed by the research team.

Ethical Considerations

This study was reviewed and approved by the University of Missouri Health Care Institutional Review Board (IRB #2078424 MU). The protocol was deemed to be no greater than minimal risk. Written informed consent was obtained from all participants, including disclosure of the study goals. Participants could opt out at any time. Nonessential identifying information has been removed for publication. Screenshots and examples included in the manuscript were deidentified so that no individual could be identified directly or indirectly. Participants were compensated with a US $50 cash card.

Results

Patient Characteristics and Thematic Analysis Findings From Interviews

Patient Characteristics

Overall, 11 patients participated, recruited from a specialty diabetes center. The average age was 72 (SD 4.9; range 66 - 83) years, 6 were female (55%), and most were non-Hispanic White (10/11, 91%). Nearly half (5/11, 45.5%) had some college education. The mean duration of diabetes was 26 (SD 15; range 3 - 47) years, with a mean hemoglobin A1c (HbA_{1C}) of 7.6% (SD 1.2%; range 6.1% - 10.3%). Most patients were insulin users (9/11, 82%) and routinely accessed their clinic notes via patient portals (10/11, 91%), typically on their own computers (Table 1).

Table . Characteristics of patient participants (n=11).

Characteristics	Values, n (%)
Clinic location	
Cosmopolitan International Diabetes and Endocrinology Center	11 (100)
Age (years), mean (SD; range)	71.6 (4.9; 66-83)
Sex	
Male	5 (45.5)
Female	6 (54.5)
Hispanic or Latino	
No	11 (100)
Yes	0 (0)
Race	
Non-Hispanic White	10 (90.9)
Asian	1 (9.1)
Education	
Some college credit, no degree	5 (45.5)
Associate degree	2 (18.2)
High school graduate, diploma, or equivalent	1 (9.1)
Bachelor’s degree	1 (9.1)
Trade/technical/vocational training	1 (9.1)
Higher than a bachelor’s degree	1 (9.1)
Diabetes duration (years), mean (SD; range)	25.6 (15; 3-47)
HbA_{1c}, mean (SD; range)	7.6 (1.2; 6.1 - 10.3)
Insulin	
No	2 (18.2)
Yes	9 (81.8)
Access patient portal	
No	1 (9.1)
Yes	10 (90.9)
How (n=10)	
Yourself	9 (90)
With help from someone else	1 (10)
Devices (n=10)	
Computer	8 (80)
I appreciate the large screen (n=2)	—[a]
It’s easy (n=2)	—
Mobile devices	2 (20)
My phone is always with me (n=1)	—
Read clinic notes	
No	1 (9.1)
Yes	10 (90.9)

[a]Not applicable.

Readability

Most participants described the SEE-Diabetes statements as straightforward and easy to read due to plain language and clear structure. For example, a 74-year-old woman (HbA_{1c} 6.9%) highlighted that the section:

> *gives you the information about any testing that you have had and the results from it.*

While an 83-year-old man remarked it was:

> *well written and easily understood.*

However, some participants suggested adopting stronger motivational phrasing that better reflected a patient's voice to encourage action, such as statements:

> [to get them to take something seriously 71-year-old man, HbA1c 6.7%]

Helpfulness

Perceptions of helpfulness were mixed. Several participants valued the content as a practical reminder between visits*:*

> [It makes it a whole lot easier… to remember what I'm supposed to be doing 66-year-old woman, HbA1c 10.3%]

or as a motivator to improve self-care (74-year-old woman, HbA_{1c} 6.9%).

Others, especially those with long-standing diabetes, perceived limited incremental benefit, describing the information as:

> [not new 69-year-old man, HbA1c 9.2%]

> [or too broad… not specific enough to make any difference 69-year-old man, HbA1c 9.2%.]

One participant raised concerns about documentation practices, noting frustration with

> [cut and paste… especially when the information is inaccurate 74-year-old woman, HbA1c 7%]

Overall, participants viewed helpfulness as dependent on personalization, specificity, and avoidance of redundant content.

Perceived Value

Patient views on added value also varied. Some appreciated the consolidation of practical information:

> [They don't have to go online and google it. The facts are here 68-year-old woman, HbA1c 6.1%]

and emphasized that SEE-Diabetes could complement physician communication, which was sometimes perceived as incomplete:

> [Doctors aren't the best at communicating all the information. I think those notes actually cover the information… better 69-year-old man, HbA1c 9.2%]

Others reported minimal added value because they were already managing well (76-year-old woman, HbA_{1c} 7.8%) or desired clearer, directive next steps:

> [If there's a diabetes education section… another section with recommendations… I would read that too 71-year-old man, HbA1c 6.7%]

In this context, participants referred to distinct thematic groupings within the SEE-Diabetes content, with actionable recommendations embedded under each of the 7 ADCES7-aligned headings rather than presented in a separate section. Several noted that regular updates and tailoring would be essential to maintain engagement and prevent redundancy. Additional illustrative quotes are provided in Multimedia Appendix 1.

Clinician Characteristics and Thematic Analysis Findings From Interviews

Clinician Characteristics

In total, 8 clinicians participated, including 5 from specialty diabetes care clinics and family medicine (primary care) settings. The average age was 49 (SD 13.5; range 32 - 65) years, and 7 were female (88%). Most were non-Hispanic White (6/8, 75%). The average experience in diabetes care was 13 (SD 12.7; range 2 - 30) years. Most clinicians were familiar with ADCES7 (5/8, 63%) and DSMES guidelines (6/8, 75%) (Table 2).

Table . Characteristics of clinician participants and knowledge of diabetes self-management education and support and Association of Diabetes Care & Education Specialists 7 (n=8).

Characteristics	n (%)
Clinic location	
Cosmopolitan International Diabetes and Endocrinology	5 (62.5)
Keene Family Medicine	2 (25)
Ashland Family Medicine	1 (12.5)
Age (years), mean (SD; range)	48.6 (13.5; 32-65)
Sex	
Male	1 (12.5)
Female	7 (87.5)
Hispanic or Latino	
No	8 (100)
Yes	0 (0)
Race	
Non-Hispanic White	6 (75)
Asian	2 (25)
Work experience (years), mean (SD; range)	12.7 (9.5; 2-30)
Knowledge about DSMES[a] and ADCES7[b] guidelines	
Familiar with ADCES7	
No	3 (37.5)
Yes	5 (62.5)
Familiar DSMES	
No	2 (25)
Yes	6 (75)

[a]DSMES: diabetes self-management education and support.

[b]ADCES7: Association of Diabetes Care & Education Specialists 7 Self-Care Behavior.

Clarity and Concise

Most clinicians agreed that the SEE-Diabetes statements were concise, free of jargon, and written in accessible language. A 40-year-old diabetes specialist noted that the notes "use simple language, no medical jargon, and [are] easy to read." Similarly, a primary care physician with 2 years' experience described the information as "short and easy to understand." However, some clinicians highlighted areas of ambiguity. For instance, a diabetes specialist (8 y experience) observed that the phrasing around medication timing and weight control was confusing and insufficiently specific, suggesting that clearer targets, such as "work on weight loss of 5%," would enhance patient comprehension.

Sufficiency of Content

Several clinicians endorsed the adequacy of the content, describing it as "pretty thorough and self-explanatory (diabetes specialist, 8 y experience). However, others raised concerns that some sections, particularly related to medication adherence, lacked clarity and risked confusing patients. A primary care physician (2 y experience) noted difficulty interpreting the statement regarding bringing medications to the next appointment, whereas another clinician emphasized the importance of ensuring that each educational category adequately addressed patient priorities.

Clinical Usefulness

Clinicians generally recognized the clinical utility of SEE-Diabetes in supporting patient education and reinforcing self-care. Several reported that the tool aligned with common teaching practices, such as educating patients about blood glucose monitoring, interpreting results, and linking lifestyle behaviors with outcomes (diabetes specialist, 8 y experience). Others saw potential value in emphasizing diabetes-specific goals during visits that are often crowded with competing priorities (primary care physician, 30 y experience). Nonetheless, some cautioned that time constraints may limit consistent use in busy practices. Additionally, suggestions for refinement included offering more concrete examples, such as defining portion sizes in relatable terms (diabetes specialist, 22 y experience), to maximize patient engagement and comprehension. Additional illustrative quotes are provided in Multimedia Appendix 1.

Discussion

Principal Findings

This study validated the content of SEE-Diabetes, an EHR-integrated patient education tool designed to support self-management among older adults with diabetes. By incorporating both expert review and direct feedback from patients and clinicians, we assessed the clarity, relevance, and clinical utility of the educational content. Our findings indicate that SEE-Diabetes has strong potential to address documentation and communication gaps in delivering DSMES and to facilitate more personalized, actionable communication during routine outpatient care. Importantly, SEE-Diabetes is not intended to replace formal DSMES, which remains an ongoing, person-centered process grounded in the assessment of individual learning needs and preferences. Participants may have received varying levels of diabetes education through prior DSMES or routine clinician-provided counseling; however, the amount and modality of such education were not assessed. Accordingly, SEE-Diabetes was evaluated as a complementary, EHR-integrated tool to reinforce routine self-management communication rather than as a measure of DSMES exposure or delivery.

Content validation was conducted using a multimethod approach that combined expert opinion, end-user perspectives, and alignment with the ADCES7 framework [15]. This strategy ensured SEE-Diabetes is grounded in scientific evidence and the practical realities of diabetes care. While content validation is sometimes overlooked in digital health tool development, it plays a critical role in ensuring safety, relevance, and usability. For instance, Patel et al [16] created a clinical decision support system for patients with serious mental illness and diabetes but relied mainly on in silico validation due to the complexity of real-world testing. Such computational methods are useful for assessing technical performance; however, they can delay clinical implementation and may overlook usability issues in practice [17]. In contrast, our study prioritized real-world applicability by engaging both patients and clinicians in the evaluation process, thereby strengthening the credibility and adaptability of SEE-Diabetes in routine care.

Readability and understandability of the educational content emerged as a central theme in the feedback from both patients and clinicians. This aligns with prior evidence that older adults, who may experience cognitive decline or limited health literacy, benefit significantly from materials presented in straightforward, jargon-free language [18]. Communicating health information in clear, familiar terms (eg, using plain language and avoiding medical jargon) significantly improves comprehension and engagement [18]. Participant feedback in our study consistently reinforced the value of plain language in promoting understanding, highlighting the ongoing need for patient-centered communication strategies across health care settings [19]. Ensuring educational content is easily digestible is especially critical for older adults, as it can empower them to more actively participate in their care.

Clinicians viewed SEE-Diabetes as a concise, efficient tool for delivering self-care guidance in time-constrained clinic visits, consistent with prior research showing that brief, targeted educational interventions can be effective in busy health care environments [20-22]. At the same time, some clinicians suggested further refining certain statements (particularly in the "Taking Medication" domain) to enhance clarity and better motivate patients. For example, one provider commented, "Bringing meds to the visit does not ensure the patient will take them regularly between visits." Such feedback underscores the importance of iterative development and continuous user input to ensure that tools like SEE-Diabetes remain clinically relevant, context-sensitive, and adaptable [23]. Incorporating provider and patient suggestions in subsequent revisions will help address these nuances and improve the tool's effectiveness.

Our analysis also identified a remaining gap in the delivery of patient-centered education during routine diabetes follow-up visits. This finding echoes prior studies indicating that although DSMES is widely implemented, it often lacks the personalization necessary to meet individual patient needs [6,7,11,24]. In our previous work, we observed that standard follow-up clinic notes frequently lacked patient-centered education for patients with diabetes [7]. SEE-Diabetes directly addresses this gap by embedding personalized educational content directly into the clinic note (which nearly 80% of our older patients reported reading via the patient portal [7]). By aligning educational messages with each patient's unique context and self-management goals, this approach supports the broader movement toward patient-centered care. Such individualized interventions are expected to enhance patient engagement and treatment adherence and ultimately improve outcomes in diabetes management.

Strengths and Limitations

A key strength of this study lies in its user-centered validation approach, which engaged both patients and clinicians across primary care and specialty care settings. By involving real-world end users in the design and evaluation process, we ensured that SEE-Diabetes content is not only evidence-based but also practical, readable, and clinically relevant. The use of tailored clinical scenarios, combined with in-depth qualitative interviews, provided rich insights into the clarity, usefulness, and perceived value. This multistakeholder engagement enhances the credibility of our findings and supports the tool's adaptability across diverse workflows, thereby strengthening its potential for real-world implementation. Notably, our approach aligns with UCD principles that emphasize iterative development and continuous involvement of target users [9]. By continuously incorporating feedback from both providers and patients, we aimed to develop an educational tool that meets users' needs in everyday practice.

Limitations of this study include a small sample size and a lack of racial and geographic diversity in our participants. Because the majority of participants were non-Hispanic White and recruitment was limited to a single academic health center, the generalizability of our findings may be constrained. This homogeneity is consistent with the demographic profile of the Midwestern United States, where approximately 73% of the population identifies as non-Hispanic White, which likely influenced the composition of our sample [25]. Future work should evaluate SEE-Diabetes in larger and more diverse populations and test its implementation across various clinical settings and regions. Despite these limitations, our study supports the feasibility and potential value of integrating personalized education into routine care through tools like SEE-Diabetes. The structured, user-informed content provided by SEE-Diabetes may help improve patient-provider communication, support patient self-management, and ultimately contribute to more patient-centered chronic disease care.

Future Directions

Beyond the current implementation, SEE-Diabetes has potential for broader scalability across diverse care settings. While this study focused on EHR-based delivery, future work could explore parallel formats such as printable summaries or patient-facing handouts to support clinics without advanced EHR functionality, including rural and resource-limited programs. Additionally, situating SEE-Diabetes within national DSMES Standards and the 4 critical times for DSMES delivery may help align its use with formal education pathways while reinforcing self-management communication during routine care.

Conclusions

This study validated SEE-Diabetes, a patient-centered tool that embeds tailored diabetes self-management support into EHR notes for older adults. Both patients and clinicians confirmed that the content is clear, relevant, and feasible for integration into primary and specialty care. Embedding plain-language education within routine documentation may strengthen communication, reinforce self-care, and support chronic disease management in aging populations. Future work should evaluate implementation across diverse settings and its impact on clinical outcomes, engagement, and scalability.

Acknowledgments

The authors are grateful to the patients and clinicians at the University of Missouri Health Care who generously contributed their time and insights to this study. The authors also thank the interdisciplinary team members and research assistants who supported data collection and analysis.

Funding

This work was supported by the Agency for Healthcare Research and Quality [R21HS028032]; and the National Institute of Diabetes and Digestive and Kidney Disorders [P30DK092950].

Data Availability

The datasets generated or analyzed during this study are not publicly available due to the confidential nature of the qualitative interview data but are available from the corresponding author on reasonable request.

Authors' Contributions

Conceptualization: MSK (lead), UK (equal), MD (supporting), and SAB (supporting)
Data curation: PN
Formal analysis: PN (lead) and MSK (supporting)
Funding acquisition: MSK
Investigation: PN (lead), SD (equal), UK (equal), MD (equal), MSK (supporting), and SAB (supporting)
Methodology: MSK (lead), PN (supporting), SD (supporting), UK (supporting), MD (supporting), and SAB (supporting)
Project administration: MSK (lead)
Resources: MSK
Supervision: MSK
Validation: MSK
Visualization: PN (lead), SD (supporting), and UK (supporting)
Writing – original draft: PN (lead)
Writing – review and editing: MSK (lead), SD (supporting), UK (supporting), MD (supporting), SAB (supporting), and EJS (supporting)

Conflicts of Interest

None declared.

Multimedia Appendix 1
Representative participant quotations illustrating perceptions of clarity, relevance, usability, and clinical usefulness of SEE-Diabetes content, organized by stakeholder group (patients and clinicians) and ADCES7-aligned domains.
[DOCX File, 18 KB - diabetes_v11i1e83448_app1.docx]

References

1. National diabetes statistics report. Centers for Disease Control and Prevention. 2024. URL: https://www.cdc.gov/diabetes/php/data-research/index.html [accessed 2026-01-05]
2. Davis J, Fischl AH, Beck J, et al. 2022 National standards for diabetes self-management education and support. Diabetes Care 2022 Feb 1;45(2):484-494. [doi: 10.2337/dc21-2396] [Medline: 35050365]
3. Chrvala CA, Sherr D, Lipman RD. Diabetes self-management education for adults with type 2 diabetes mellitus: a systematic review of the effect on glycemic control. Patient Educ Couns 2016 Jun;99(6):926-943. [doi: 10.1016/j.pec.2015.11.003] [Medline: 26658704]
4. He X, Li J, Wang B, et al. Diabetes self-management education reduces risk of all-cause mortality in type 2 diabetes patients: a systematic review and meta-analysis. Endocrine 2017 Mar;55(3):712-731. [doi: 10.1007/s12020-016-1168-2] [Medline: 27837440]
5. Alliston P, Jovkovic M, Khalid S, Fitzpatrick-Lewis D, Ali MU, Sherifali D. The effects of diabetes self-management programs on clinical and patient reported outcomes in older adults: a systematic review and meta-analysis. Front Clin Diabetes Healthc 2024;5:1348104. [doi: 10.3389/fcdhc.2024.1348104] [Medline: 38952998]
6. Narindrarangkura P, Boren SA, Khan U, Day M, Simoes EJ, Kim MS. SEE-diabetes, a patient-centered diabetes self-management education and support for older adults: findings and information needs from providers' perspectives. Diabetes Metab Syndr 2022 Sep;16(9):102582. [doi: 10.1016/j.dsx.2022.102582] [Medline: 35963033]
7. Narindrarangkura P, Boren SA, Khan U, Day M, Simoes EJ, Kim MS. SEE-Diabetes, a patient-centered diabetes self-management education and support for older adults: findings and information needs from patients' perspectives. Prim Care Diabetes 2022 Jun;16(3):395-403. [doi: 10.1016/j.pcd.2022.02.008] [Medline: 35227635]
8. ElSayed NA, McCoy RG, Aleppo G. 5. Facilitating positive health behaviors and well-being to improve health outcomes: standards of care in diabetes-2025. Diabetes Care 2025 Jan 1;48(1 Suppl 1):S86-S127. [doi: 10.2337/dc25-S005] [Medline: 39651983]
9. Graham AK, Munson SA, Reddy M, et al. Integrating user-centered design and behavioral science to design a mobile intervention for obesity and binge eating: mixed methods analysis. JMIR Form Res 2021 May 10;5(5):e23809. [doi: 10.2196/23809] [Medline: 33970114]
10. Freeman S, Lecka L, Davis K, et al. Leveraging user-centered design and usability and skills testing for a novel diabetes survival skills toolkit. J Endocr Soc 2025 Sep;9(9):bvaf111. [doi: 10.1210/jendso/bvaf111] [Medline: 40799775]
11. Ye Q, Patel R, Khan U, Boren SA, Kim MS. Evaluation of provider documentation patterns as a tool to deliver ongoing patient-centred diabetes education and support. Int J Clin Pract 2020 Mar;74(3):e13451. [doi: 10.1111/ijcp.13451] [Medline: 31769903]
12. Ghorayeb A, Darbyshire JL, Wronikowska MW, Watkinson PJ. Design and validation of a new Healthcare Systems Usability Scale (HSUS) for clinical decision support systems: a mixed-methods approach. BMJ Open 2023 Jan 30;13(1):e065323. [doi: 10.1136/bmjopen-2022-065323] [Medline: 36717136]
13. Facts/figures. MU Health Care. URL: https://www.muhealth.org/for-media/resources/factsfigures [accessed 2026-01-05]
14. Guest G, MacQueen KM, Namey EE. Applied Thematic Analysis: SAGE Publications; 2011. [doi: 10.4135/9781483384436]
15. American Association of Diabetes Educators. An effective model of diabetes care and education: revising the AADE7 self-care behaviors®. Diabetes Educ 2020 Apr;46(2):139-160. [doi: 10.1177/0145721719894903] [Medline: 31928334]
16. Patel D, Msosa YJ, Wang T, et al. An implementation framework and a feasibility evaluation of a clinical decision support system for diabetes management in secondary mental healthcare using CogStack. BMC Med Inform Decis Mak 2022 Apr 14;22(1):100. [doi: 10.1186/s12911-022-01842-5] [Medline: 35421974]
17. Wasylewicz A, Scheepers-Hoeks A. Clinical decision support systems. In: Dumontier M, Kubben P, Dekker A, editors. Fundamentals of Clinical Data Science: Springer, Cham; 2018:153-169. [doi: 10.1007/978-3-319-99713-1_11] [Medline: 31314237]
18. Shi Z, Du X, Li J, Hou R, Sun J, Marohabutr T. Factors influencing digital health literacy among older adults: a scoping review. Front Public Health 2024;12:1447747. [doi: 10.3389/fpubh.2024.1447747] [Medline: 39555039]
19. Johnson A. Health literacy, does it make a difference? Aust J Adv Nurs 2014;31(3):39-45. [doi: 10.37464/2016.334.1596]
20. Joyce MF, Berg S, Bittner EA. Practical strategies for increasing efficiency and effectiveness in critical care education. World J Crit Care Med 2017 Feb 4;6(1):1-12. [doi: 10.5492/wjccm.v6.i1.1] [Medline: 28224102]
21. Klein JW, Jackson SL, Bell SK, et al. Your patient is now reading your note: opportunities, problems, and prospects. Am J Med 2016 Oct;129(10):1018-1021. [doi: 10.1016/j.amjmed.2016.05.015] [Medline: 27288854]
22. Zhang A, Wang J, Wan X, et al. Community-based intelligent blood glucose management for older adults with type 2 diabetes based on the health belief model: randomized controlled trial. JMIR Mhealth Uhealth 2025 Jun 19;13:e60227. [doi: 10.2196/60227] [Medline: 40536934]
23. Stamate A, Marzan MD, Velciu M, Paul C, Spiru L. Advancing user-centric design and technology adoption for aging populations: a multifaceted approach. Front Public Health 2024;12:1469815. [doi: 10.3389/fpubh.2024.1469815] [Medline: 39712308]
24. Kim MS, Khan U, Boren SA, Narindrarangkura P, Ye Q, Simoes EJ. Transforming AADE7 for use in an evaluation framework for health information technology in diabetes mellitus. J Diabetes Sci Technol 2022 May;16(3):764-770. [doi: 10.1177/1932296820985842] [Medline: 33435720]

25. Census profile: midwest region. Census Reporter. 2021. URL: http://censusreporter.org/profiles/02000US2-midwest-region/ [accessed 2026-01-05]

Abbreviations

ADCES7: Association of Diabetes Care & Education Specialists 7 Self-Care Behaviors
DSMES: diabetes self-management education and support
EHR: electronic health record
SEE-Diabetes: Support-Engage-Empower-Diabetes
UCD: user-centered design

Edited by S Li; submitted 03.Sep.2025; peer-reviewed by AH Fischl, AD Keetharuth; revised version received 06.Jan.2026; accepted 07.Jan.2026; published 06.Feb.2026.

Please cite as:
Narindrarangkura P, Dejhansathit S, Khan U, Day M, Boren SA, Simoes EJ, Kim MS
Content Validation of an Electronic Health Record–Based Diabetes Self-Management Support Tool for Older Adults With Type 2 Diabetes: Qualitative Study
JMIR Diabetes 2026;11:e83448
URL: https://diabetes.jmir.org/2026/1/e83448
doi:10.2196/83448

Inter-Regional Center for Automated Insulin in Diabetes (CIRDIA) and Hospital-Based Approaches to Closed-Loop Therapy in Type 1 Diabetes: Cost-Effectiveness Analysis

Mercia Napame[1], MHE; Sylvie Picard[2], MD, PhD; Tony Foglia[3], PhD; Anne Rubenstrunk[1], PhD; Florence Baudoux[1], MD; Francoise Giroud[4], MD; Sandrine Lablanche[5], MD, PhD; Sophie Borot[6], MD, PhD

[1]Santelys, Loos, France
[2]CIRDIA, Point Medical, Rond-Point de la Nation, Dijon, France
[3]URPS-ML-BFC, Burgundy-Europe University, Dijon, France
[4]CDOM21, CIRDIA, Dijon, France
[5]Centre Hospitalier Universitaire de Grenoble, Grenoble, France
[6]Centre Hospitalier Universitaire de Besançon, Besançon, France

Corresponding Author:
Sylvie Picard, MD, PhD
CIRDIA, Point Medical, Rond-Point de la Nation, Dijon, France

Abstract

Background: Closed-loop insulin delivery is the new standard of care for patients with type 1 diabetes (T1D). However, in France, its implementation remains predominantly hospital based. Expanding access to this treatment through alternative care models looks essential.

Objective: This study (cost-effectiveness analysis) compares 2 care models for people with T1D implementing a closed-loop system in France: outpatient care in the Inter-Regional Center for Automated Insulin in Diabetes (CIRDIA) and inpatient care.

Methods: We conducted a cost-effectiveness analysis using retrospective observational data from individuals with T1D aged 16 years and older from the implementation of the closed loop to a 12-month follow-up either in the CIRDIA (CIRDIA group) or in a hospital center setting (hospital center [HC] group). The cost analyses were based on patient records and public databases: the French Medical Information Systems Program and the French General Nomenclature of Professional Acts. Closed-loop efficacy was assessed using a time in range (TIR) of 70 to 180 mg/dL, and closed-loop safety was assessed using the glycemia risk index (GRI), a single indicator that represents the risk of hypoglycemia or hyperglycemia and ranges from 0 (minimal risk) to 100 (maximal risk).

Results: A total of 201 patients were included: 128 in the CIRDIA group and 73 in the HC group. The mean (SD) age was 43 (14) years and 46 (15) years, respectively. Mean (SD) baseline TIR was 52.9% (16%) in the CIRDIA group versus 65.9% (15.1%) in the HC group (*P*<.001), whereas mean (SD) baseline GRI was 56.4 (21) in the CIRDIA group versus 37.8 (19.8) in the HC group (*P*<.001). After 12 months, both groups achieved similar efficacy and safety outcomes with a mean (SD) TIR at 72.7% (11.6%) in the CIRDIA group versus 71.9% (10.5%) in the HC group, and a mean GRI at 30.1 (14.1) versus 30.3 (13), respectively. There were no significant between-group differences (*P*=.60 for TIR; *P*=.91 for GRI). However, the CIRDIA was associated with significantly lower management costs with a mean cost of €8373.12 (SD €427.30; €1=US $1.10 at the time of the study) per patient in the CIRDIA group versus €8814.32 (SD €192) per patient in the HC group (*P*<.001). The estimated saving was €626 per percentage point of increase in TIR and €2011 per point of reduction in GRI, indicating that the HC closed-loop initiation was dominated by the CIRDIA. The CIRDIA was less costly than HC in 8600 (86%) out of 10,000 simulations in a probabilistic sensitivity analysis.

Conclusions: These findings suggest the potential of the CIRDIA to represent a viable alternative organizational model for closed-loop initiation in France, achieving comparable effectiveness at lower cost in our population. Further research with longer follow-up is warranted. From a policy perspective, the resources saved could be at least partly reallocated to support out-of-hospital closed-loop initiation centers.

(JMIR Diabetes 2026;11:e86690) doi:10.2196/86690

KEYWORDS

closed-loop; cost-effectiveness analysis; follow-up studies; hospital; treatment outcome; type 1 diabetes

Introduction

Background

Diabetes is a chronic disease characterized by persistent hyperglycemia, resulting from either a relative or absolute deficiency in insulin secretion or an impairment in its action. It represents a major public health challenge because of its increasing prevalence, its impact on patients' quality of life, and the substantial economic burden on health care systems [1]. As of 2024, 588.8 million adults (aged 20 - 79) worldwide were living with diabetes, a number projected to increase to approximately 853 million by 2050 [2]. In France, more than 4.5 million people are living with diabetes [3]. Among the different forms of diabetes, type 1 diabetes (T1D) is an autoimmune disease that is often diagnosed in children, adolescents, or young adults. Overall, about 7.4 million people are living worldwide with T1D, and in France, T1D accounts for approximately 320,000 individuals [3]. The management of T1D requires lifelong insulin therapy, frequent or nowadays continuous glucose monitoring (CGM), and structured patient therapeutic education [4]. Over the past decades, technological advances have progressively transformed diabetes care from multiple daily injections to external insulin pumps and subsequently to CGM, enabling the real-time tracking of glycemia [5]. These innovations have paved the way for the development of closed-loop (CL) systems, which integrate a glucose sensor, an insulin pump, and an adaptive control algorithm [5].

Prior Work

While numerous studies have established the clinical benefits of CL systems on glycemic outcomes, evidence on the models of care for their initiation and follow-up remains limited [6-10]. The recent reimbursement of CL systems in France, and the relative novelty of studying organizational rather than purely clinical outcomes, may explain this evidence gap [11].

In France, approximately 2 years after the first reimbursement, only about 15,000 eligible patients had received CL systems—roughly a 5% coverage—despite the benefits for glycemic control [12]. This low rate is partly attributable to the centralization of CL initiation in hospital-based clinics, where waiting times are often long [13].

The Inter-Regional Center for Automated Insulin in Diabetes (CIRDIA) was developed in 2023 mainly to improve access to CL among persons with T1D. The CIRDIA is a multisite CL initiation center regrouping highly trained diabetologists, mostly in private practice. The CIRDIA—like hospital-based CL initiation centers—is based on the guidelines of the French-Speaking Diabetes Society (SFD) [4]. However, as this is a new concept of care in France, its cost-effectiveness had to be evaluated and compared to usual hospital-based care.

Study Objectives

Evidence on the cost-effectiveness of alternative organizational models of CL initiation, such as out-of-hospital–based pathways, remains scarce. This raises the question of whether initiating CL systems in out-of-hospital settings, such as the CIRDIA, could represent a cost-effective alternative to hospital center (HC)–based initiation.

This study aimed to estimate the 1-year cost-effectiveness of CIRDIA-based CL initiation compared to HC-based initiation among patients with T1D in France from a French National Health Insurance perspective. We hypothesized that out-of-hospital–based initiation could achieve comparable effectiveness and safety while reducing costs. Evaluating this organizational model could determine whether or not the CIRDIA represents a viable alternative for the French health care system and provide the data that may be transferable to other health care systems worldwide.

Methods

Study Design

This is a cost-effectiveness analysis based on retrospective observational data collected between 2023 and 2024 with a 12-month follow-up as part of the routine monitoring of patients with T1D initiating CL in France. We compared 2 modes of health care delivery: the CIRDIA setting and the HC setting. The cost-effectiveness analysis compared the net monetary costs of health care intervention with a measure of its clinical effectiveness.

Accordingly, the evaluation was conducted from the perspective of the French National Health Insurance (Assurance Maladie), considering all costs covered by the payer, with a 1-year time horizon. No modeling was conducted, as all analyses relied on real-world data extracted from patient records (follow-up consultations) and public databases: the Agency for Information on Hospital Care and the French Health Insurance [14].

Recruitment

The study included persons living with T1D, 16 years of age or older, starting for the first time a CL system. Patients with missing continuous glucose monitoring data were excluded. Participants were allocated to 1 of the 2 groups based on their care pathway: those managed directly by the CIRDIA center (CIRDIA group) and those initiated and followed by the hospital center outpatient clinic (HC group). The 2 models of care were mutually exclusive and could not be used simultaneously.

Participants from the CIRDIA group were consecutive patients who started CL between May 2, 2023, and March 30, 2024, and had at least a 12-month follow-up. Devices (insulin pump, infusion sets, insulin reservoirs, and glucose sensors) were provided by different home health care providers, as it is the rule in France. Registered nurses specialized in diabetes care and working for home health care providers are usually responsible for the technical education of the patient and connectivity issues. Participants in the HC group had CL initiated in 2023 or 2024 in 1 of the 5 HCs located in the north of France ("Haut-de-France" region) and were the patients for whom devices and technical education were provided by Santelys, a nonprofit organization acting as a home health care provider.

Ethical Considerations

This study used retrospective observational data collected as part of the routine monitoring of persons with T1D managed on CL therapy. No additional intervention occurred beyond usual care. All data were fully anonymized before analysis in accordance with the General Data Protection Regulation. No patient could be identified directly or indirectly [15]. In line with current regulations regarding research not involving human persons, no specific ethics committee approval was required [16].

All participants had received oral and written information at the time of CL initiation about the potential use of their anonymized clinical data for research purposes. Written consent or non-opposition was obtained in accordance with French data protection and ethical regulations. This study complied with the principles of the Declaration of Helsinki and relevant national guidelines regulating the secondary use of health data.

Interventions

The CIRDIA is a new concept in France of a multisite health care model that performs CL initiation most often during a long (about 1 h) office visit or occasionally during a day hospitalization (DH) outside of university hospitals. Its activity complies with the position statement issued by the SFD and the French National Health Authority (HAS) [17]. The main objective of the CIRDIA is to expand access to care for people living with T1D while reducing the burden on HC. Furthermore, initiating CL systems in the out-of-hospital sector is considered a strategic lever to support the sustainability of out-of-hospital diabetes care. Nevertheless, since CL initiation is predominantly performed in hospital settings, hospital-based care is considered the reference strategy. The out-of-hospital sector initiation remains underdeveloped and must demonstrate its effectiveness.

In the CIRDIA arm, CL initiation was usually followed by 3 teleconsultations and 3 consultations over 1 year. For some patients (those initiated after January 1, 2024), an additional 3-month telemonitoring period could be implemented. In the HC arm, CL initiation was carried out during DH, followed by 3 teleconsultations and 3 follow-up visits, coupled with 3 months of telemonitoring for patients initiated after January 1, 2024. In both settings, CGM data were available for the diabetologist (or the diabetes care team) to optimize patient adherence to the device [18].

Efficacy and Safety Inputs

Because CL initiation and the 1-year time horizon did not affect mortality or lifespan, we selected an alternative measure for effectiveness. However, due to incomplete data on comorbidities and complications in 1 of the 2 study arms (HC), adverse events could not be included in the analysis. Instead, effectiveness was assessed by improvement in the time in range (TIR) 70 - 180 mg/dL, while safety was assessed through a reduction in the glycemia risk index (GRI). The GRI is a composite metric that reflects both hypoglycemia and hyperglycemia risks by integrating the time spent below range (<54 mg/dL and 54 - 69 mg/dL) and the time spent above range (181 - 250 mg/dL and >250 mg/dL). Notably, although hemoglobin A_{1c} is frequently used as an efficacy outcome in similar studies, it is no longer systematically measured during routine consultations [19].

Cost Inputs

We conducted the economic evaluation from a health care payer perspective, including all direct medical and nonmedical expenses reimbursed by the French National Health Insurance, expressed in euros for the year 2024. Costs were estimated using a bottom-up micro-costing approach, which is considered the gold standard in health technology cost assessment according to HAS recommendations. Because T1D belongs to the list of fully covered diseases by the French National Health Insurance, no out-of-pocket expense was considered. Moreover, because the time horizon was limited to 1 year, no discount rate was applied. Cost components were identified and calculated in line with the HAS and SFD recommendations [20,21].

Outpatient procedures and consultations were valued according to the prices from the General Classification of Professional Acts and the Common Classification of Medical Acts. Biological analyses were valued according to the Common Nomenclature of Medical Biology Acts. In addition, CL-related costs were valued in accordance with the List of Products and Services of the French National Health Insurance. The cost of DH was calculated using the Homogeneous Group of Patients with the principal diagnosis code Z451 (“Adjustment and maintenance of an infusion pump”), associated with the Hospital Stay Tariff 1794, based on prices provided by the Agency for Information on Hospital Care [22-26].

Incremental Cost-Effectiveness Ratio

The results of a cost-effectiveness analysis were expressed in terms of incremental cost-effectiveness ratios (ICERs) and were calculated as the ratio of incremental costs to incremental health outcomes between the 2 groups. Specifically, ICERs were expressed as the additional cost per percentage point of increase in TIR and per unit of reduction in the GRI. In line with International Society for Pharmacoeconomics and Outcomes Research recommendations, negative ICERs were interpreted as situations of dominance or dominated strategies rather than reported as such. A strategy was considered to be dominated if it was more costly and less effective or more costly and equally effective. We designed, conducted, and reported this evaluation in accordance with the CHEERS (Consolidated Health Economic Evaluation Reporting Standards) guidelines [27].

Sensitivity Analysis

As this study was based on real-world observational data rather than modeled parameters, some uncertainty may still arise from the data, potentially leading to biased estimates. According to the International Society for Pharmacoeconomics and Outcomes Research [28], deterministic sensitivity analysis was not applicable in this context. Instead, robustness was explored through subgroup analyses and through a probabilistic sensitivity analysis (PSA) to test whether the conclusions of the base-case analysis held under parameter uncertainty. A PSA was performed using 10,000 Monte Carlo simulations in which all parameters were varied simultaneously. Parameter values were sampled from predefined probability distributions: truncated normal for efficacy and safety outcomes (bounded between 0 and 100) and gamma for costs [28].

Statistical Analysis

Data were collected using Excel (2016, Microsoft Inc.), and statistical analyses were performed with the R software version 4.4.2 (2024). Means and SDs were calculated for quantitative variables. To verify comparability between the groups, we conducted a Shapiro-Wilk test to check the normality of our variables. For normally distributed variables, we used a 2-tailed Student *t* test, and for non-normally distributed variables, the Wilcoxon signed-rank test was used. The threshold for statistical significance was set at *P*<.05.

Results

Overview

Overall, 201 patients aged 16 to 80 years were included in this study, including 128 CL initiations by the CIRDIA and 73 by HC. Baseline characteristics of the 2 groups are shown in Table 1. The mean age of patients initiated at CL by the CIRDIA was 43 (SD 14) years and 46 (SD 15) years for the HC arm. The gender distribution was 52% (n=66 and n=38) women and 48% (n= 62 and n=35) men in both arms, and the average BMI was 27.5 (SD 4.9) and 27.2 (SD 5.2) kg/m^2, respectively. In the CIRDIA arm, only 17 (13%) CL initiations were performed during DH, while the remaining initiations were conducted during 1-hour office visits.

Table . Baseline characteristics of the patients included in the study by group (CIRDIA[a] vs HC[b]).

Parameters	CIRDIA (n=128)	HC (n=73)	*P* value (*t* test/Wilcoxon test)
BMI (kg/m^2), mean (SD)	27.5 (4.9)	27.2 (5.2)	.71
Weight (kg), mean (SD)	78.5 (14.8)	79.8 (15.3)	.55
Height (cm), mean (SD)	169 (7.9)	171 (9.3)	.048
Gender, n (%)			
Men	62 (48)	35 (48)	.95
Women	66 (52)	38 (52)	—[c]
Age class (y), n (%)			
<25	13 (10.2)	7 (11)	—
25-45	60 (46.9)	27 (37)	—
45-65	46 (35.9)	31 (42.5)	—
>65	9 (7)	7 (9.6)	—
Age (y), mean (SD)			
At pump initiation	34 (15)	45 (15)	<.001
At closed-loop initiation	43 (14)	46 (15)	.15
Pump model, n (%)			
Medtronic 780G (with Guardian 4 sensor)	99 (77)	67 (92)	—
"Control IQ" (Tandem Slim 2X pump, Dexcom G6 sensor)	17 (9)	6 (8)	—
"CamAPS" (Ypsopump, Dexcom G6 sensor)	12 (13)	—	—
Baseline glucose control, mean (SD)			
TIR[d] (%)	52.9 (16)	65.9 (15.1)	<.001[e]
GRI[f]	56.4 (21)	37.8 (19.8)	<.001[e]

[a]CIRDIA: Inter-Regional Center for Automated Insulin in Diabetes.

[b]HC: hospital center.

[c]Not applicable.

[d]TIR: time in range 70-180 mg/dL.

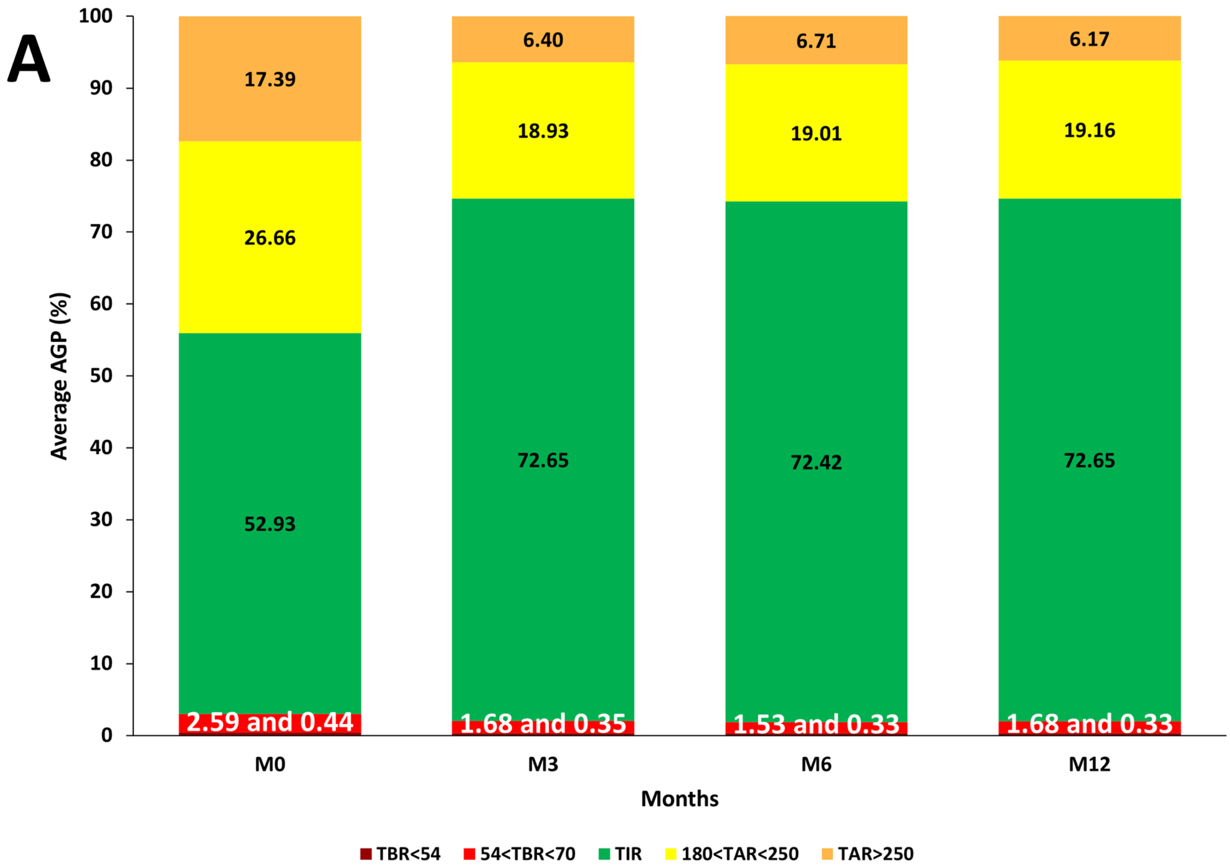
[e]Wilcoxon test values.

[f]GRI: glycemia risk index.

Efficacy and Safety Outcomes

Figure 1 illustrates the changes in the ambulatory glucose profile from baseline to 1 year after initiation.

Figure 1. Ambulatory glucose profile for both comparison arms (A: CIRDIA group, B: hospital centers group) at baseline (M0) and after 3 months (M3), 6 months (M6), and 12 months (M12) of closed-loop use. AGP: ambulatory glucose profile; CIRDIA: Inter-Regional Center for Automated Insulin in Diabetes; TAR: time above range; TBR: time below range; TIR: time in range.


A
Average AGP (%)
100
90
80
70
60
50
40
30
20
10
0
17.39
26.66
52.93
2.59 and 0.44
6.40
18.93
72.65
1.68 and 0.35
6.71
19.01
72.42
1.53 and 0.33
6.17
19.16
72.65
1.68 and 0.33
M0
M3
M6
M12
Months
TBR<54
54<TBR<70
TIR
180<TAR<250
TAR>250

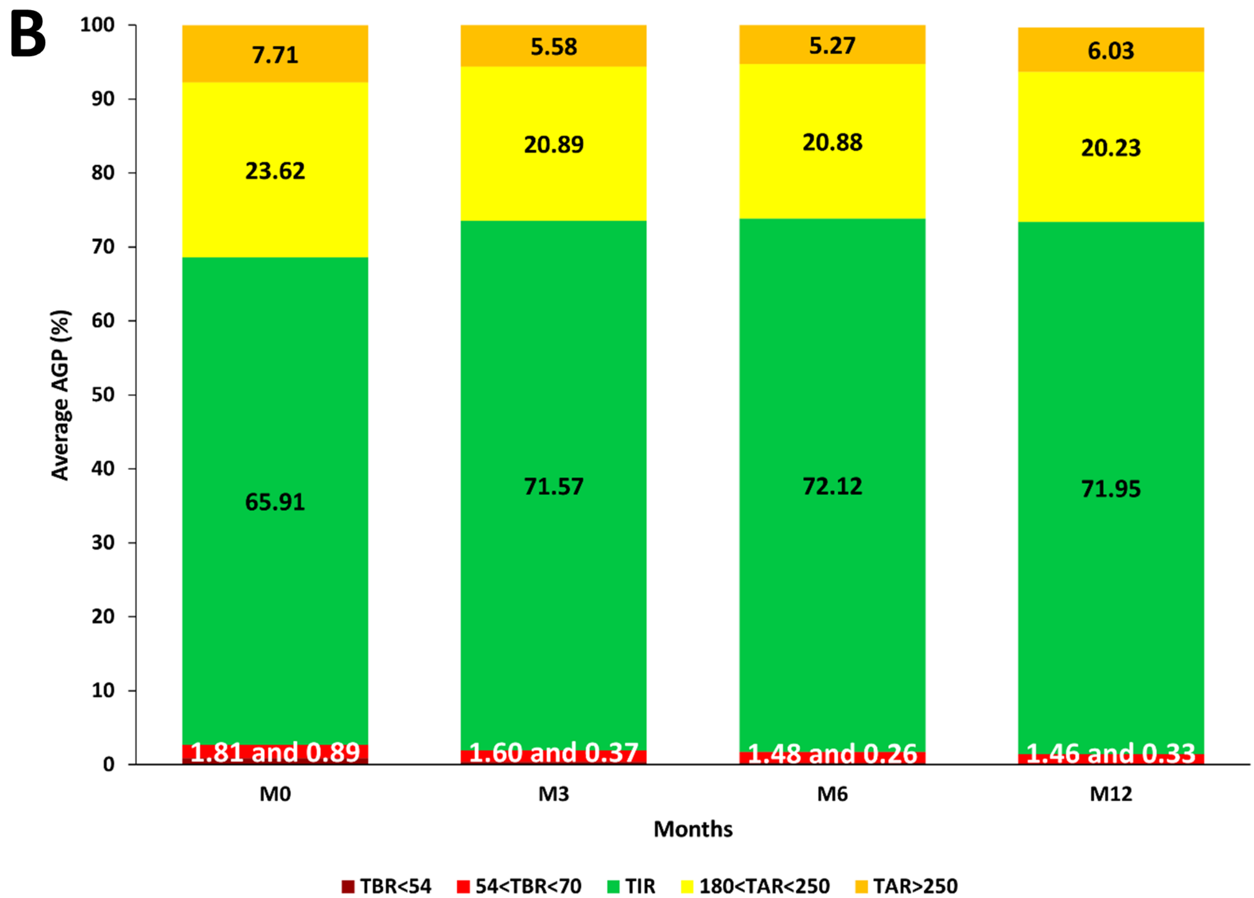

B
Average AGP (%)
100
90
80
70
60
50
40
30
20
10
0
7.71
23.62
65.91
1.81 and 0.89
5.58
20.89
71.57
1.60 and 0.37
5.27
20.88
72.12
1.48 and 0.26
6.03
20.23
71.95
1.46 and 0.33
M0
M3
M6
M12
Months
TBR<54
54<TBR<70
TIR
180<TAR<250
TAR>250

Figure 2 presents GRI grids showing glycemic risk zones over the same period (zone A: minimal hypo- or hyperglycemia risk; zone E: maximal hypo- or hyperglycemia risk). At baseline, 79% (101/128) of the patients from the CIRDIA group were in the intermediate risk (zone C) or high-risk zones (zones D and E). After 1 year on CL, only 21% (27/128) remained in these GRI zones. In the HC arm, 34% (25/73) of the patients were in zones C, D, and E at baseline, and 25% (18/73) remained in these zones after 1 year.

Figure 2. Glycemia risk index (GRI) grids at baseline (M0) and 1 year after closed-loop initiation (M12). Upper grids: Inter-Regional Center for Automated Insulin in Diabetes (CIRDIA) group; lower grids: hospital center (HC) group. Each participant is identified by a blue circle and their identification number.

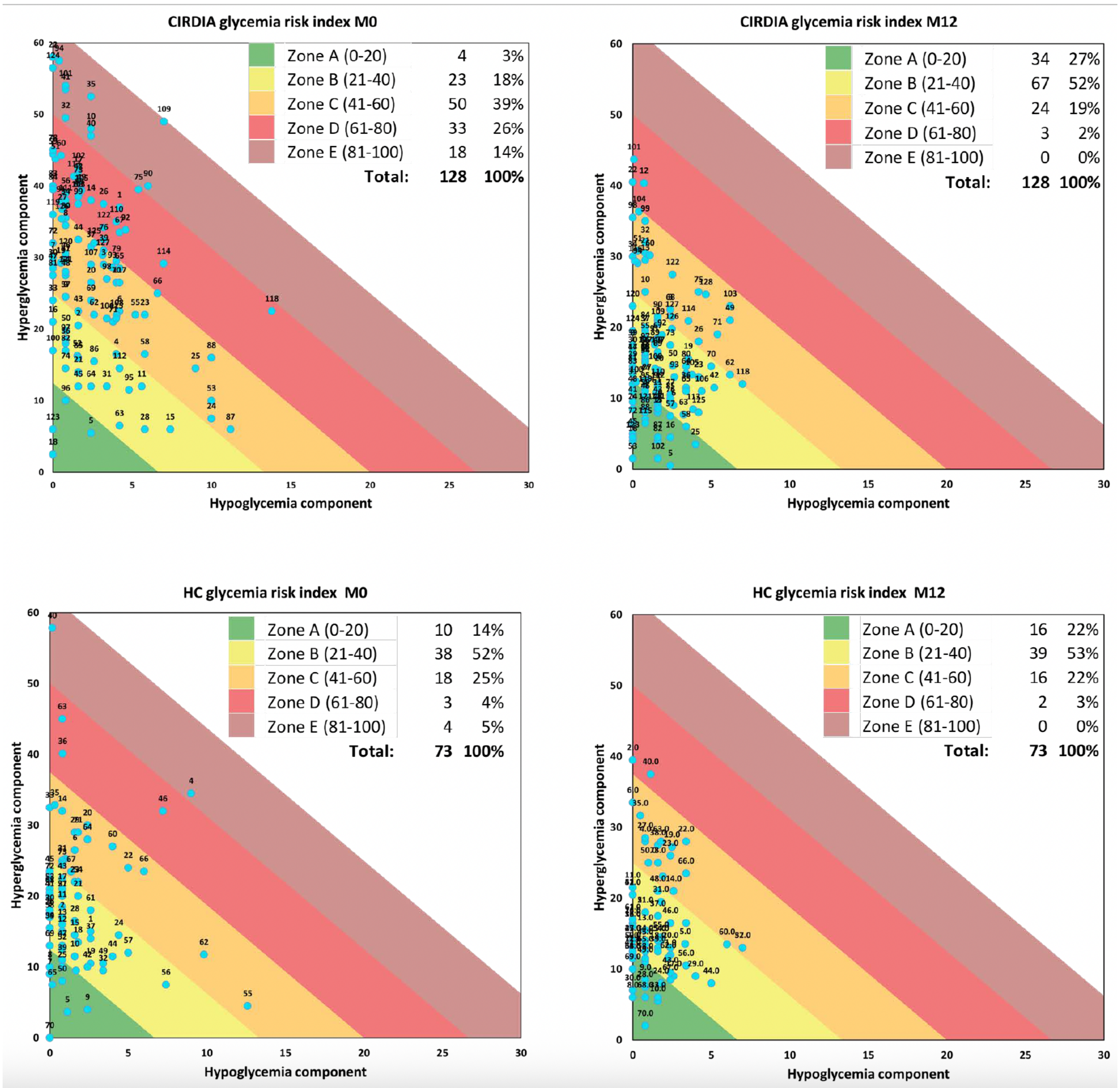


Table 2 summarizes the overall effectiveness and safety results for the total population and according to age at inclusion. After 12 months, the mean (SD) TIR increased by 19.8 points in the CIRDIA group (from 52.9% [16] to 72.7% [11.6]) and by 6 points in the HC group (from 65.9% [15.1] to 71.9% [10.5]). Although baseline differences were significant (*P*<.001), no significant difference between groups was observed at 12 months (*P*=.60). The GRI decreased in both groups, by 26.3 points in the CIRDIA group, from 56.4 (21) to 30.1 (14.1), and by 7.5 points in the HC arm, from 37.8 (19.8) to 30.3 (13). No significant difference between groups was observed at 12 months (*P*=.91).

Table . Glycemic outcomes at baseline (M0) and 12 months (M12) after closed-loop initiation for the total population and according to age class at inclusion[a].

Parameters	CIRDIA[b] (n=128), mean (SD)	HC[c] (n=73), mean (SD)	*P* value
M0[d]			
TIR[e]			
Total	52.9 (16)	65.9 (15.1)	<.001
Age class (y)			
<25	50.4 (11.4)	67.8 (14.7)	.01
25-45	48.4 (15.6)	64.7 (16.7)	<.001
45-65	57.6 (16.3)	65.8 (14)	.02
≥65	62.9 (14.2)	68.9 (14.2)	.30
GRI[f]			
Total	56.4 (21)	37.8 (19.8)	<.001
Age class (y)			
<25	59.6 (16.9)	35.3 (17.9)	.007
25-45	63 (20.7)	40.2 (22.1)	<.001
45-65	50.2 (20.1)	36.6 (17.7)	.002
≥65	39.3 (16.6)	36.7 (24.3)	.70
M12[g]			
TIR			
Total	72.7 (11.6)	71.9 (10.5)	.60
Age class (y)			
<25	70 (11.8)	77.9 (11.8)	.20
25-45	69.1 (11.7)	72 (7.9)	.20
45-65	76.7 (10.6)	72.2 (10.8)	.09
≥65	79 (6.2)	63.6 (14)	.03
GRI			
Total	30.1 (14.1)	30.3 (13)	.91
Age class (y)			
<25	33.6 (14.5)	23.5 (15.1)	.10
25-45	34.4 (13.9)	29.7 (10.7)	.12
45-65	25.3 (13.1)	30.2 (13)	.13
≥65	21.2 (7.4)	40.9 (15.3)	.01

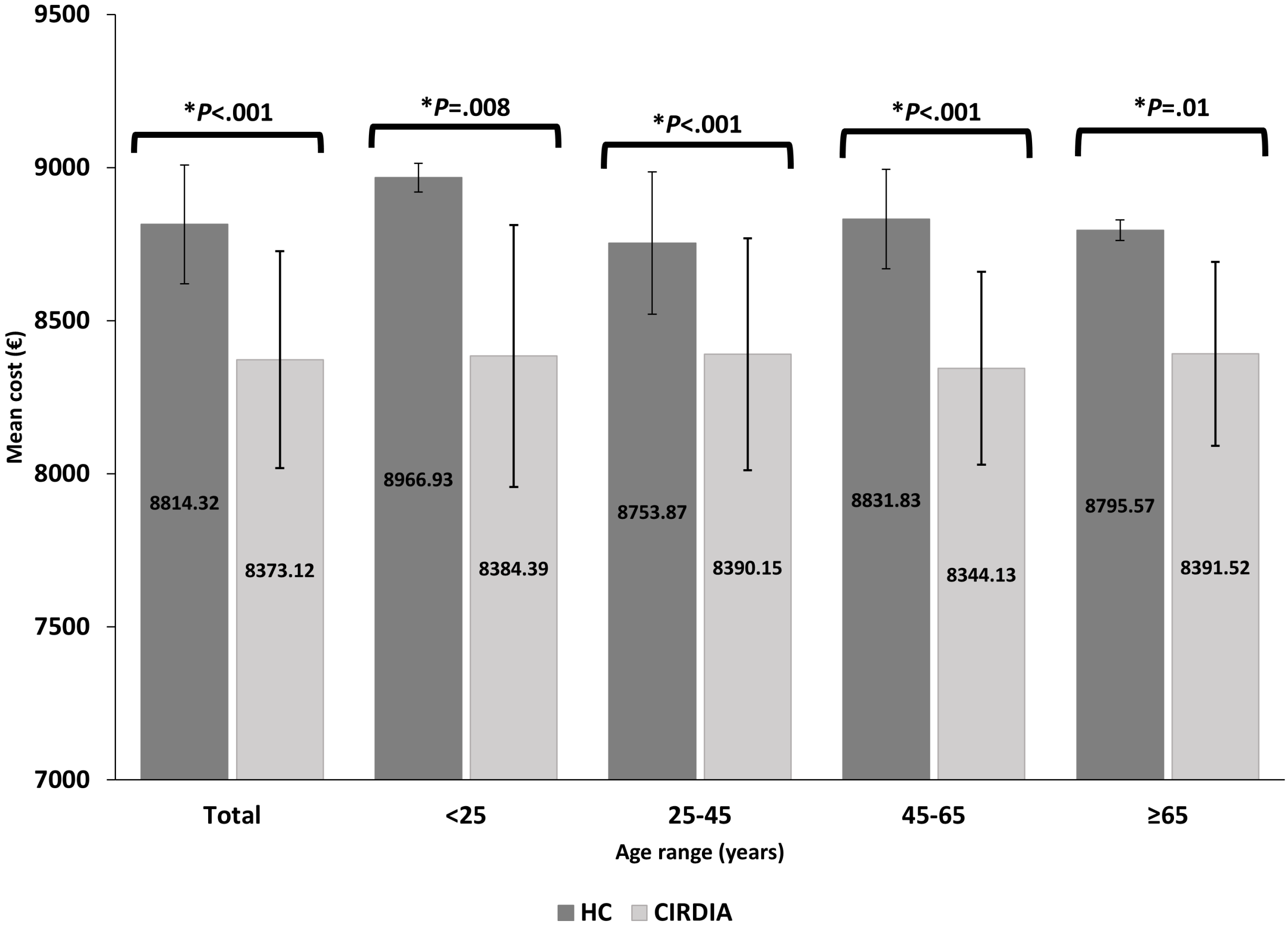
[a]Values were compared using the Wilcoxon rank sum test.
[b]CIRDIA: Inter-Regional Center for Automated Insulin in Diabetes.
[c]HC: hospital center.
[d]M0: closed-loop initiation.
[e]TIR: time in range 70 - 180 mg/dL.
[f]GRI: glycemia risk index.
[g]M12: 12 months after closed-loop initiation.

Subgroup analyses revealed no statistically significant differences between the CIRDIA and HC at M12, except among patients older than 65 years, for whom CIRDIA participants had higher TIR and lower GRI values (*P*=.03 and *P*=.01, respectively).

Costs Outcomes

We combined all cost items by type of procedure, year, and data source. Costs are expressed in euros from the French National Health Insurance perspective, and Figure 3 shows the mean costs for both comparison arms and by subgroup.

Figure 3. Average costs per patient according to the care setting. €1=US $1.10 at the time of the study. CIRDIA: Inter-Regional Center for Automated Insulin in Diabetes; HC: hospital centers.



The total cost of CL insulin therapy management was €1,077,231 (1 €=1.10 US $ at the time of the study) for 128 patients initiated in the CIRDIA, which was a mean cost of €8373.12 (SD 427.3) per patient. In the HC group, the total cost was €645,991 for 73 patients, which was a mean cost of €8814.32 (SD 192) per patient. Out-of-hospital–based management was associated with significantly lower costs (*P*<.001). All cost components are shown in Table S1 in Multimedia Appendix 1.

Incremental Cost-Effectiveness Ratio

The base-case analysis, using mean parameter values, indicated that the CIRDIA was less costly while achieving comparable effectiveness and safety to HC. This situation corresponds to dominance, with an estimated saving of €626 per additional percentage point of TIR and €2011 per point reduction in GRI, indicating that CL initiation in HC is dominated by the CIRDIA. The detailed results are presented in Table 3.

Table . Base-case cost-effectiveness and cost-safety results.

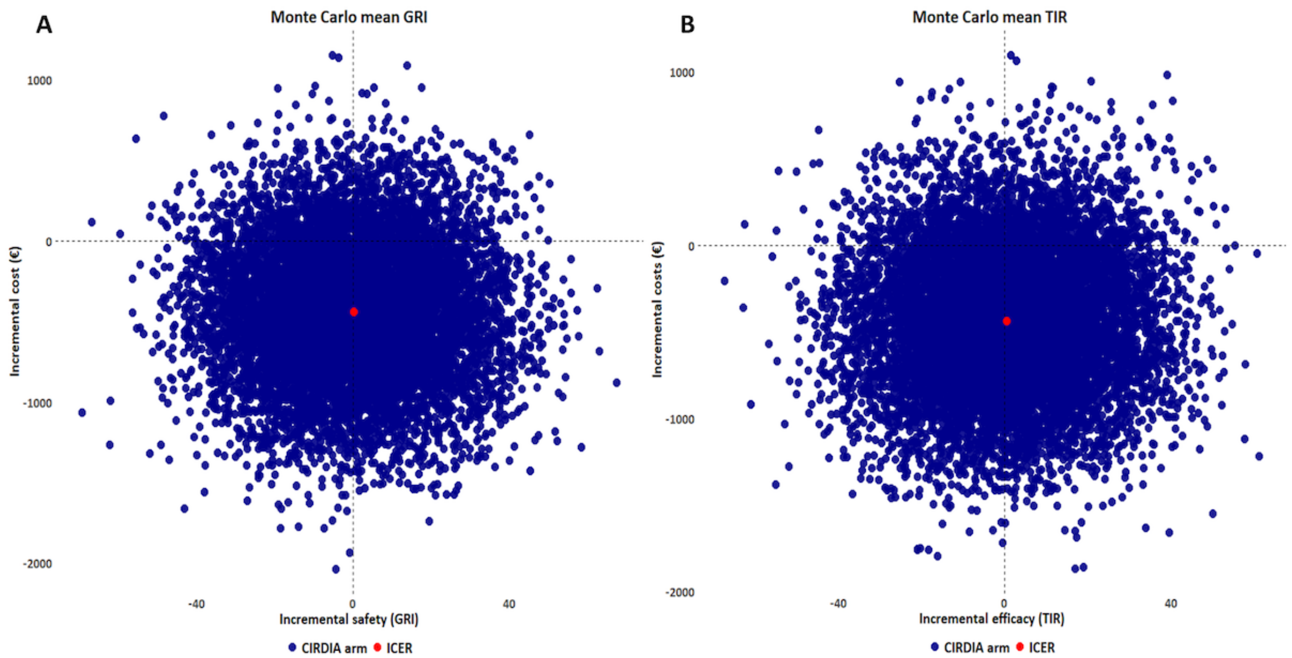
Parameters	CIRDIA[a]	HC[b]
Costs per patient (€[c])	8373.12	8814.32
Incremental costs (€)	−441.20	N/A[d]
Mean efficacy (TIR[e])	72.65	71.95
Incremental efficacy (increase in TIR)	0.70	N/A
Mean safety (reduction in GRI[f])	30.11	30.33
Incremental safety	−0.22	N/A
ICER[g]	−625.83	N/A
ICSR[h]	2011.02	N/A

[a]CIRDIA: Inter-Regional Center for Automated Insulin in Diabetes.
[b]HC: hospital center.
[c]€1=US $1.10 at the time of the study.
[d]N/A: not applicable.
[e]TRI: time in range.
[f]GRI: glycemia risk index.
[g]ICER: incremental cost-effectiveness ratio (based on the increase of time in range).
[h]ICSR: incremental cost-safety ratio (based on the decreased of glycemia risk index).

Sensitivity Analysis

In the PSA (10,000 simulations), the CIRDIA was less costly in 8600 (86%) of the cases. Strong dominance (less costly and more effective) was observed in 4340 (43.4%) of the simulations, while in 4270 (42.7%) of the simulations, the CIRDIA was less costly but less effective. The probability of being more or less effective was generally consistent with the base-case results. Only 1400 (14%) of the simulations placed the CIRDIA in a more costly position, being either less effective (n=700, 7.0%) or more effective (n=690, 6.9%). The scatter plot of the incremental cost-effectiveness plane is presented in Figure 4.

Figure 4. Probabilistic sensitivity analysis of 10,000 Monte Carlo simulations using glycemia risk index (GRI) (A) or time in range (TIR) (B) Inter-Regional Center for Automated Insulin in Diabetes (CIRDIA). €1=US $1.10 at the time of the study. ICER: incremental cost-effectiveness ratio.



Discussion

Principal Findings

The aim of this study was to assess the cost-effectiveness of CL initiation by the CIRDIA and HC from a health care payer perspective. We examined the relationship between initiation models and the increase in TIR or reduction in GRI and whether or not an out-of-hospital CL initiation and follow-up can be achieved in a cost-effective manner compared to the usual hospital management. In our cohort, CL initiation through the CIRDIA was associated with comparable TIR and GRI values at 12 months compared to HC initiation (P=.60 and P=.91, respectively), while being consistently less costly (P<.001), although baseline TIR was lower and baseline GRI was higher in the CIRDIA group. Sensitivity analyses further supported these results, confirming that the CIRDIA generally remained less costly than HC across a wide range of parameter variations.

Prior Work

To our knowledge, this is the first cost-effectiveness analysis comparing hospital-based and out-of-hospital CL initiation. However, our findings are consistent with previous studies, such as Böhme et al [29], which reported no significant differences in effectiveness between outpatient care and hospital settings in therapeutic education programs for patients with type 2 diabetes in France. Similarly, Cavassini et al [30] reported that the outpatient management of gestational diabetes was more cost-beneficial than hospital-based care in Brazil, underlining the potential economic advantages of ambulatory strategies. In the United Kingdom, Pulleyblank et al [31] also found that treatment setting had a significant impact on costs in patients with type 2 diabetes, with outpatient follow-up being less resource-intensive than hospital-based management.

Moreover, recent studies have shown that transitioning to CL reduces the GRI at 1 year [32-34], which is consistent with the trend observed in our cohort.

Strength

One major strength of this study is the use of real-world French data, but many published economic evaluations of CL systems have so far relied mainly on modeled analyses conducted in the United States and the United Kingdom. Furthermore, the use of TIR and GRI as primary end points is relatively novel in economic evaluations, allowing for the integration of a clinically relevant weighting of risk in the assessment of glycemic control [19,35].

Finally, sensitivity analyses and subgroup explorations provided additional insights into the robustness of our results, supporting the finding that CIRDIA and HC achieved broadly comparable outcomes in terms of TIR and GRI, whereas at baseline, TIR was lower and GRI higher in the CIRDIA participants. This also underlines that prior to CL initiation, patients followed in out-of-hospital settings do not have better glucose control than those followed in hospital centers, at least in our population.

Limitations

This study has several limitations.

First, the relatively small sample size limits the representativeness of the study population and, consequently, the robustness of the conclusions.

Second, there was an imbalance in baseline efficacy and safety outcomes between groups, which could have led to selection bias. To address this, we performed inverse probability of treatment weighting to adjust for sociodemographic characteristics as well as baseline efficacy and safety measures. After weighting, the cost advantage of the CIRDIA was maintained, and the results on effectiveness and safety suggested a potential benefit, although these should be interpreted cautiously given the limited sample size (data not shown). However, the patients who chose to start CL therapy in the CIRDIA setting might be different from the patients from the HC group in terms of prior education or other characteristics. A prospective study with better characterizations of these items will be needed.

Third, the 1-year time horizon restricts the evaluation to the short term and does not allow assessment of long-term effectiveness or costs, although this choice was justified by the specific objective of analyzing the initiation phase of CL.

Fourth, because the costs were assessed using French Health Insurance (Assurance Maladie) rates, the results may not be generalizable to other health care systems. However, this study suggests that CL initiation in an outpatient setting is feasible, safe, and probably less expensive than the inpatient setting, regardless of the health care system.

Fifth, we cannot exclude a bias in the recruitment of HC patients as it is possible that the patients sent to Santelys home health care provider by the hospital teams might have a different (here better) control compared to other HC patients. However, as patients are from 5 different hospitals, it is unlikely that this happened in all of the hospitals.

Finally, missing information on complications and comorbidities in the hospital arm may have led to an underestimation of certain costs (eg, retinopathy-related tests), although this does not appear to alter the overall trend observed.

Nevertheless, the data from the French Closed-Loop Observatory (OB2F) indicate that outpatient initiation is already widespread, reinforcing the relevance of investigating this organizational model [18].

Conclusion

This cost-effectiveness analysis compared 2 models of CL initiation for patients with T1D: a conventional hospital-based model and an out-of-hospital–based model supported by the CIRDIA.

Although baseline TIR was lower and baseline GRI was higher in the CIRDIA out-of-hospital setting compared to the HC setting, our results showed no significant differences in efficacy or safety outcomes between the 2 approaches. However, the CIRDIA setting was associated with lower management costs. While the patients who choose to initiate a CL system in the CIRDIA setting are probably not the same as the patients who choose to initiate CL in hospitals, these real-life findings suggest that the CIRDIA may represent a viable alternative organizational model for CL initiation in France, as it combines efficacy and savings.

Future research should assess whether these results hold over longer time horizons (eg, 5 or even 10 y) and from broader perspectives, such as a societal perspective that incorporates quality of life and indirect costs. Such work would enable cost-utility analyses to complement our cost-effectiveness findings.

From a policy perspective, the resources saved through out-of-hospital CL initiation could be reallocated to organizations such as the CIRDIA, which bring together highly trained diabetologists and uphold high-quality standards. This would allow persons living with T1D to choose their CL initiation setting, ensure early access to new technologies, and benefit the overall health care system through a cost-effective model.

Acknowledgments

We are grateful to (by alphabetical order) Dr Fabienne-Amiot-Chapoutot (Inter-Regional Center for Automated Insulin in Diabetes [CIRDIA], Troyes, France), Dr Véronique Averous (Douai Hospital), Dr Blandine Courbebaisse (CIRDIA, Fontenay-sous-Bois, France), Dr Vincent Degros (Valenciennes Hospital), Dr Joëlle Dupont (CIRDIA and HPL, St Etienne, France), Dr Salha Fendri (Amiens University Hospital), Dr Emmanuelle Lecornet-Sokol (CIRDIA and IPE, Paris, France), Dr Christine Lemaire (Bethune hospital), Dr Estelle Personeni (CIRDIA, Chatenoy-le-Royal, France), Dr Sylvie Picard (CIRDIA, Dijon, France), Pr Anne Vambergue (Lille University Hospital), and Dr Rihab Caillier (Santélys, Loos, France) for their assistance with the recruitment phase and Imane Boukhers, PhD (Santelys, Loos, France), for the editorial support. This study would not have been possible without the support of the CIRDIA.

Funding

This work was supported by a research grant from Santélys, Loos, France. Outside of this work, the CIRDIA has also received a research grant form the Clement-Drevon Foundation, Dijon, France; an award from the FENAREDIAM; and donations from the French Diabetic Association Burgundy-Franche-Comté and Côte d'Or and from the French Mutualité Jura.

Data Availability

The data used in this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Authors' Contributions

Conceptualization: FG, SB, SP, TF
Data curation: AR, FB, MN, SP
Formal analysis: AR, MN, TF
Funding acquisition: SP
Investigation: FB, SP
Methodology: AR, FB, SB, SL, TF
Project administration: SP
Supervision: AR, SB, TF
Validation: AR, FB, FG, SB, SL, TF
Writing – original draft: MN
Writing – review and editing: SB, SL, SP, TF

Conflicts of Interest

None declared related to this study. Outside of this work, SP has received consulting and/or speaking fees from Abbott, Air Liquide, Asdia, Dexcom, Insulet, Isis Diabete, Lilly, Medtronic, NHC, Novo Nordisk, Orkyn, Sanofi, and VitalAire. SL has received consulting and/or speaking fees from Abbott, Insulet, Medtronic, and Dexcom. SB has received consulting and/or speaking fees from Dexcom, Insulet, Lilly, and Medtrum.

Multimedia Appendix 1
Costs components for closed-loop system management.
[DOCX File, 18 KB - diabetes_v11i1e86690_app1.docx]

References

1. Diabetes. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/diabetes [accessed 2026-01-14]
2. People with diabetes, in 1,000s. International Diabetes Federation. URL: https://diabetesatlas.org/data-by-indicator/diabetes-estimates-20-79-y/people-with-diabetes-in-1000s/ [accessed 2026-01-14]

3. Jarraud C, Piotto E, Leclerc S, Thomassin C, Galmiche H. Le Diabète en France: nouvelles données épidémiologiques [Article in French]. J Epidemiol Popul Health 2025 Mar;73:202923. [doi: 10.1016/j.jeph.2025.202923]
4. Franc S, Schaepelynck P, Tubiana-Rufi N, et al. Mise en place de l'insulinothérapie automatisée en boucle fermée: position d'experts français [Article in French]. Med Mal Metab 2020 Sep;14(5S):S1-S40. [doi: 10.1016/S1957-2557(20)30003-1]
5. Renard E. L'insulinothérapie dans le diabète de type 1: un paysage en constante évolution [Article in French]. Med Mal Metab 2023 Feb;17(1):58-63. [doi: 10.1016/j.mmm.2022.12.001]
6. Benhamou PY, Lablanche S, Vambergue A, et al. The beneficial effects of closed-loop insulin delivery in patients with highly unstable type 1 diabetes eligible for islet transplantation are maintained over 6 months: an extension study of the DBLHU-WP10 trial. Diabetes Obes Metab 2022 May;24(5):956-961. [doi: 10.1111/dom.14654] [Medline: 35099098]
7. Murphy HR, Kumareswaran K, Elleri D, et al. Safety and efficacy of 24-h closed-loop insulin delivery in well-controlled pregnant women with type 1 diabetes: a randomized crossover case series. Diabetes Care 2011 Dec;34(12):2527-2529. [doi: 10.2337/dc11-1430] [Medline: 22011408]
8. Wang LR, Malcolm J, Arnaout A, Humphrey-Murto S, LaDonna KA. Real-world patient experience of long-term hybrid closed-loop insulin pump use. Can J Diabetes 2021 Dec;45(8):750-756. [doi: 10.1016/j.jcjd.2021.02.006] [Medline: 33958309]
9. Choudhary P, Arrieta A, van den Heuvel T, Castañeda J, Smaniotto V, Cohen O. Celebrating the data from 100,000 real-world users of the MiniMed. Diabetes Technol Ther 2024 Mar;26(S3):32-37. [doi: 10.1089/dia.2023.0433] [Medline: 38377326]
10. New data highlights the clinical impact of the MiniMed. Medtronic. 2024 Nov 18. URL: https://news.medtronic.com/New-data-highlights-the-clinical-impact-of-the-MiniMed-TM-780G-system-in-Asia-Pacific [accessed 2026-01-14]
11. Le remboursement du 1er système de boucle fermée hybride: retour sur un long chemin pour l'amélioration de la qualité de vie des personnes diabétiques [Article in French]. Fédération Française des Diabétiques. 2021 Nov 18. URL: https://www.federationdesdiabetiques.org/federation/actualites/le-remboursement-du-1er-systeme-de-boucle-fermee-hybride-retour-sur-un-long-chemin-pour-lamelioration-de-la-qualite-de-vie-des-personnes-diabetiques [accessed 2026-01-14]
12. Raffier N. Le pancréas artificiel allège le fardeau du diabète [Article in French]. Medscape. 2023 May 3. URL: https://francais.medscape.com/voirarticle/3610024 [accessed 2026-01-14]
13. Picard S, Personeni E, Dupont J, et al. Près de la moitié des personnes vivant avec un diabète de type 1 atteignent les objectifs internationaux de glucose 6 mois après l'initiation d'une boucle fermée hybride hors hôpital: résultats du CIRDIA [Article in French]. Ann Endocrinol (Paris) 2024 Oct;85(5):371-372. [doi: 10.1016/j.ando.2024.08.075]
14. Choix méthodologiques pour l'évaluation économique à la HAS [Report in French]. : Haute Autorité de Santé; 2020 Jul 2 URL: https://www.has-sante.fr/upload/docs/application/pdf/2020-07/guide_methodologique_evaluation_economique_has_2020_vf.pdf [accessed 2026-01-14]
15. Le règlement général sur la protection des données - RGPD [Article in French]. CNIL - Commission Nationale de l'Informatique et des Libertés. URL: https://www.cnil.fr/fr/reglement-europeen-protection-donnees [accessed 2026-01-14]
16. Recherches n'impliquant pas la personne humaine, études et évaluations dans le domaine de la santé [Report in French]. : Commission Nationale de l'Informatique et des Libertés; 2025 URL: https://www.cnil.fr/sites/default/files/2024-01/mr-004.pdf [accessed 2026-01-14]
17. Picard S, Lecornet-Sokol E. Le CIRDIA, un nouveau concept d'organisation des soins pour le diabète de type 1 en France [Article in French]. Med Mal Metab 2025 May;19(3):235-241. [doi: 10.1016/j.mmm.2025.01.013]
18. Raffier N. SFD 2024 – en « vie réelle », le pancréas artificiel fait ses preuves ! [Article in French]. Univadis. 2024 Apr 8. URL: https://www.univadis.fr/viewarticle/sfd-2024-vie-r%25C3%25A9elle-pancr%25C3%25A9-artificiel-fait-ses-2024a10006ea [accessed 2026-01-14]
19. Raffier N. TITR et GRI: les critères mesure continue du glucose de demain? [Article in French]. Medscape. 2025 Apr 25. URL: https://francais.medscape.com/voirarticle/3612807 [accessed 2026-01-14]
20. ALD n° 8 - diabète de type 1 chez l'adulte [Report in French]. : Haute Autorité de Santé; 2014 Apr 24 URL: https://www.has-sante.fr/jcms/c_419399/fr/ald-n-8-diabete-de-type-1-chez-l-adulte [accessed 2026-01-14]
21. Mallone R, Bismuth E, Thivolet C, et al. Dépistage et prise en charge du diabète de type 1 préclinique, stade 1–2. Prise de position d'experts français [Article in French]. Med Mal Metab 2024 Sep;18(5):405-432. [doi: 10.1016/j.mmm.2024.06.003]
22. Tarifs conventionnels des médecins généralistes et spécialistes [Article in French]. L'Assurance Maladie. 2026 Jan 15. URL: https://www.ameli.fr/roubaix-tourcoing/medecin/exercice-liberal/facturation-remuneration/consultations-actes/tarifs/tarifs-conventionnels-medecins-generalistes-specialistes [accessed 2026-01-14]
23. CCAM en ligne [Article in French]. L'Assurance Maladie. URL: https://www.ameli.fr/accueil-de-la-ccam/index.php [accessed 2026-01-14]
24. Téléchargement de la TNB version 101 [Article in French]. L'Assurance Maladie. URL: http://www.codage.ext.cnamts.fr/codif/nabm/telecharge/index_tele.php?p_site=AMELI [accessed 2025-09-8]
25. Liste des produits et prestations [Article in French]. L'Assurance Maladie. URL: http://www.codage.ext.cnamts.fr/codif/tips/index_presentation.php?p_site=AMELI [accessed 2026-01-14]

26. Tarifs MCO et HAD [Article in French]. Agence technique de l'information sur l'hospitalisation. URL: https://www.atih.sante.fr/tarifs-mco-et-had [accessed 2026-01-14]
27. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. Value Health 2022 Jan;25(1):3-9. [doi: 10.1016/j.jval.2021.11.1351] [Medline: 35031096]
28. Briggs AH, Weinstein MC, Fenwick EAL, et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–6. Value Health 2012;15(6):835-842. [doi: 10.1016/j.jval.2012.04.014] [Medline: 22999133]
29. Böhme P, Durain-Sieffert D, Fay R, et al. P89 Efficacité à court terme d'un réseau de ville centré sur l'éducation thérapeutique du patient diabétique de type 2: comparaison avec un service hospitalo-universitaire [Article in French]. Diabetes Metab 2010 Mar;36:A60. [doi: 10.1016/S1262-3636(10)70237-8]
30. Cavassini ACM, Lima SAM, Calderon IMP, Rudge MVC. Cost-benefit of hospitalization compared with outpatient care for pregnant women with pregestational and gestational diabetes or with mild hyperglycemia, in Brazil. Sao Paulo Med J 2012;130(1):17-26. [doi: 10.1590/s1516-31802012000100004] [Medline: 22344355]
31. Pulleyblank R, Laudicella M, Olsen KR. Cost and quality impacts of treatment setting for type 2 diabetes patients with moderate disease severity: hospital- vs. GP-based monitoring. Health Policy 2021 Jun;125(6):760-767. [doi: 10.1016/j.healthpol.2021.03.009] [Medline: 33789816]
32. Picard S, Courbebaisse B, Dupont J, et al. Glycemia Risk Index (GRI) and international glucose targets before and 6 months after initiation of hybrid closed loop system in the CIRDIA, a French multisite out-of-hospital center. Diabetes Metab 2025 Mar;51(2):101617. [doi: 10.1016/j.diabet.2025.101617] [Medline: 39922278]
33. Resmini E, Zarra E, Dotti S, et al. Impact on Glycemia Risk Index and other metrics in type 1 adult patients switching to Advanced Hybrid Closed-Loop systems: a one-year real-life experience. Eur J Med Res 2024 Jul 15;29(1):365. [doi: 10.1186/s40001-024-01946-w] [Medline: 39004734]
34. Castorani V, Rigamonti A, Frontino G, et al. Turning the tides: achieving rapid and safe glucose control in adolescents with suboptimally controlled type 1 diabetes using advanced hybrid closed loop systems. Front Endocrinol (Lausanne) 2024;15:1243565. [doi: 10.3389/fendo.2024.1243565] [Medline: 38628580]
35. Battelino T, Alexander CM, Amiel SA, et al. Continuous glucose monitoring and metrics for clinical trials: an international consensus statement. Lancet Diabetes Endocrinol 2023 Jan;11(1):42-57. [doi: 10.1016/S2213-8587(22)00319-9] [Medline: 36493795]

Abbreviations

CGM: continuous glucose monitoring
CHEERS: Consolidated Health Economic Evaluation Reporting Standards
CIRDIA: Inter-Regional Center for Automated Insulin Therapy
CL: closed-loop
DH: day hospitalization
GRI: glycemia risk index
HAS: French National Health Authority
HC: hospital center
ICER: incremental cost-effectiveness ratio
PSA: probabilistic sensitivity analysis
SFD: French-Speaking Diabetes Society
T1D: type 1 diabetes
TIR: time in range

Edited by A Stone; submitted 28.Oct.2025; peer-reviewed by B Detournay, J Lu; revised version received 28.Dec.2025; accepted 07.Jan.2026; published 29.Jan.2026.

Please cite as:
Napame M, Picard S, Foglia T, Rubenstrunk A, Baudoux F, Giroud F, Lablanche S, Borot S
Inter-Regional Center for Automated Insulin in Diabetes (CIRDIA) and Hospital-Based Approaches to Closed-Loop Therapy in Type 1 Diabetes: Cost-Effectiveness Analysis
JMIR Diabetes 2026;11:e86690
URL: https://diabetes.jmir.org/2026/1/e86690
doi:10.2196/86690

Exploring the REACHOUT Mental Health Support App for Type 1 Diabetes From the Perspectives of Recipients and Providers of Peer Support: Qualitative Study

Debbie Lam[1], MSc; Poonamdeep Jhajj[2], MD; Diana Sherifali[3,4,5], RN, CDE, PhD; Frances S Chen[6], PhD; Tricia S Tang[7], PhD, RPsych

[1]Experimental Medicine Program, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[2]Research Institute, BC Children's Hospital, Vancouver, BC, Canada
[3]School of Nursing, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
[4]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
[5]Heather M. Arthur Population Health Research Institute/Hamilton Health Sciences Chair in Inter-Professional Health Research, Hamilton, ON, Canada
[6]Department of Psychology, Faculty of Arts, University of British Columbia, Vancouver, BC, Canada
[7]Division of Endocrinology, Faculty of Medicine, University of British Columbia, 2775 Laurel Street, Vancouver, BC, Canada

Corresponding Author:
Tricia S Tang, PhD, RPsych
Division of Endocrinology, Faculty of Medicine, University of British Columbia, 2775 Laurel Street, Vancouver, BC, Canada

Abstract

Background: Existing qualitative research in peer support interventions has largely focused on the recipients of support rather than those delivering support. Exploring the perspectives of both roles may provide a holistic understanding of the peer support experience.

Objective: This study elicits the experiences of recipients and providers of support who participated in REACHOUT, a 6-month peer-led mental health support intervention delivered via mobile app for adults with type 1 diabetes. REACHOUT offered multiple support delivery modalities (one-on-one, group-based texting, and virtual face-to-face small group sessions) that could be customized by recipients.

Methods: A total of 32 study participants (recipients and peer supporters) attended focus group discussions following the completion of REACHOUT. Thematic analysis was performed in an inductive approach.

Results: Four major themes were identified by thematic analysis: (1) need for a sense of community and belonging, (2) factors to enhance the recipient-peer supporter experience, (3) key aspects of the peer supporter experience, and (4) importance of personalizing the user experience while using the REACHOUT mobile app. REACHOUT successfully fostered connectedness by bringing together adults with type 1 diabetes who previously felt isolated. Recipients felt greater agency when given the opportunity to self-select a peer supporter. The main factors considered during the matching process included insulin delivery and glucose monitoring systems, duration of diabetes, shared hobbies, life stage, and age. While support was designed to be unidirectional from peer supporter to recipient, the former also derived benefits. Peer supporters expressed the need for greater guidance around navigating boundaries and responding to emotionally charged conversations. Finally, the REACHOUT app was able to accommodate a heterogeneity of support needs by offering one-on-one and group support across multiple communication platforms including text, audio, and video.

Conclusions: The success of peer-led mental health support interventions such as REACHOUT is likely associated with the recipient-peer supporter dynamic. By offering a range of support delivery and communication modalities, participants can better personalize solutions to meet their unique support needs. Understanding the perspectives of both recipients and peer supporters is essential to refining interventions and optimizing digitally delivered mental health support models.

(JMIR Diabetes 2026;11:e72779) doi:10.2196/72779

KEYWORDS

diabetes; mental health; mHealth; mobile app; mobile health; peer support; qualitative; thematic analysis; type 1; type 1 diabetes

Introduction

Peer support is a promising self-management strategy to improve emotional health in chronic illness care [1-4]. In the context of diabetes, several systematic reviews of adults with diabetes (both type 1 and type 2 diabetes) have found peer support interventions to be associated with improved clinical, behavioral, and psychosocial (quality of life, perceived social support) outcomes [5-8]. However, to better understand the processes underlying these positive changes, it is important to explore the qualitative experience of giving and receiving peer support.

While qualitative research on peer support interventions has focused largely on the experiences of those who receive support [9-11], there has been a notable increase in studies focused on the individuals who deliver support [12-19]. However, the optimal model for understanding the peer support experience is to explore the perspectives of both parties involved. To date, there have been 4 qualitative studies that have investigated the experiences of both recipients of support and peer supporters in the context of diabetes [20-23]. Of these studies, only 1 recruited adults with type 1 diabetes (T1D) as part of a larger sample [21], while the other investigations targeted adults with type 2 diabetes [20,22,23].

In the era of digital health, peer support models in diabetes have been made more accessible through the shift to virtual platforms such as mobile apps. Such digital peer support programs are especially valuable in rural and remote areas, where access to traditional peer networks and diabetes programs can be limited [24-26]. A systematic review of in-person and technology-mediated peer support for adults with diabetes found that peer support was beneficial in reducing isolation and increasing social support for recipients [27]. However, none of these studies were specific to T1D only. Interestingly, in a review of technology for peer support intervention for adolescents with chronic illness, rather than adults, T1D was the most represented condition [28]. Generally, adolescents with T1D experienced benefits in emotional support and diabetes management [29]. Of the few studies utilizing mobile or web apps for T1D adults, peer support was a secondary feature to self-management behavior education or one of multiple intervention components rather than the main focus [30-33]. As T1D is a lifelong condition, it is important to offer ongoing mental health support to adults living with T1D, especially those facing geographical or resource barriers.

Methods

Study Aim

This study aimed to explore the experiences of and perspectives from recipients and providers of support on REACHOUT, a peer-led mental health support intervention for adults with T1D living in rural and remote regions of British Columbia, Canada.

Study Design

Following the completion of the pilot trial titled REACHOUT, which investigated the feasibility and acceptability of peer-led mental health support intervention delivered by a mobile app, we conducted focus groups with participants of the study. The reporting of methods and findings adheres to the COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist (Checklist 1) [34].

REACHOUT Intervention Description

Described in detail elsewhere, the REACHOUT pilot investigated the impact of a mobile app that delivered mental health support to adults with T1D living in Interior British Columbia over a period of 6 months. REACHOUT offered multiple support delivery modalities (one-on-one, group-based texting, and virtual face-to-face small group sessions that could be customized by recipients) [35]. Participants include individuals who receive support (recipients) and those who provide support (peer supporters). In this paper, the term "participants" will only be used when addressing both recipients and peer supporters. The eligibility criteria for recipients were as follows: (1) be diagnosed with T1D, (2) be at least 18 years or older, (3) speak English, (4) have access to the internet and/or a smartphone, (5) live in the interior region of British Columbia, and (6) have a mean subscale score of ≥2 on the type 1 Diabetes Distress Scale [36]. Peer supporters had similar requirements with the exceptions of criteria 5 and 6. They also had to be willing to complete a 6-hour training program. Training components and competency evaluation are published elsewhere [37]. It should be noted that if asked a medical question by recipients, peer supporters were instructed to refrain from answering and defer to the diabetes nurse educator.

The REACHOUT app offered multiple support delivery modalities including one-on-one support provided by a recipient-selected peer supporter, group texting support via the 24/7 chat room, and small group face-to-face support via video huddles and happy hours. Recipients were encouraged to use any or all modalities as frequently as desired. Peer supporters were invited to attend virtual wellness sessions to debrief their experiences as well as receive their own emotional support. Finally, the ongoing monitoring of group-based communication exchanges was performed by the research team, and fidelity assessments were conducted at 1, 3, and 5 months of the intervention with all participants.

Ethical Considerations

This qualitative descriptive study was approved by the University of British Columbia Behavioural Research Ethics Board (H20-00276). Prior to focus groups, participants provided e-informed consent using REDCap (Research Electronic Data Capture) electronic data capture tools hosted at the University of British Columbia [38,39]. To maintain privacy and confidentiality, recordings were anonymized to omit personal identifying information and stored securely. Only the study team could access study data. Upon completion, participants received a CAD $25 (approximately US $18) e-gift card.

Participant Recruitment and Sampling

Following the completion of the pilot trial REACHOUT, all those in recipient roles were contacted by a research assistant and invited to the postintervention focus groups to share their experience with the REACHOUT program and app and suggestions for improvement. Only peer supporters who had been paired with recipients were invited to join the postintervention focus groups. Those who provided consent were interviewed.

Data Collection and Analysis

Focus groups were conducted online using Zoom; video and audio were recorded and later transcribed. Led by a female researcher (TST), focus groups were stratified into recipient versus peer supporter-only membership with approximately 6 individuals per group. The interview guide (Multimedia Appendix 1) used open-ended questions and prompts to elicit discussion around their experience in the program, peer support interactions, and app usage. Follow-up questions were posed if clarification or explanation was needed.

Recordings were transcribed verbatim and anonymized with participant roles (recipient or peer supporter) identified to capture perspectives from both groups. Transcripts were analyzed using NVivo V.14 software package [40]. Guided by an interpretivist research paradigm, which centers around subjective experiences [41], we selected an inductive thematic approach to support the possible variation of participant perceptions. Following Braun and Clarke's 6 phases of thematic analysis, 1 coder (DL) participated in transcribing the data and another coder (PJ) who had no involvement in the interview guide development, interviews, and transcription familiarized themselves with the transcripts [42]. Both coders discussed initial ideas before independently performing open coding. The coders discussed the findings after every round of coding to enhance reflexivity and iteratively refine a unified codebook. Independently coded transcripts were combined, and codes were sorted and combined to form themes and subthemes. Themes and subthemes were reviewed and refined with clear definitions and names. Findings and any discrepancies were discussed with the principal investigator (TST) and another coauthor member (DS) who was not involved in the interview guide creation and interviews. Moreover, this was a recursive process where analysis phases moved back and forth as needed [42].

Positionality Statement

Our multidisciplinary team comprises cisgender, heterosexual women from East Asian, South Asian, and European settler backgrounds. TST has over 25 years of experience working in peer support, and her research focuses on developing models to improve mental health outcomes in high-risk and medically underserved communities. DS has over 25 years of research working in diabetes self-management at the community and provider level. FSC has over 20 years of experience working on topics related to stress, social support, and social connection and contributes a behavioral science perspective. DL and PJ are early-career researchers with master's and medical graduate training. All authors are living in urban centers and are cognizant of their own privileges and practice reflexivity to ensure that priorities of the diabetes community are represented throughout the research process.

Results

Description of Sample

In total, 32 study participants (17 recipients and 15 peer supporters) who completed the REACHOUT intervention were recruited and interviewed from August to October 2022. The characteristics between focus group participants compared to nonrespondents in the pilot study population are noted in Multimedia Appendix 2. There were 9 focus groups lasting 60 - 90 minutes, 4 recipient-only groups, and 5 peer supporter-only groups. As summarized in Table 1, participants were predominantly women and Caucasian, with a mean age of 48 (SD 16.3; range 23 - 76) years and an average of 24 (SD 18.1; range 0 - 65) years living with diabetes. Most participants received postsecondary education and had a household income greater than CAD $70,000 (approximately US $50,505).

Table . Interviewed recipients' and peer supporters' baseline characteristics.

	Total focus group participants (n=32)	Recipients (n=17)	Peer supporters (n=15)
Age (y), mean (SD)	48 (16.3)	48 (16.6)	50 (16.4)
Diabetes duration (y), mean (SD)	24 (18.1)	25 (18.5)	23 (18.2)
Women, n (%)	26 (81)	15 (88)	11 (73)
Marital status, n (%)			
Never married	9 (28)	6 (35)	3 (20)
Married or living with a partner	20 (63)	10 (59)	10 (67)
Separated or divorced or Widow	3 (9)	1 (6)	2 (13)
Ethnicity, n (%)			
Aboriginal	1 (3)	1 (6)	0 (0)
Aboriginal/Caucasian	1 (3)	1 (6)	0 (0)
East Asian (Chinese, Korean, Japanese)	1 (3)	0 (0)	1 (7)
Caucasian	29 (91)	15 (88)	14 (93)
Education, n (%)			
High school graduate (or equivalent)	3 (9)	3 (18)	0 (0)
Some college or technical school	7 (22)	4 (24)	3 (20)
College graduate	10 (31)	3 (18)	7 (47)
Graduate degree	12 (38)	7 (41)	5 (33)
Pretax household income (CAD $), n (%)			
<70,000 (approximately US $50,505)	10 (31)	7 (41)	3 (20)
>70,000 (approximately US $50,505)	17 (53)	5 (29)	12 (80)
Declined to answer	5 (16)	5 (29)	0 (0)
Employment, n (%)			
Full-time job	12 (38)	6 (35)	6 (40)
Part-time job	6 (19)	5 (29)	1 (7)
Retired	6 (19)	2 (12)	4 (27)
Other	7 (22)	4 (24)	3 (20)
Declined to answer	1 (3)	0 (0)	1 (7)

Themes

Four overarching themes were identified and related to participants' experiences in the peer support intervention and on their user experience with the mobile app delivery (Table 2).

Table . Four major themes were identified by thematic analysis with subthemes that capture similarities and differences within and across recipient and peer supporter groups.

Theme	Recipient	Peer supporter	Both group
Need for a sense of community and belonging	—[a]	—	• Giving and receiving unconditional support • Reducing isolation in rural communities • Learning from real-life experiences of T1D peers
Factors to enhance the recipient-peer supporter experience	• Ability to select a peer supporter	—	• Modality and frequency of communication
Key aspects of the peer supporter experience	—	• Supporting peer supporters in their role • Benefits of being a peer supporter • Challenges of being a peer supporter	—
Importance of personalizing the user experience while using the REACHOUT mobile app	—	—	• Varied preferences in peer support • Adapting the mobile app to fit user expectations

[a]Not applicable.

Theme 1: Need for a Sense of Community and Belonging

For recipients and peer supporters, REACHOUT created a safe environment to build and strengthen connections with other adults who shared the lived experience of T1D. This sense of belonging and community spirit manifested in different ways.

Subtheme A: Giving and Receiving Unconditional Support

The intervention created a space to express concerns without fear of judgment or rejection. Participants who had felt completely alone in the past finally found their "tribe"—a community that experienced and understood the same fears, frustrations, and emotional burdens of T1D.

> *The whole thing has been just so rewarding and I think it's kind of brought me out a little bit too. Like being able to be who I am and not be judged it's like – it's just this community. Being able to kind of hop into the chat and say, "Oh yeah this is what happened to me" or you know, just that common sharing. It's been huge.* [Peer supporter 5-2]

Initially, some participants were hesitant to engage in group activities such as face-to-face virtual sessions because the possibility of meeting peers who were managing their diabetes "perfectly" could trigger feelings of inadequacy or resentment. However, once the intervention started, they realized others were willing to be vulnerable. For example, when some participants disclosed perceived self-management failures in the 24/7 chat room, they were met with empathy and validation. After this precedent was established, others felt safe to reveal moments of insecurity and self-blame.

> *It was really nice to know when you're like, "I'm doing everything possible to keep my blood sugar stable right now and for the life of me they're on the higher side. I don't know why." But knowing other people are like, "Yeah, isn't that frustrating," like they get it because they live it. It's not like your [endocrinologist], it's nice to hear it from somebody who lives it, I don't feel so alone in the world.* [Recipient 6-3]

Subtheme B: Reducing Isolation in Rural Communities

Coming from rural and small communities across Interior British Columbia, many recipients and peer supporters had never encountered another T1D adult in their local community. This sense of loneliness was particularly pronounced for individuals diagnosed late in life (eg, 45 years and older).

> *It seems like we grew up in a smaller town, and there wasn't anybody that had diabetes that I knew, and then going through the other parts of my life, I didn't have really anybody to talk to.* [Recipient 4-2]

Although REACHOUT was a virtual intervention, participants were comforted knowing that peers resided in nearby towns. When browsing through the peer supporter library, participants were able to identify the general location where each peer supporter lived and, therefore, felt reassured that face-to-face support was accessible if needed. As part of the REACHOUT community, participants were not left to cope with the struggles of T1D on their own.

> *I thought it was really nice to connect with people, maybe not totally in my community. But certainly, there have been a great number of people within an hour's drive that's connected with and there's just something about that to know that you're not alone in your little portion of the world.* [Recipient 6-2]

Subtheme C: Learning From Real-Life Experiences of Peers With T1D

With REACHOUT, participants had direct access to the most reliable and high-quality T1D information including "real-world" experiences from adults who used insulin, insulin pumps, and continuous glucose monitors daily. The mobile app offered different mechanisms to obtain the knowledge needed. For instance, in the 24/7 chat room, participants posted updates regarding changes to health insurance coverage or, during the COVID-19 period, shortages in various diabetes supplies. This platform was also a place to pose questions and elicit differing perspectives from both recipients and peer supporters. For example, participants who were considering transitioning to a different insulin pump or continuous glucose monitoring device could hear opinions from peers from diverse lifestyles and backgrounds.

> *It was cool to hear firsthand information from somebody's experience, say about the Omnipod or the Medtronic or Dexcom or whatever. I think that's invaluable, rather than just going to a doctor or endocrinologist and just a medical professional, which is still really good information but to get the user's perspective on something is kind of for sure.* [Peer supporter 7-1]

Notably, how participants preferred to learn varied. Those who were not comfortable posting messages or disclosing personal experiences still enjoyed reading the discussion threads and exchanges in the 24/7 chat room. Many participants routinely checked the app to read the most recent conversation and updates. While not directly participating, participants who passively monitored the exchange of dialogue derived substantial benefits.

> *In my journey over the years with diabetes, I just felt so alone, so this app has been – just knowing it's there has been huge. I'm kind of a classic introvert – I don't really go on and participate actively on it, but I do on in and I read the conversations and just I love it. Please don't underestimate power of that because it's really been a big thing for me.* [Peer supporter 7-2]

Theme 2: Factors to Enhance the Recipient-Peer Supporter Experience

Factors related to one's experience with REACHOUT were largely dependent on the quality of the recipient-peer supporter relationship. Many found their peer supporter extremely helpful and valued their time, but the strength of their relationship was influenced by various contributing factors.

Subtheme A: Ability to Select a Peer Supporter

Recipients felt empowered by the opportunity to choose their peer supporter. Some sought identical counterparts, while others envisioned their peers as potential mentors. The criteria that each recipient used to choose their peer supporter were unique and personal. The main factors included diabetes management system, duration of diabetes, shared activities, life stage, and age.

According to some recipients, diabetes and management-related factors weighed heavily into the selection process. For example, some recipients were seeking a peer supporter who had been living with diabetes for as long, if not longer, than themselves. Others felt a greater kinship with peer supporters using the same continuous glucose monitoring or insulin pump.

> *I looked at not necessarily insulin type, but just device that they might be using. And for me, the Dexcom was new so I wanted somebody who knew and used the Dexcom. So that was some of my criteria when I started to go through the list. I don't need to read the other fifteen that don't use a Dexcom, that was a clear priority for me.* [Recipient 6-3]

Lifestyle factors also factored in prominently when selecting a peer supporter. For instance, recipients who enjoyed exercising or engaging in outdoor sports preferred an equally active peer supporter. Having shared hobbies enhanced the quality of recipient-peer supporter relationships and extended conversations beyond the boundaries of diabetes. In contrast, in the absence of similar interests, some recipients found it difficult to establish meaningful and sustained rapport with their peer supporters.

> *Device for me wasn't as important. Cause I've been on both injections and pump. So for me, mostly activities and hobbies. And someone that liked to travel as well, cause I always find that quite daunting but I want to do more of that so yeah. I found a good person for that.* [Recipient 8-4]

The stage of life was equally important. For example, young mothers gravitated toward selecting peer supporters who were also raising children. As expected, navigating both diabetes and parenthood created strong connections. Similarly, older recipients who were retired understood the priorities and pace of others who were also no longer in the workforce.

> *I picked someone who was in a similar life stage as me, cause I've had diabetes for 30 years I don't really need advice on how to treat my diabetes. For me, it was much more the mental health connection and then transition to this new part of my life of being a mom. Because stuff would come up and I'd be like oh, my gosh, how do you deal with this? How do you prioritize a crying baby verses a low? So that for me was great.* [Recipient 9-1]

Age and/or length of diabetes experience emerged as critical factors in the selection process. Some recipients intentionally chose older peer supporters who had a lifelong journey with diabetes as they envisioned having a mentor who could provide insight on what challenges to expect over time. Rarely did recipients choose peer supporters who were much younger than themselves.

> *Someone [who] was male, and older than me. So I can relate to what they're going through, and someone who has had diabetes for longer than I have. So it's quite focused of what I was looking for. I was able to be paired up with someone who was in my*

position, but a couple years down the road. [Recipient 8-3]

Subtheme B: Modality and Frequency of Communication

Video conferencing was the most preferred modality, as it allowed for the 2 parties to observe facial expressions and body language. Different communication methods were utilized for different functions. Direct messaging, texting, and emails were ideal for quick communication such as check-ins and meeting coordination. If both parties were amenable to investing greater effort and commitment, more substantial conversations took place through video conferencing or phone calls.

Consistency formed the foundation of a strong recipient-peer supporter relationship. Initially, weekly communication was needed to establish and build rapport. However, as the relationship matured, for some, the frequency of contact slowed down as people had other competing life demands such as full-time jobs or home responsibilities. Mid-intervention, many acknowledged that the ideal schedule was contact once every 2 weeks.

> *I liked that it was once a week in the beginning. I think it gave you a lot of opportunity to get to know each other, tell each other your diagnosis story and then from there on. I think I did realize with my peer supporter when we started, when we were meeting every week that we almost were running out of things to update each other on or talk about. And then every two weeks was really great and then we had some things to share over the last two weeks.* [Recipient 8-1]

Theme 3: Key Aspects of the Peer Supporter Experience

The cornerstone of a peer-led intervention is the peer supporters who deliver mental health support. Although the goal of REACHOUT was to provide support to recipients, the sustained quality of the 6-month intervention provided opportunities for peer supporters to be nurtured as well.

Subtheme A: Supporting Peer Supporters in Their Role

To function effectively in their role, peer supporters underwent a 6-hour training. According to peer supporters, the most instrumental training activity was "role-plays." Not only did role-plays allow trainees to practice newly developed skills, but these simulated scenarios helped build their self-confidence and preparedness.

During the intervention, peer supporters appreciated having a workbook with structured activities to lead their recipients through. These activities served as a valuable foundation for conversations that would not occur organically—for example, identifying personal values and exploring sources of diabetes distress.

Furthermore, peer supporters benefited from attending wellness sessions hosted by the research team. Wellness sessions were Zoom-based and provided the opportunity for peer supporters to share stories, voice concerns, and pose questions to one another. Moreover, these discussions fostered camaraderie among peer supporters while navigating inherent challenges in their support roles.

> *I think every [Wellness] session – I found important, because there's always something new that you can take away. And then, if there's a question that I have, [I] can actually ask during those sessions. "Okay, you know. Great. I'm on the right track," you know as well and then, "I'm following what I supposed to be following and doing what I'm supposed to be doing with the peers."* [Peer supporter 1-1]

Subtheme B: Benefits of Being a Peer Supporter

Peer supporters derived deep satisfaction and intrinsic reward from their role, finding genuine fulfillment from providing mental health support to other adults with T1D. Through acts of altruism and compassion for the T1D community, they experienced satisfaction knowing that their contribution added meaning and value to the lives of their recipient.

Many peer supporters realized that their relationship with their matched recipient was mutually beneficial. Not only did peer supporters deliver emotional support, but recipients also shared their knowledge, coping strategies, and perspectives. Additionally, many peer supporters discovered a renewed connection with their own diabetes journey and engaged in self-reflection and self-development.

> *[My recipient] was fairly newly diagnosed, within the last year, and it's been 11 years for me. I benefited a lot from talking with her. It kind of re-engaged me in diabetes. I think I realized I've been coasting, and I needed to kind of re-engage, and I think that was really important for me.* [Peer supporter 2-2]

Subtheme C: Challenges of Being a Peer Supporter

Not all peer supporters had recipients who reciprocated with the same level of enthusiasm. Rather than feeling rejected if their recipient did not respond immediately, some peer supporters did not take it personally. Moreover, peer supporters found it challenging to sustain consistent communication with their recipients, especially in the last half of the intervention. Peer supporters tried to understand their recipient's perspectives by acknowledging the demands of personal and professional lives.

> *I found sending a text- something, I felt like I was chasing her. And I would think, "Oh maybe she doesn't want to talk to me anymore," "Maybe she's had enough," or, "Maybe I'm doing something wrong," but it wasn't anything like that at all. It was just she was busy; she has a job and family.* [Peer supporter 5-3]

Some peer supporters struggled to deepen their conversations when recipients appeared to be reluctant to broach more sensitive topics. At times, peer supporters adhered to surface-level conversations so as to not "over-step." As such, peer supporters suggested having more guidance on how to navigate boundaries and tips for gauging the depth recipients seek from relationships.

> *I didn't bring up the underlying issues as much as I would have expected, perhaps because I wasn't quite versed in how to bring those up. I didn't know if it was appropriate for me to kind of prod a little bit. […] I felt a little bit at a loss of how to bring up like these big concepts, psychological issues and things like that. There was definitely stuff going on, but it was hard for me to get them to speak about some of those things.* [Peer supporter 2-1]

Conversely, some peer supporters encountered recipients who openly shared their feelings and concerns, which posed a different challenge as it triggered feelings of worry and inadequacy. Peer supporters were seeking greater instructions on how to navigate these emotionally charged conversations. Two potential solutions suggested were (1) establishing clear guidelines on how to respond to questions requiring escalation to a health professional and (2) providing a set of prepared questions to ask when these situations arose.

> *I'm not gonna lie, I was a little bit stressed if this person was really in distress, because I don't know if I was like, "Jeez, like I don't know if I can be the guy that's going to help this person." But I was pretty fortunate, [my recipient] just wants someone to talk to, basically, which worked out well for me.* [Peer supporter 7-1]

Theme 4: Importance of Personalizing the User Experience While Using the REACHOUT Mobile App

Participants (recipients and peer supporters) had four ways to engage with others on the REACHOUT mobile app: (1) direct messaging, (2) 24/7 chat room, (3) virtual happy hours, and (4) virtual huddles.

Subtheme A: Varied Preferences in Peer Support

The 24/7 group chat room served as a central feature of the app with a significant amount of activity. Most participants referred to the 24/7 chat room to pose questions, share stories and updates, and initiate discussions. The high level of engagement led many participants to habitually check the chat to stay informed. For some, monitoring the 24/7 chat room was a part of their daily routine, as participants could obtain new information as well as be exposed to a diverse range of topics.

Alternatively, some found the continuous flow of information in the 24/7 chat room to be overwhelming. Specifically, it was burdensome to sift through a high number of messages to find discussions of personal relevance. For example, while the majority of participants discussed insulin pumps, it alienated the few individuals who used multiple daily injections. In extreme cases, some participants deactivated the notifications setting for the 24/7 chat room.

> *Like it was overwhelming right at the beginning [from the 24/7 chat room], and so I turned off the notifications but then I got it out of the habit of checking, so I missed a whole bunch of stuff, me and my mentor were communicating through text, so I didn't really have to worry about going back into the app.* [Recipient 9-1]

Virtual huddles and virtual happy hours were 2 additional support delivery mechanisms offered. The former was a larger-group interactive webinar led by peer supporters and/or professionals and required fewer social demands or active participation. The latter involved a smaller, intimate group discussion led by a peer supporter and fostered open and relaxed conversations beyond their one-on-one peer support relationship. These 2 support modalities cater to diverse personality types and needs.

Subtheme B: Adapting the Mobile App to Fit User Expectations

App usability issues centered largely around the lack of logical structure and flow of exchanges within the 24/7 chat room. Because participants had the option of responding within a thread or creating a new thread, conversations often seemed disjointed. As a result, many suggested creating more topic-focused discussion boards as "exit ramps" from the 24/7 chat room, allowing participants to select personally relevant information in a structured way. Participants also suggested a keyword search feature. This element would streamline the process of finding specific information without the need to scroll through recent posts. To increase accessibility for people with different reading abilities, participants suggested that the app be available on bigger devices such as tablets or computers.

Finally, the mobile app experienced various bugs. For research purposes, this app was launched on a testing platform that required participants to log in with their credentials every 3 months. This issue led to widespread frustration and confusion among participants who lost access unexpectedly. Additionally, there were bugs in the video feature, which made it difficult for participants to connect unless they used platforms outside of the mobile app (eg, Zoom, Facetime). Future improvements to fix these bugs would ensure a smoother and more reliable user experience.

> *I guess I went to log on the other day I wasn't sure when it ended, and I was quite sad when I didn't have access anymore, to go on and read the stuff I was used to reading each day so that was kind of, that was nice. Well, it wasn't nice that I couldn't get on but it was nice, yeah.* [Recipient 4-2]

Discussion

Principal Findings

This study explored recipients' and peer supporters' experiences with and perspectives on REACHOUT, a peer-led mental health support intervention for adults with T1D living in rural and remote regions of British Columbia. Our results identified four major themes: (1) Need for a sense of community and belonging, (2) Factors to enhance the recipient-peer supporter experience, (3) Key aspects of the peer supporter experience, and (4) Importance of personalizing the user experience while using the REACHOUT mobile app.

Comparison to Prior Work

Consistent with our findings, the need for community and belonging, especially for geographically marginalized individuals, has also been reported in the literature. For instance, a systematic review of 12 qualitative studies on health care access for rural patients with chronic diseases found that a sense of group connection in rural areas mitigates feelings of vulnerability [43]. Similarly, Joensen et al [44] noted that while a feeling of inclusion contributes to health promotion, it is often lacking in daily life for individuals with T1D. Thus, a mobile app such as REACHOUT is especially valuable in addressing these gaps in remote and underserved communities.

With REACHOUT, recipients had the agency to choose a peer supporter based on personally relevant factors. This choice-based model deserves consideration, as it may optimize the recipient-peer supporter match [12,35]. Our data also suggest that successful pairs often referred to one another as "friends," which supports the idea that effective emotional support is built upon friendship and trust [17,22]. To enhance participant satisfaction, future peer support studies should adopt this recipient-driven matching process as recipients are in the best position to understand their own unique support needs.

While the one-on-one and group support delivery mechanisms address different support needs, many recipients expressed greater value for the former. The advantages of personalized individual relationships address the limitations inherent in group settings. For example, in an intervention of peer support meetings for adults with T1D focusing on insulin pumps, dissatisfied participants reported a lack of relevance in the discussion topics, hindering their ability to speak about topics that mattered to them [45]. Incorporating modalities that allow recipients to seek both group-based and one-on-one peer support within the same intervention promotes greater support customization for each user. Subsequent mental health support models should prioritize flexible delivery options that balance individualized support with opportunities for group engagement.

As participation in group activities within the mobile app was optional, we observed varying levels of engagement. Passive participation, characterized by viewing (vs posting) 24/7 chat room exchanges, was the most common. Participants engaged in "lurking" behavior, which involved routinely checking the chat room, gleaning value in reading anecdotes and being exposed to new topics related to T1D. "Lurking" was also observed in an online community–based peer support forum for in-hospital patients. This study found that 7 of 30 participants opted not to post yet still experienced a positive impact on emotional well-being [46]. Additionally, Tang et al [47] found that adults with T1D who passively engaged with the digital support platform (ie, 'lurkers') reported greater reductions in stigma-related distress compared to active posters. These findings highlight the role of passive engagement in mental health interventions as a strategy for mitigating "social risks" [47,48]. An in-depth examination of the mental health benefits associated with passive participation on digital platforms is warranted.

While not anticipated by peer supporters, the flow of support with recipients was bidirectional. However, the content of the "give and take" exchange likely encompassed a range of topics not necessarily diabetes-specific. Nonetheless, this opportunity for mutual sharing was also cited in a systematic review of qualitative peer support studies for chronic diseases [49]. Recognizing this reciprocity as an unintentional intervention, peer support studies should routinely assess changes in outcomes for both recipients and peer supporters. Clearly, peer support fosters emotional well-being for both parties.

Ensuring ongoing support for peer supporters beyond the initial training phase is essential for peer supporter effectiveness and well-being. Our intervention addressed this need by offering peer wellness sessions, a space for peer supporters to share successes and challenges. Not surprisingly, emotional investment leading to exhaustion can harm the mental health of peer supporters [50]. Thus, having an environment to express frustrations in real time such as how to deal with nonresponsive recipients or navigate emotionally charged conversations could potentially prevent burnout or dissatisfaction. Therefore, implementing regular communication or check-ins could enhance peer supporters' experience and overall intervention effectiveness.

Limitations

First, this study only recruited matched peer supporters (vs unmatched peer supporters). Perspectives from unmatched peer supporters were not captured. Future studies should consider interviewing those peer supporters who did not participate in the one-on-one support component but had access to other support delivery features. Second, this sample was self-selected and possibly more engaged and enthusiastic than other participants. This may limit the representativeness of the original REACHOUT cohort. While we compared the characteristics of the consenting and nonconsenting sample, future studies should ensure representation across different levels of engagement. Third, the socioeconomic background for participants was relatively high. Because we did not overrecruit for individuals with lower levels of income or education, the diversity of experiences captured may be skewed. Finally, the study targeted the rural and remote communities of Interior British Columbia; therefore, the results may not be generalizable to other geographically marginalized populations in BC or Canada.

Conclusions

Peer support is increasingly recognized as a critical component for mental health interventions in T1D. While research has focused largely on recipients of support, our study also considered perspectives of individuals delivering support, providing a holistic view. More importantly, it is the recipient-peer supporter dynamic that most likely drives the success of the implementation of the REACHOUT program and impacts mental health outcomes. Only by understanding the experiences of both parties can we refine our interventions to provide the optimal mental health support model.

Acknowledgments

We would like to acknowledge Interior Health, Diabetes Education Centres, health care professionals, and research team who supported the recruitment of recipients and peer supporters. This study was supported by Breakthrough T1D International 2-SRA-2020-986-S-B. The authors declare that there was no use of generative artificial intelligence in the writing, editing, or development of this paper. The focus group recordings were transcribed using the Zoom transcription service and the authors reviewed all transcripts to ensure accuracy and deidentified personal information.

Funding

TST's program of research is funded by Michael Smith Health Research BC/Breakthrough T1D Canada Health Professional-Investigator Award #HPI-2021-2359.

Data Availability

The dataset generated and analyzed during this study is not publicly available to ensure participants' privacy. For questions about the dataset, contact the corresponding author.

Authors' Contributions

TST contributed to funding acquisition, study conceptualization, design, implementation, supervision, manuscript review, editing, and revision and is the study guarantor. DL contributed to the study implementation, data collection, analysis, interpretation, original manuscript preparation, editing, and revision. PJ contributed to the data analysis and interpretation. DS contributed to the data interpretation, manuscript review, and editing. FSC contributed to data interpretation, manuscript review, and editing. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Interview guide created by the principal investigator and research team to guide focus group discussions.

[DOCX File, 20 KB - diabetes_v11i1e72779_app1.docx]

Multimedia Appendix 2

Interviewed participants' baseline characteristics compared to nonrespondents from the larger pilot participant population. Mann-Whitney *U* tests were applied for continuous variables and the Fisher exact test for categorical variables.

[DOCX File, 18 KB - diabetes_v11i1e72779_app2.docx]

Checklist 1

COREQ: 32-item checklist.

[DOCX File, 27 KB - diabetes_v11i1e72779_app3.docx]

References

1. Bennett PN, St Clair Russell J, Atwal J, Brown L, Schiller B. Patient-to-patient peer mentor support in dialysis: improving the patient experience. Semin Dial 2018 Sep;31(5):455-461. [doi: 10.1111/sdi.12703] [Medline: 29722062]
2. Kiemen A, Czornik M, Weis J. How effective is peer-to-peer support in cancer patients and survivors? A systematic review. J Cancer Res Clin Oncol 2023 Sep;149(11):9461-9485. [doi: 10.1007/s00432-023-04753-8] [Medline: 37120782]
3. Longley RM, Harnedy LE, Ghanime PM, et al. Peer support interventions in patients with kidney failure: a systematic review. J Psychosom Res 2023 Aug;171:111379. [doi: 10.1016/j.jpsychores.2023.111379] [Medline: 37270909]
4. Thompson DM, Booth L, Moore D, Mathers J. Peer support for people with chronic conditions: a systematic review of reviews. BMC Health Serv Res 2022 Mar 31;22(1):427. [doi: 10.1186/s12913-022-07816-7] [Medline: 35361215]
5. Azmiardi A, Murti B, Febrinasari RP, Tamtomo DG. The effect of peer support in diabetes self-management education on glycemic control in patients with type 2 diabetes: a systematic review and meta-analysis. Epidemiol Health 2021;43:e2021090. [doi: 10.4178/epih.e2021090] [Medline: 34696569]
6. Chen C, Zhou Y, Xu JY, Song HY, Yin XW, Gu ZJ. Effect of peer support interventions in patients with type 2 diabetes: a systematic review. Patient Educ Couns 2024 May;122(108172). [doi: 10.1016/j.pec.2024.108172] [Medline: 38320444]
7. Krishnamoorthy Y, Sakthivel M, Sarveswaran G, Eliyas SK. Effectiveness of peer led intervention in improvement of clinical outcomes among diabetes mellitus and hypertension patients—a systematic review and meta-analysis. Prim Care Diabetes 2019 Apr;13(2):158-169. [doi: 10.1016/j.pcd.2018.11.007] [Medline: 30528217]

8. Lu S, Leduc N, Moullec G. Type 2 diabetes peer support interventions as a complement to primary care settings in high-income nations: a scoping review. Patient Educ Couns 2022 Nov;105(11):3267-3278. [doi: 10.1016/j.pec.2022.08.010] [Medline: 36038395]
9. Grant V, Litchfield I. Acceptability of community health worker and peer supported interventions for ethnic minorities with type 2 diabetes: a qualitative systematic review. Front Clin Diabetes Healthc 2024;5:1306199. [doi: 10.3389/fcdhc.2024.1306199] [Medline: 38836261]
10. Okoro FO, Veri S, Davis V. Culturally appropriate peer-led behavior support program for African Americans with type 2 diabetes. Front Public Health 2018;6:340. [doi: 10.3389/fpubh.2018.00340] [Medline: 30533408]
11. Shiyanbola OO, Maurer M, Schwerer L, et al. A culturally tailored diabetes self-management intervention incorporating race-congruent peer support to address beliefs, medication adherence and diabetes control in African Americans: a pilot feasibility study. Patient Prefer Adherence 2022;16:2893-2912. [doi: 10.2147/PPA.S384974] [Medline: 36317056]
12. Afshar R, Askari AS, Sidhu R, et al. Out of the mouths of peer leaders: perspectives on how to improve a telephone-based peer support intervention in type 2 diabetes. Diabet Med 2022 Sep;39(9):e14853. [doi: 10.1111/dme.14853] [Medline: 35437815]
13. Allen NA, Litchman ML, Greenwood DA, Ng A, Sanchez-Birkhead A. Hispanic diabetes peer facilitator training needs and connections to Hispanics with type 2 diabetes. Hisp Health Care Int 2023 Dec;21(4):221-227. [doi: 10.1177/15404153231175998] [Medline: 37221843]
14. Ayala GX, Canale K, Ibarra L, et al. Facilitators of peer leader retention within a type 2 diabetes intervention trial for US Latinos/Hispanics of Mexican origin. Health Educ Res 2023 May 22;38(3):204-219. [doi: 10.1093/her/cyad006] [Medline: 36848036]
15. Goldman ML, Ghorob A, Eyre SL, Bodenheimer T. How do peer coaches improve diabetes care for low-income patients?: a qualitative analysis. Diabetes Educ 2013;39(6):800-810. [doi: 10.1177/0145721713505779] [Medline: 24168838]
16. Holman D, Simmons D, Ockenden N, Graffy JP. Learning from the experience of peer support facilitators and study nurses in diabetes peer support: a qualitative study. Chronic Illn 2021 Sep;17(3):269-282. [doi: 10.1177/1742395319873378] [Medline: 31495199]
17. Kowitt SD, Urlaub D, Guzman-Corrales L, et al. Emotional support for diabetes management: an international cross-cultural study. Diabetes Educ 2015 Jun;41(3):291-300. [doi: 10.1177/0145721715574729] [Medline: 25722064]
18. Tarfa A, Nordin J, Mott M, Maurer M, Shiyanbola O. A qualitative exploration of the experiences of peer leaders in an intervention to improve diabetes medication adherence in African Americans. BMC Public Health 2023 Jan 20;23(1):144. [doi: 10.1186/s12889-023-15059-2] [Medline: 36670394]
19. Wen MJ, Salihu EY, Yang C, Maurer M, Shiyanbola OO. Peer ambassador perspectives in a culturally tailored self-management intervention for African Americans with type 2 diabetes: a qualitative study. Pharmacy (Basel) 2024 May 8;12(3):75. [doi: 10.3390/pharmacy12030075] [Medline: 38804467]
20. Dale J, Caramlau I, Sturt J, Friede T, Walker R. Telephone peer-delivered intervention for diabetes motivation and support: the telecare exploratory RCT. Patient Educ Couns 2009 Apr;75(1):91-98. [doi: 10.1016/j.pec.2008.09.014] [Medline: 19013741]
21. Kim C, MacMillan F, Osuagwu UL, Simmons D. Peer support facilitator and peer perspectives of an inner-regional Australian diabetes prevention and self-management peer support program. Health Soc Care Community 2022 Nov;30(6):e4051-e4064. [doi: 10.1111/hsc.13798] [Medline: 35318758]
22. Lott BD, Dicks TN, Keddem S, Ganetsky VS, Shea JA, Long JA. Insights into veterans' perspectives on a peer support program for glycemic management. Diabetes Educ 2019 Dec;45(6):607-615. [doi: 10.1177/0145721719879417] [Medline: 31596174]
23. Paul G, Keogh K, D'Eath M, Smith SM. Implementing a peer-support intervention for people with type 2 diabetes: a qualitative study. Fam Pract 2013 Oct;30(5):593-603. [doi: 10.1093/fampra/cmt027] [Medline: 23836093]
24. Ashrafi S, Taylor D, Tang TS. Moving beyond "don't ask, don't tell": mental health needs of adults with type 1 diabetes in rural and remote regions of British Columbia. Diabet Med 2021 May;38(5):e14534. [doi: 10.1111/dme.14534] [Medline: 33524209]
25. Bujanda-Sainz de Murieta A, Escalada-Hernández P, San Martín-Rodríguez L, Soto-Ruiz N. Determinants of the adoption of an online community for people with diabetes: a qualitative study. Sci Rep 2025 Oct 22;15(1):36944. [doi: 10.1038/s41598-025-20960-4] [Medline: 41125712]
26. Malik FS, Lind C, Duncan S, Mitrovich C, Pascual M, Yi-Frazier JP. Augmenting traditional support groups for adolescents with type 1 diabetes using Instagram: mixed methods feasibility study. JMIR Diabetes 2021 Oct 21;6(4):e21405. [doi: 10.2196/21405] [Medline: 34673527]
27. Litchman ML, Oser TK, Hodgson L, et al. In-person and technology-mediated peer support in diabetes care: a systematic review of reviews and gap analysis. Diabetes Educ 2020 Jun;46(3):230-241. [doi: 10.1177/0145721720913275] [Medline: 32321370]
28. Berkanish P, Pan S, Viola A, Rademaker Q, Devine KA. Technology-based peer support interventions for adolescents with chronic illness: a systematic review. J Clin Psychol Med Settings 2022 Dec;29(4):911-942. [doi: 10.1007/s10880-022-09853-0] [Medline: 35147830]

29. Titoria R, Fung A, Tang TS, Amed S. Systematic review of technology-mediated peer support interventions in paediatric type 1 diabetes care. Diabet Med 2023 Oct;40(10):e15172. [doi: 10.1111/dme.15172] [Medline: 37428650]
30. Berg M, Linden K, Adolfsson A, Sparud Lundin C, Ranerup A. Web-based intervention for women with type 1 diabetes in pregnancy and early motherhood: critical analysis of adherence to technological elements and study design. J Med Internet Res 2018 May 2;20(5):e160. [doi: 10.2196/jmir.9665] [Medline: 29720365]
31. Linden K, Berg M, Adolfsson A, Sparud-Lundin C. Person-centred, web-based support in pregnancy and early motherhood for women with type 1 diabetes mellitus: a randomized controlled trial. Diabet Med 2018 Feb;35(2):232-241. [doi: 10.1111/dme.13552] [Medline: 29171071]
32. Liu Z, Wang C, Yang D, et al. High engagement in mobile peer support is associated with better glycemic control in type 1 diabetes: a real-world study. J Diabetes Investig 2022 Nov;13(11):1914-1924. [doi: 10.1111/jdi.13870] [Medline: 35708894]
33. Xie LF, Housni A, Nakhla M, et al. Adaptation of an adult web application for type 1 diabetes self-management to youth using the behavior change wheel to tailor the needs of health care transition: qualitative interview study. JMIR Diabetes 2023 Apr 26;8:e42564. [doi: 10.2196/42564] [Medline: 37121571]
34. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
35. Tang TS, Klein G, Görges M, et al. Evaluating a mental health support mobile app for adults with type 1 diabetes living in rural and remote communities: the REACHOUT pilot study. Diabet Med 2025 Mar;42(3):e15451. [doi: 10.1111/dme.15451] [Medline: 39538423]
36. Fisher L, Polonsky WH, Hessler DM, et al. Understanding the sources of diabetes distress in adults with type 1 diabetes. J Diabetes Complications 2015;29(4):572-577. [doi: 10.1016/j.jdiacomp.2015.01.012] [Medline: 25765489]
37. Tang TS, Yip AKW, Klein G, et al. Training peers to deliver mental health support to adults with type 1 diabetes using the REACHOUT mobile app. Diabet Med 2024 Mar;41(3):e15210. [doi: 10.1111/dme.15210] [Medline: 37634222]
38. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
39. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208. [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
40. NVivo – leading qualitative data analysis software. Lumivero. URL: https://lumivero.com/products/nvivo/ [accessed 2026-01-07]
41. Ryan G. Introduction to positivism, interpretivism and critical theory. Nurse Res 2018 Mar 16;25(4):14-20. [doi: 10.7748/nr.2018.e1466] [Medline: 29546962]
42. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
43. Brundisini F, Giacomini M, DeJean D, Vanstone M, Winsor S, Smith A. Chronic disease patients' experiences with accessing health care in rural and remote areas: a systematic review and qualitative meta-synthesis. Ont Health Technol Assess Ser 2013;13(15):1-33. [Medline: 24228078]
44. Joensen LE, Filges T, Willaing I. Patient perspectives on peer support for adults with type 1 diabetes: a need for diabetes-specific social capital. Patient Prefer Adherence 2016;10:1443-1451. [doi: 10.2147/PPA.S111696] [Medline: 27536076]
45. Joensen LE, Meldgaard Andersen M, Jensen S, Nørgaard K, Willaing I. The effect of peer support in adults with insulin pump-treated type 1 diabetes: a pilot study of a flexible and participatory intervention. Patient Prefer Adherence 2017;11:1879-1890. [doi: 10.2147/PPA.S142204] [Medline: 29138541]
46. Haldar S, Mishra SR, Kim Y, Hartzler A, Pollack AH, Pratt W. Use and impact of an online community for hospital patients. J Am Med Inform Assoc 2020 Apr 1;27(4):549-557. [doi: 10.1093/jamia/ocz212] [Medline: 31986197]
47. Tang TS, Seddigh S, Halbe E, Vesco AT. Testing 3 digital health platforms to improve mental health outcomes in adults with type 1 diabetes: a pilot trial. Can J Diabetes 2024 Feb;48(1):18-25. [doi: 10.1016/j.jcjd.2023.08.006] [Medline: 37625504]
48. Mazuro C, Rao N. Online discussion forums in higher education: is 'lurking' working? Int J cross-disciplinary Subj Educ 2011;2(2):364-371. [doi: 10.20533/ijcdse.2042.6364.2011.0051]
49. Embuldeniya G, Veinot P, Bell E, et al. The experience and impact of chronic disease peer support interventions: a qualitative synthesis. Patient Educ Couns 2013 Jul;92(1):3-12. [doi: 10.1016/j.pec.2013.02.002] [Medline: 23453850]
50. Kane L, Portman RM, Eberhardt J, et al. Peer supporters' mental health and emotional wellbeing needs: key factors and opportunities for co-produced training. Health Expect 2023 Dec;26(6):2387-2395. [doi: 10.1111/hex.13836] [Medline: 37565576]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative Research
REDCap: Research Electronic Data Capture
T1D: type 1 diabetes

Edited by A Stone; submitted 17.Feb.2025; peer-reviewed by D Tilden, L Fried, M Pulier; revised version received 19.Dec.2025; accepted 22.Dec.2025; published 21.Jan.2026.

Please cite as:
Lam D, Jhajj P, Sherifali D, Chen FS, Tang TS
Exploring the REACHOUT Mental Health Support App for Type 1 Diabetes From the Perspectives of Recipients and Providers of Peer Support: Qualitative Study
JMIR Diabetes 2026;11:e72779
URL: https://diabetes.jmir.org/2026/1/e72779
doi:10.2196/72779

Artificial Intelligence in Diabetic Kidney Disease Research: Bibliometric Analysis From 2006 to 2024

Xingyuan Li, MS; Liming Xiao, BS; Fenghao Yang, MS; Fang Liu, MD, PhD

Department of Nephrology, West China Hospital, Sichuan University, 37 Guoxue Alley, Wuhou District, Chengdu, Sichuan, China

Corresponding Author:
Fang Liu, MD, PhD
Department of Nephrology, West China Hospital, Sichuan University, 37 Guoxue Alley, Wuhou District, Chengdu, Sichuan, China

Abstract

Background: Diabetic kidney disease (DKD) is a major complication of diabetes and the leading cause of end-stage renal disease globally. Artificial intelligence (AI) technologies have shown increasing potential in DKD research for early detection, risk prediction, and disease management. However, the landscape of AI applications in this field remains incompletely mapped, especially in terms of collaboration networks, thematic evolution, and clinical translation.

Objective: This study aims to perform a comprehensive bibliometric and translational analysis of AI-related DKD research published between 2006 and 2024, identifying publication trends, research hotspots, key contributors, collaboration patterns, and the extent of clinical validation and explainability.

Methods: A systematic search of the Web of Science Core Collection was conducted to identify English-language original articles applying AI technologies to DKD. Articles were screened following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines. Bibliometric visualization was performed using CiteSpace and VOSviewer to assess coauthorship, institutional and country collaboration, keyword evolution, and citation bursts. A qualitative review was conducted to evaluate clinical validation, model explainability, and real-world implementation.

Results: Out of 1158 retrieved records, 384 studies met the inclusion criteria. Global publications on AI in DKD increased rapidly after 2019. China led in publication volume, followed by the United States, India, and Iran. Keyword analysis showed a thematic transition from early biomarker and proteomic research to deep learning, clinical prediction models, and management tools. Despite methodological advances, few studies included external validation or explainability frameworks. Notable translational efforts included DeepMind's acute kidney injury predictor and a chronic kidney disease prediction model developed by Sumit, yet widespread real-world integration remains limited.

Conclusions: AI research in DKD has grown substantially over the past 2 decades, with expanding international collaboration and diversification of research themes. However, challenges persist in clinical applicability, model transparency, and global inclusivity. Future research should prioritize explainable AI, multicenter validation, and integration into clinical workflows to support effective translation of AI innovations into DKD care.

(JMIR Diabetes 2026;11:e72616) doi:10.2196/72616

KEYWORDS

artificial intelligence; diabetic kidney disease; bibliometric analysis; clinical validation; explainable AI; global collaboration

Introduction

Diabetic kidney disease (DKD) is the most prevalent microvascular complication of diabetes mellitus and a leading cause of end-stage renal disease globally, accounting for a substantial proportion of dialysis and transplantation burdens worldwide [1]. The pathophysiological progression of DKD is complex, often involving chronic hyperglycemia-induced glomerular injury, hemodynamic changes, inflammation, and fibrosis. Early-stage DKD is typically asymptomatic, and by the time clinical markers such as proteinuria or a decline in glomerular filtration rate become apparent, irreversible kidney damage may have already occurred [2]. Therefore, early detection and individualized risk stratification are essential for improving patient outcomes and alleviating long-term health care burdens.

In this context, artificial intelligence (AI) has emerged as a transformative approach in biomedical research and clinical practice. With capabilities in data-driven pattern recognition, predictive modeling, and real-time decision support, AI techniques—including machine learning, deep learning, and neural networks—have been increasingly explored to address key challenges in DKD research and management [3,4]. Applications range from biomarker discovery and disease classification to risk modeling and personalized treatment optimization. Despite the growing enthusiasm for AI, there is wide variability in the methodological rigor, clinical applicability, and translational maturity of these studies.

While several narrative and systematic reviews have highlighted specific AI models used in nephrology, there remains a lack of comprehensive evaluation of how the field has evolved thematically over time, which countries and institutions are leading its development, how collaborative efforts are shaping knowledge production, and to what extent the proposed AI solutions are being validated and implemented in real-world clinical settings. Moreover, important dimensions such as model explainability, equity in global research representation, and translational readiness are often underexamined.

This study aims to address these gaps by conducting a bibliometric and translational landscape analysis of AI-related DKD research published from 2006 to 2024. By integrating quantitative bibliometric mapping with qualitative evaluation of translational attributes—including clinical validation, model transparency, and implementation potential—we aim to provide a comprehensive overview of this rapidly evolving field and offer insights to inform future research, clinical integration, and policy development.

Methods

Literature Search and Eligibility Criteria

A systematic literature search was conducted using the Web of Science Core Collection to identify studies related to the application of AI in DKD from January 1, 2006, to April 30, 2024. The search strategy included combinations of terms for DKD ("diabetic kidney disease," "diabetic nephropathy," "DKD," or "DN") and AI ("artificial intelligence," "machine learning," "deep learning," or "neural network"). Only English-language articles were considered. The search was limited to original research articles involving human-related data, excluding reviews, editorials, letters, conference abstracts, and purely experimental or theoretical reports without clinical relevance.

Eligible articles were those that applied AI techniques to DKD in a clinical, translational, or predictive context. Studies that involved image processing, signal detection, or statistical models unrelated to DKD-specific diagnostic or prognostic tasks were excluded. To ensure the reliability of inclusion, 2 reviewers (XL and FY) independently screened titles and abstracts for relevance, followed by full-text assessment. Discrepancies were resolved by consensus or consultation with a third reviewer (LX).

Bibliometric Mapping and Analysis Tools

Bibliometric data were exported from the Web of Science platform (Multimedia Appendix 1) and analyzed using CiteSpace (v6.1.R6) and VOSviewer (v1.6.18; Leiden University's Centre for Science and Technology Studies; Multimedia Appendix 2). These tools enabled visualization and quantification of publication trends, author and institutional productivity, international collaboration networks, and thematic keyword clusters. CiteSpace was used to generate timeline visualizations and detect emergent research topics through keyword burst detection. VOSviewer was applied to construct network maps illustrating coauthorship patterns and co-occurrence frequencies. Centrality scores and citation frequencies were used to identify influential authors, institutions, and countries within the research landscape.

Translational and Thematic Evaluation

In addition to bibliometric analysis, a qualitative assessment was performed to evaluate the translational significance of the included studies. This review focused on identifying whether AI models were externally validated or tested across different cohorts, whether explainable AI methods were incorporated, and whether any studies reported or discussed clinical integration or real-world implementation. Studies that mentioned the use of interpretability frameworks such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) were noted. The presence of multicenter datasets, ethnically diverse populations, or cross-national data integration was also considered as indicators of generalizability and applicability. This dual approach—combining quantitative mapping with thematic content analysis—allowed for a multidimensional perspective on both the scientific growth and translational depth of AI research in DKD.

Ethical Considerations

This study involved no human participants, animals, or patient data, and therefore did not require ethical approval. The data used were retrieved from publicly available bibliographic databases and do not involve any sensitive or identifiable personal information.

Results

Study Selection

A total of 1158 records were initially identified following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 framework. After the removal of 0 duplicates, 1158 records were screened based on their titles and abstracts. Of these, 251 records were excluded as irrelevant. The remaining 907 full-text articles were assessed for eligibility, resulting in 384 articles included in the quantitative synthesis, and an additional 78 articles included in the qualitative thematic review. Ultimately, these articles were included in the subsequent bibliometric and qualitative synthesis.

Publication Growth Over Time

The global volume of publications related to AI in DKD remained low and relatively stagnant between 2006 and 2016. A notable increase in research output began in 2019, followed by a rapid rise during the years 2022 to 2024 (Figure 1). This pattern reflects the growing integration of AI techniques into biomedical research and the rising urgency of addressing DKD in the context of the global diabetes epidemic. The sharp upward trend in recent years suggests an increasing recognition of AI as a valuable tool for advancing DKD risk prediction, diagnosis, and management (Figure 2A).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for literature screening. AI: artificial intelligence; DKD: diabetic kidney disease.

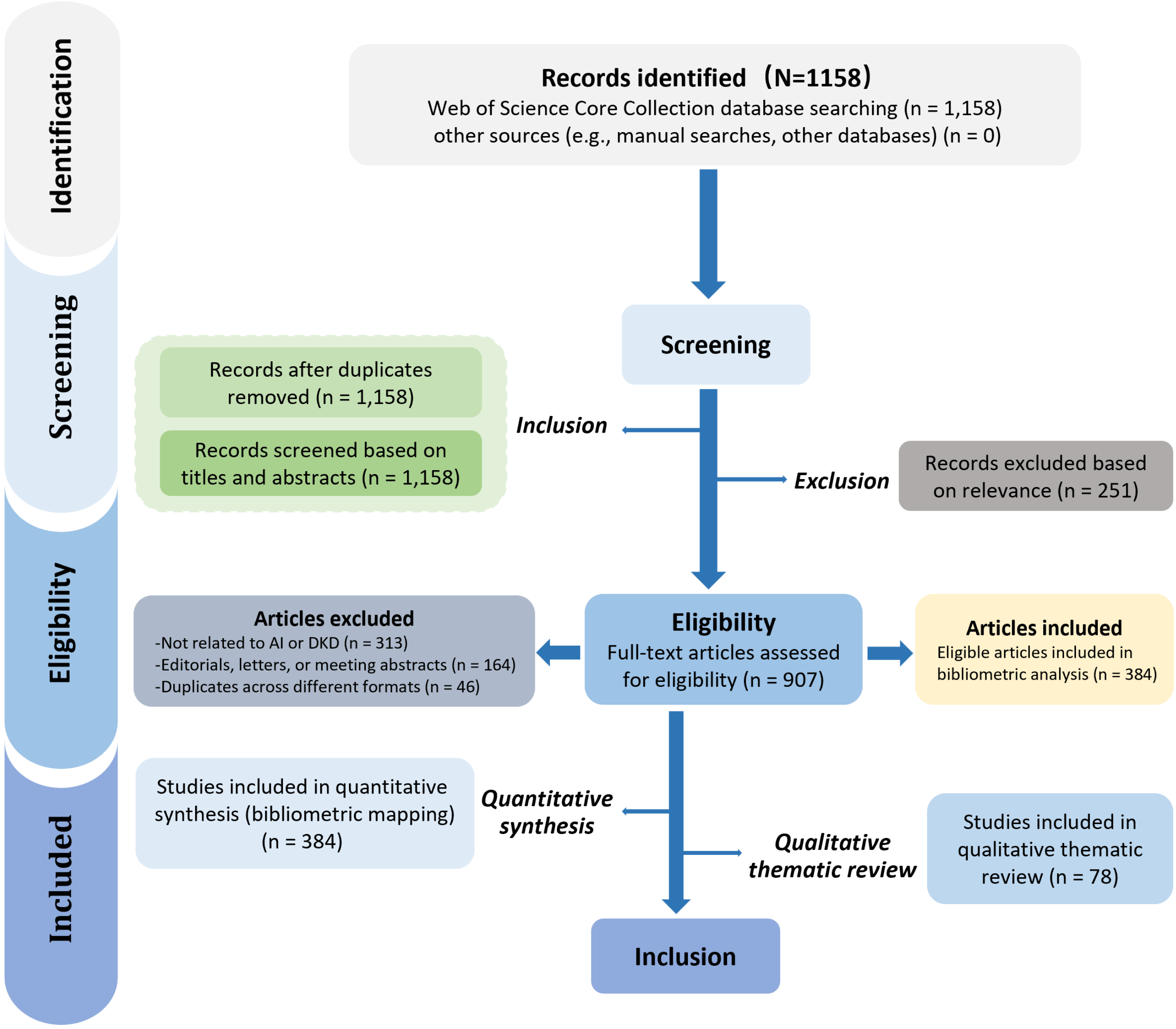
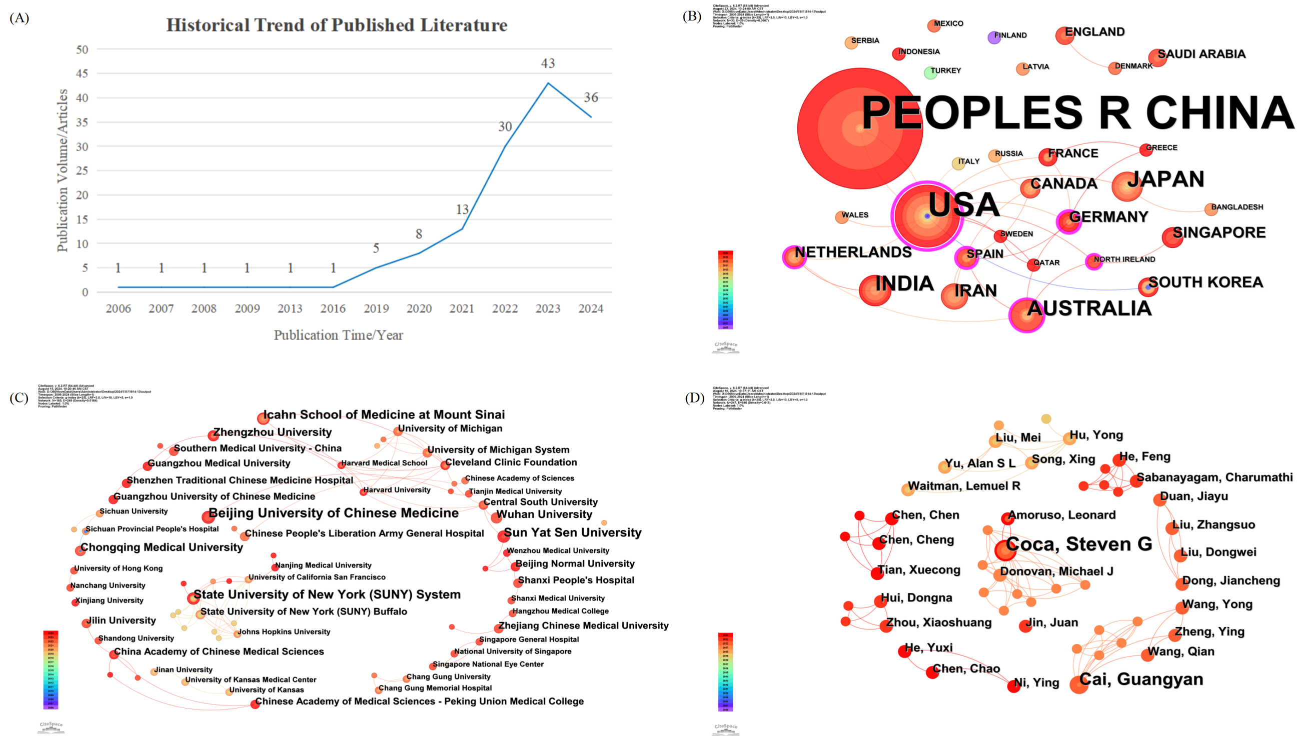


Figure 2. Analysis of the publication trends in artificial intelligence research on diabetic kidney disease from 2006 to 2024: (A) timeline of annual publications, (B) co-occurrence network of research countries, (C) co-occurrence network of research institutions, and (D) co-occurrence network of authors.



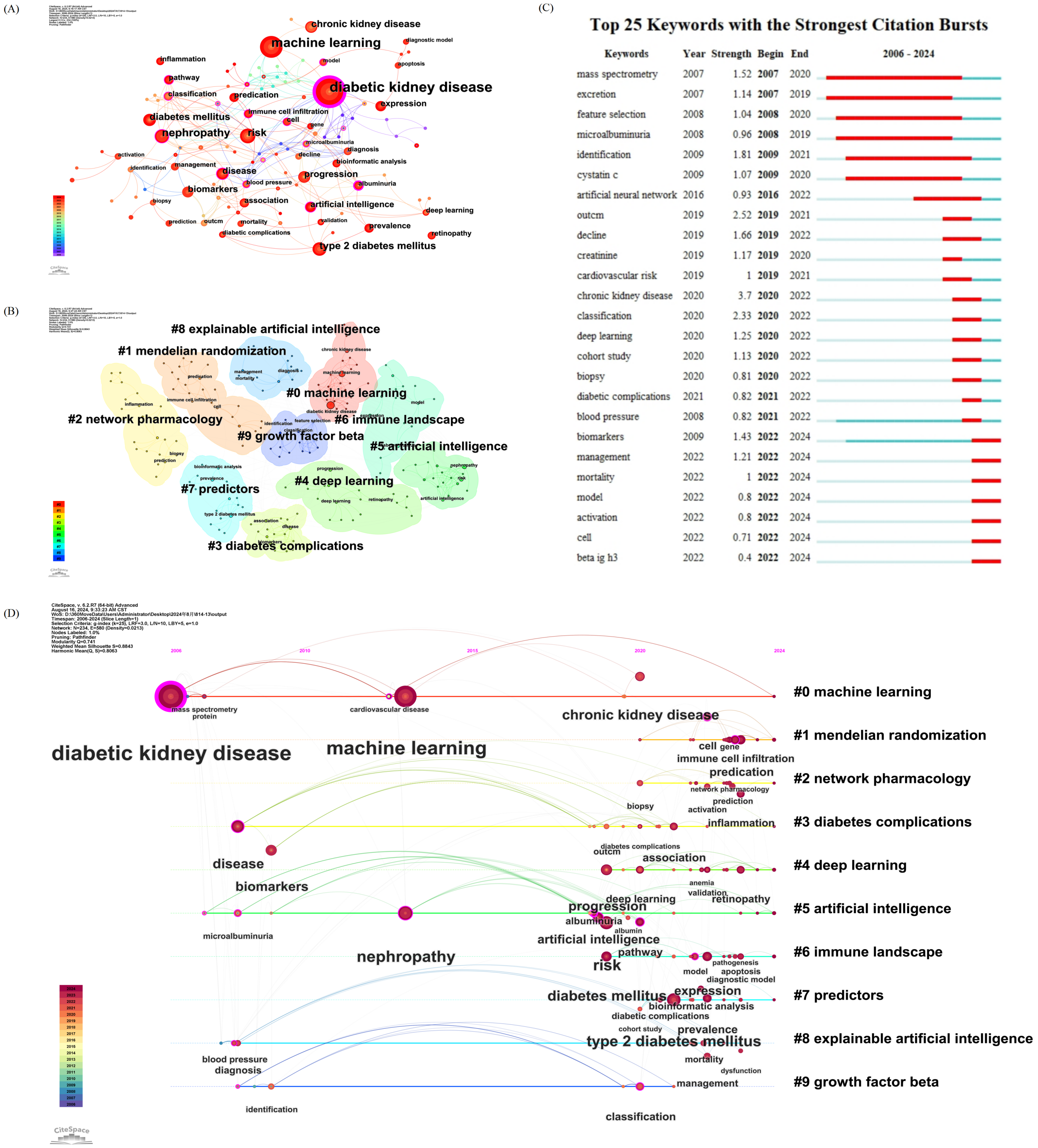
Geographic and Institutional Contributions

China emerged as the leading contributor in terms of publication volume, accounting for nearly half of all included studies. Key institutions such as Beijing University of Chinese Medicine, Sun Yat-sen University, and Central South University were among the most prolific. The United States ranked second, with prominent contributions from institutions such as the Icahn School of Medicine at Mount Sinai. India, Iran, and Australia also made notable contributions, reflecting a broader international interest in the intersection of AI and nephrology. Collaboration patterns showed that high-output countries often published independently, although intercontinental partnerships—particularly between East Asia, North America, and parts of Europe—have been increasing in frequency and visibility (Figure 2B-D).

Keyword Evolution and Research Hotspots

Analysis of keyword co-occurrence and burst terms revealed distinct phases in the thematic development of the field. During the early period (2006 - 2012), research was focused primarily on pathology, biomarker identification, and proteomic analysis, often using conventional statistical tools. Between 2013 and 2018, machine learning began to emerge as a prominent analytical method, with keywords such as “support vector machine” and “feature selection” gaining prominence. From 2019 onward, deep learning became a dominant theme, as reflected by the increasing frequency of terms such as “convolutional neural network,” “risk prediction,” and “decision support system.” Thematic clustering and citation bursts also indicated a growing interest in explainability, model integration, and individualized risk stratification, marking a shift toward clinical application and interpretability (Figure 3A-D).

Figure 3. Co-occurrence analysis of keywords in bibliometric studies: (A) keyword co-occurrence network, (B) keyword clustering, (C) keywords with the strongest citation bursts, and (D) timeline of keyword trends in artificial intelligence research on diabetic kidney disease from 2006 to 2024.



Collaboration Networks Among Authors and Institutions

Coauthorship network visualization demonstrated that the field remains highly fragmented, with a large number of small, loosely connected research groups. The most central nodes in the institutional network were located in China, the United States, and Singapore, reflecting both productivity and cross-institutional engagement. Although multicenter projects were occasionally identified, most AI models were developed and tested within single-center or regional datasets. Cross-national research, while increasing, often lacked shared validation protocols or harmonized data structures, limiting direct comparisons and large-scale model generalizability.

Model Validation, Explainability, and Translational Readiness

A review of the included studies showed that only a limited proportion of AI models underwent external validation using independent cohorts. Most models were based on retrospective data from a single institution or health system, with internal cross-validation as the primary method of evaluation. Very few studies implemented explainability frameworks such as SHAP or LIME, and even fewer offered insights into how model outputs could be integrated into clinical decision-making processes. Notable exceptions included studies that incorporated prospective testing or demonstrated integration with electronic health records, although these remained rare. DeepMind's acute kidney injury prediction system, while not DKD-specific, was often cited as a prototype for nephrology-focused AI applications [5]. Similarly, Sumit's [6] deep learning–based model for chronic kidney disease risk prediction represented an example of real-world implementation relevant to diabetic populations. However, the lack of consistent attention to explainability, real-time integration, and regulatory considerations suggests that most AI-DKD research remains in a pretranslational stage.

Discussion

Principal Findings

This bibliometric and thematic analysis presents a comprehensive overview of research trends, international collaborations, and translational depth in the application of AI to DKD from 2006 to 2024. The temporal trend reveals a slow developmental phase lasting more than a decade, followed by a surge in research activity from 2019 onward. This acceleration corresponds with the broader adoption of AI in medicine and the urgent need for precision tools to combat the rising global burden of diabetes-related complications.

China and the United States have emerged as the primary contributors to this field, with China leading in publication quantity and institutional productivity. However, the dominance of single-country studies and weak international collaboration networks suggests a lack of unified global efforts in AI-DKD research. While some cross-border cooperation exists, it has not yet reached the level necessary to support large-scale model generalization or multiethnic validation. Future research should prioritize open data sharing, transnational model calibration, and harmonized validation protocols to promote reproducibility and clinical readiness across diverse populations.

Keyword analysis and thematic clustering indicate a clear evolution in research focus. Early studies emphasized molecular and pathological mechanisms of DKD, typically using traditional regression models or biomarker discovery tools. From 2015 onward, a shift occurred toward applying machine learning algorithms to structured clinical data, including risk prediction and feature selection. Since 2019, the field has seen a rapid proliferation of deep learning–based applications, especially convolutional neural networks for imaging and time-series data analysis. However, the transition from computational innovation to clinical implementation remains incomplete. Most studies prioritize model development and internal validation, while relatively few undertake real-world testing or prospective evaluation.

One major limitation identified is the scarcity of externally validated and clinically integrated AI models. Despite rapid algorithmic progress, few studies reach the level of clinical translation demonstrated by landmark systems such as DeepMind's acute kidney injury prediction algorithm, which was prospectively validated and tested in hospital settings [5]. Similarly, the work by Sumit [6], which developed and validated a deep learning model for predicting chronic kidney disease progression, represents an exemplar of real-world application. These examples underscore the importance of incorporating prospective design, external datasets, and health system integration early in the research pipeline to ensure that AI tools can transition beyond proof-of-concept stages.

Moreover, the "black box" nature of many AI models presents a significant barrier to clinical trust and regulatory approval. Although explainable artificial intelligence methods such as SHAP and LIME have been proposed and applied in other medical domains, they are seldom used in DKD-related research. This gap not only limits interpretability but also hinders integration into clinical workflows where explainability is essential for physician adoption and patient safety. The increasing interest in interpretable models and hybrid systems—combining clinical rules with machine learning outputs—may offer a promising path forward.

Another noteworthy observation is the underrepresentation of research from low- and middle-income countries, apart from China and India. Given the global prevalence of diabetes and its complications, this imbalance may reflect disparities in AI infrastructure, research funding, and access to large-scale clinical data. Efforts to democratize AI research—such as open-access datasets, international consortia, and capacity-building initiatives—are critical to avoid reinforcing health inequities through algorithmic bias.

Limitations and Future Work

This study also has limitations. The analysis was based solely on the Web of Science database, which, while comprehensive, may omit relevant studies indexed elsewhere, such as in Scopus or PubMed. The decision to focus on English-language articles may have further excluded important regional research. Additionally, bibliometric tools such as CiteSpace and VOSviewer, while effective in mapping research landscapes, cannot capture the full context or nuance of each study's methodological rigor or clinical relevance. Therefore, the qualitative thematic analysis presented here serves as a complementary lens, but further domain-specific review is warranted to assess clinical impact.

In conclusion, the field of AI in DKD is rapidly expanding, with increasing interest from diverse geographic regions and institutions. However, the translation of AI models into clinical nephrology practice remains limited. Future research should emphasize multicenter collaboration, external validation, and interpretability to close the gap between computational innovation and real-world impact. A systematic shift toward transparent, validated, and context-aware AI systems will be essential to unlock the full potential of AI in the management of DKD.

Conclusions

This study provides a comprehensive and multidimensional analysis of the research landscape at the intersection of AI and DKD. Through bibliometric visualization and thematic synthesis, we demonstrate that although the field has experienced substantial growth in recent years—particularly with the application of deep learning technologies—the clinical translation of these innovations remains in its infancy. Most current research is confined to retrospective model development with limited external validation and minimal integration into real-world nephrology practice.

To advance the field, future efforts must prioritize methodological transparency, external validation using diverse populations, and the incorporation of explainable AI frameworks. Strengthening international collaboration and establishing multicenter consortia will be crucial for ensuring reproducibility and promoting equitable access to AI tools across health care settings. Additionally, regulatory and ethical considerations should be proactively addressed to support the safe deployment of AI in clinical decision-making.

In summary, while the promise of AI in DKD is evident, realizing its full potential will require a deliberate transition from algorithmic development to clinically meaningful, patient-centered applications. Bridging this translational gap is not only a technical challenge but also an opportunity to reshape chronic disease management in the era of intelligent medicine.

Acknowledgments

The authors thank all members of the biomedical informatics and nephrology research teams for their valuable discussions during the conceptualization of this study. We also acknowledge the developers of CiteSpace and VOSviewer for providing open-access tools that facilitated the bibliometric analysis. This study was supported by the Health Commission of Sichuan Province Program (21ZD001).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Raw bibliometric data exported from the Web of Science Core Collection (CSV format, retrieved May 2024).
[PDF File, 19886 KB - diabetes_v11i1e72616_app1.pdf]

Multimedia Appendix 2
Analysis scripts and configuration settings used in CiteSpace (version 6.2.R6) and VOSviewer (version 1.6.19), provided in TXT and VOS formats.
[ZIP File, 22 KB - diabetes_v11i1e72616_app2.zip]

References

1. Umanath K, Lewis JB. Update on diabetic nephropathy: core curriculum 2018. Am J Kidney Dis 2018 Jun;71(6):884-895. [doi: 10.1053/j.ajkd.2017.10.026] [Medline: 29398179]
2. Alicic RZ, Rooney MT, Tuttle KR. Diabetic kidney disease: challenges, progress, and possibilities. Clin J Am Soc Nephrol 2017 Dec 7;12(12):2032-2045. [doi: 10.2215/CJN.11491116] [Medline: 28522654]
3. Rodriguez-Romero V, Bergstrom RF, Decker BS, Lahu G, Vakilynejad M, Bies RR. Prediction of nephropathy in type 2 diabetes: an analysis of the ACCORD trial applying machine learning techniques. Clin Transl Sci 2019 Sep;12(5):519-528. [doi: 10.1111/cts.12647] [Medline: 31112000]
4. Makino M, Yoshimoto R, Ono M, et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. Sci Rep 2019 Aug 14;9(1):11862. [doi: 10.1038/s41598-019-48263-5] [Medline: 31413285]
5. Ravizza S, Huschto T, Adamov A, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nat Med 2019 Jan;25(1):57-59. [doi: 10.1038/s41591-018-0239-8] [Medline: 30617317]
6. Sumit K. Chronic kidney disease prediction using deep learning. Int J Sci Res Eng Manag 2025 May;9(5):1-4. [doi: 10.55041/IJSREM48077]

Abbreviations

AI: artificial intelligence
DKD: diabetic kidney disease
LIME: Local Interpretable Model-Agnostic Explanations
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SHAP: Shapley Additive Explanations

Edited by L Quinlan; submitted 13.Feb.2025; peer-reviewed by R Wen, S Banerjee, SR Polamuri; revised version received 22.Jul.2025; accepted 23.Jul.2025; published 09.Jan.2026.

Please cite as:
Li X, Xiao L, Yang F, Liu F
Artificial Intelligence in Diabetic Kidney Disease Research: Bibliometric Analysis From 2006 to 2024
JMIR Diabetes 2026;11:e72616
URL: https://diabetes.jmir.org/2026/1/e72616
doi:10.2196/72616

Privacy-Preserving Collaborative Diabetes Prediction in Heterogeneous Health Care Systems: Algorithm Development and Validation of a Secure Federated Ensemble Framework

Md Rakibul Hasan, MSc; Juan Li, PhD

Department of Computer Science, Faculty, North Dakota State University, 258 Quentin Burdick Bldg, Computer Science Department, NDSU, 1320 Albrecht Blvd, Fargo, ND, United States

Corresponding Author:
Juan Li, PhD
Department of Computer Science, Faculty, North Dakota State University, 258 Quentin Burdick Bldg, Computer Science Department, NDSU, 1320 Albrecht Blvd, Fargo, ND, United States

Abstract

Background: Diabetes prediction requires accurate, privacy-preserving, and scalable solutions. Traditional machine learning models rely on centralized data, posing risks to data privacy and regulatory compliance. Moreover, health care settings are highly heterogeneous, with diverse participants, hospitals, clinics, and wearables, producing nonindependent and identically distributed data and operating under varied computational constraints. Learning in isolation at individual institutions limits model generalizability and effectiveness. Collaborative federated learning (FL) enables institutions to jointly train models without sharing raw data, but current approaches often struggle with heterogeneity, security threats, and system coordination.

Objective: This study aims to develop a secure, scalable, and privacy-preserving framework for diabetes prediction by integrating FL with ensemble modeling, blockchain-based access control, and knowledge distillation. The framework is designed to handle data heterogeneity, nonindependent and identically distributed distributions, and varying computational capacities across diverse health care participants while simultaneously enhancing data privacy, security, and trust.

Methods: We propose a federated ensemble learning framework, FedEnTrust, that enables decentralized health care participants to collaboratively train models without sharing raw data. Each participant shares soft label outputs, which are distilled and aggregated through adaptive weighted voting to form a global consensus. The framework supports heterogeneous participants by assigning model architectures based on local computational capacity. To ensure secure and transparent coordination, a blockchain-enabled smart contract governs participant registration, role assignment, and model submission with strict role-based access control. We evaluated the system on the PIMA Indians Diabetes Dataset, measuring prediction accuracy, communication efficiency, and blockchain performance.

Results: The FedEnTrust framework achieved 84.2% accuracy, with precision, recall, and F_1-score of 84.6%, 88.6%, and 86.4%, respectively, outperforming existing decentralized models and nearing centralized deep learning benchmarks. The blockchain-based smart contract ensured 100% success for authorized transactions and rejected all unauthorized attempts, including malicious submissions. The average blockchain latency was 210 milliseconds, with a gas cost of ~107,940 units, enabling secure, real-time interaction. Throughout, patient privacy was preserved by exchanging only model metadata, not raw data.

Conclusions: FedEnTrust offers a deployable, privacy-preserving solution for decentralized health care prediction by integrating FL, ensemble modeling, blockchain-based access control, and knowledge distillation. It balances accuracy, scalability, and ethical data use while enhancing security and trust. This work demonstrates that secure federated ensemble systems can serve as practical alternatives to centralized artificial intelligence models in real-world health care applications.

(JMIR Diabetes 2026;11:e79166) doi:10.2196/79166

KEYWORDS

blockchain; decentralized health care; diabetes prediction; ensemble learning; federated learning; knowledge distillation; privacy-preserving AI; artificial intelligence; AI

Introduction

Diabetes continues to pose a growing global health burden, requiring timely prediction and proactive management to reduce complications and improve quality of life [1]. While machine learning has emerged as a powerful tool for diabetes prediction, conventional approaches often rely on centralized data repositories [2-4]. This reliance introduces serious challenges related to patient privacy, regulatory compliance (eg, Health Insurance Portability and Accountability Act (HIPAA), General

Data Protection Regulation (GDPR), and susceptibility to cyberattacks [5]. Moreover, centralized data aggregation is increasingly impractical due to fragmented data ownership across institutions and regions [6].

Real-world health care systems are inherently heterogeneous, encompassing a wide range of contributors—from large hospitals and urban clinics to wearable health devices in remote settings [7]. These entities vary significantly in data volume, quality, and computational capacity. The data are often nonindependent and identically distributed (non-IID), reflecting demographic, clinical, and behavioral diversity [8]. As a result, models trained within a single institution or on homogeneous datasets often struggle to generalize across settings, limiting their effectiveness and scalability.

To address these limitations, collaborative federated learning (FL) has emerged as a compelling solution [9]. However, applying FL to real-world diabetes prediction presents several unresolved challenges. In particular, current FL frameworks often struggle with:

- security vulnerabilities, such as model poisoning and adversarial manipulation [10]
- lack of coordination and trust, especially in decentralized, multiparty settings [11]
- performance degradation due to client heterogeneity and non-IID data distributions [12]

While several FL frameworks [13-16] have been explored for decentralized health care analytics, most assume homogeneous model architectures, single global models, or idealized trust environments and do not explicitly address lightweight or resource-constrained participants at the edge [17,18]. Existing systems, such as Biscotti [19] and Chang et al [20], rely on gradient sharing and therefore require structurally aligned models and consistent computational resources, while recent blockchain-enabled FL frameworks incorporate differential privacy but still assume homogeneous models or centralized coordination [21,22]. Furthermore, blockchain [23], a promising technology for ensuring integrity, transparency, and access control in decentralized systems, has seen limited integration with FL, especially in diabetes prediction contexts. Other blockchain-enabled approaches, such as Shalan et al [24], provide secure access control but do not incorporate mechanisms for interoperable knowledge sharing across heterogeneous local models.

In contrast, FedEnTrust introduces an integrated design that simultaneously addresses model heterogeneity, non-IID data, trust and identity verification, and secure update submission. By combining soft-label knowledge distillation with blockchain-verified RBAC, FedEnTrust enables robust collaboration across diverse health care systems while preventing unauthorized or malicious updates. FedEnTrust introduces a novel integration of:

- ensemble learning, allowing clients to train diverse local models best suited to their data and computational constraints
- soft-label knowledge distillation, enabling effective model aggregation across non-IID participants
- blockchain-based smart contracts, which provide tamper-proof coordination, role-based access control, and participant accountability

FedEnTrust represents a step forward in secure and collaborative artificial intelligence (AI) for health care, with the following key contributions:

1. Heterogeneity-aware ensemble design: Each participant trains a model tailored to its resource level, supporting real-world deployment across varied health care nodes.
2. Knowledge distillation-based aggregation: We introduce a soft-label ensemble mechanism that improves convergence and generalization across non-IID data.
3. Blockchain-enabled trust layer: Our smart contract system enforces participant registration, access control, and secure model submissions without a centralized authority.
4. Comprehensive evaluation: Using the PIMA Indians Diabetes Dataset, we demonstrate that FedEnTrust improves prediction accuracy; maintains privacy; and ensures secure, low-latency collaboration.

By addressing the intersection of privacy, trust, heterogeneity, and security, FedEnTrust provides a practical and deployable framework for AI-powered diabetes prediction in real-world, decentralized health care systems.

Methods

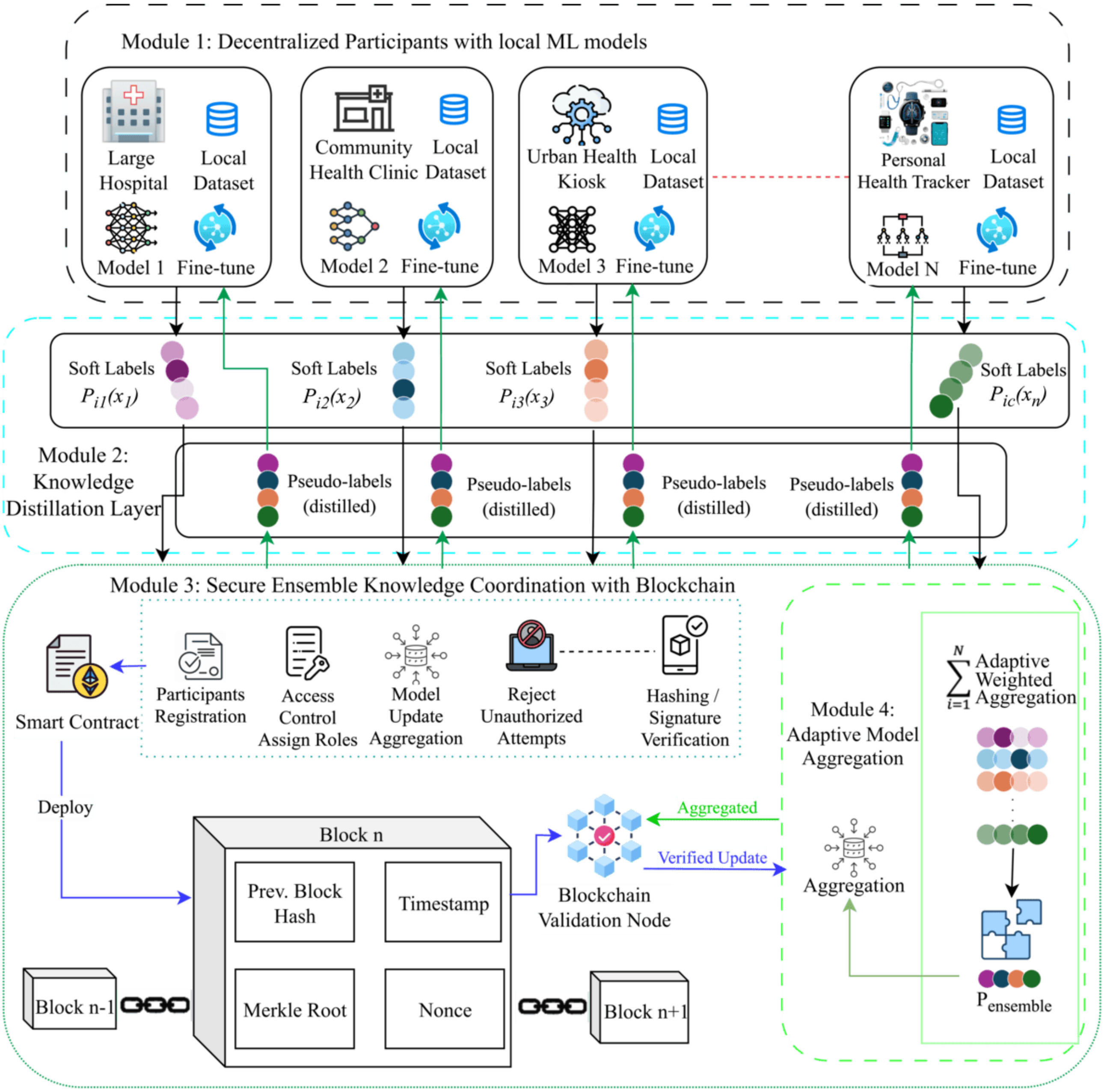
Overview of FedEnTrust

FedEnTrust is a secure, privacy-preserving federated ensemble learning framework designed to address the challenges of decentralized diabetes prediction across heterogeneous health care environments. It enables collaborative learning without centralizing sensitive patient data, accommodates diverse computational resources, and defends against malicious behaviors through a blockchain-coordinated trust infrastructure. The core modules of FedEnTrust include (1) heterogeneity-aware local model training, (2) knowledge distillation via soft label sharing, (3) blockchain-based secure coordination, and (4) adaptive ensemble aggregation.

These modules work together to realize 3 key objectives: maintaining patient privacy, enabling equitable participation across institutions with varying capacities, and ensuring secure collaboration in a decentralized system.

Figure 1 illustrates the end-to-end data flow across the 4 modules. Local raw data remain strictly on the device. Each participant trains a heterogeneous local model and generates soft-label probability vectors. These soft labels, along with accuracy metadata, are sent off-chain to the aggregator but must first pass through blockchain-based role-based access control (RBAC) validation, where the smart contract verifies participant identity, role permissions, and submission metadata. Validated soft labels are incorporated into an adaptive weighted aggregation mechanism, producing global pseudo-labels that are redistributed to all participants. The blockchain records transaction hashes and role enforcement events, ensuring traceability without revealing sensitive data.

Figure 1. Overview of the FedEnTrust architecture. Soft labels generated by local models are authenticated through blockchain-based role-based access control and combined using adaptive weighted aggregation to produce pseudo labels for continued local training. ML: machine learning.



Architectural Novelty and Comparison With Existing FL Frameworks

Real-world health care environments exhibit substantial diversity in computational capacity, data distributions, trust requirements, and security risks. To contextualize the design of FedEnTrust within this landscape, we compare its architectural capabilities against representative FL and blockchain-enabled frameworks in Table 1.

Table . Architectural comparison of FedEnTrust with representative federated learning frameworks.

Challenge in real-world health care FL	FedEnTrust (Our work)	Hasan et al [15]	Biscotti [19]	Chang et al [20]	Microcontroller FL[a] [17]
Heterogeneous compute environments (hospitals, clinics, kiosks, wearables)	Heterogeneity-aware model assignment; each node trains model matching its device capacity; ensemble aggregation aligns knowledge across disparate models	Supports ML[b] models but generally assumes similar capacity clients	Assumes all clients run comparable gradient-sharing deep models	Single model structure required; difficult for low-resource clients	Designed for ultra-low-power devices; not suitable for multitier health care
Non-IID[c] and imbalanced data across institutions	Soft-label knowledge distillation + weighted aggregation improve cross-site generalization	Local models trained independently; static averaging struggles with non-IID distributions	Gradient aggregation without distillation; non-IID data reduces convergence	DP[d]-sanitized gradients reduce signal strength on non-IID data	Very limited support for complex non-IID medical data
Cross-institution trust and secure participation	Smart contract–driven RBAC[e]; on-chain validation of model submissions; rejects malicious or unauthorized updates	Minimal security; no on-chain validation	Uses blockchain only as consensus layer, not for role-level access control	Smart contract manages DP gradients, not participation permissions	No trust or participation assurance mechanism
Protection against malicious updates (poisoning, fake uploads)	On-chain validator roles + metadata checks prevent poisoned soft labels before aggregation	No defense against malicious gradient or model uploads	Consensus prevents tampering but not model poisoning	DP reduces leakage but not poisoning	No adversarial defense features
Interoperability across model types	Soft labels unify outputs of RF[f], XGB[g], DT[h], SVM[i], KNN[j] into comparable probability space	Homogeneous ML models; limited interoperability	Requires same model structure for gradient fusion	Single-model FL; weights must match	No model interoperability
Scalability across distributed health care networks	Lightweight soft-label sharing reduces communication overhead and suits mixed-resource environments	Local model averaging; moderate scalability	Heavy blockchain consensus overhead limits scalability	DP gradient exchange increases bandwidth needs	Limited to microcontroller networks
Auditability and traceability for compliance (HIPAA[k] or GDPR[l])	Full on-chain audit log of registrations, updates, and permissions	Centralized coordination; limited auditability	All gradient updates stored on-chain—high cost	Stores only gradient summaries; limited audit transparency	Not designed for regulated health care settings

[a]FL: federated learning.
[b]ML: machine learning.
[c]IID: independent and identically distributed.
[d]DP: differential privacy.
[e]RBAC: role-based access control.
[f]RF: random forest.
[g]XGB: extreme gradient boosting.
[h]DT: decision tree.
[i]SVM: support vector machine.
[j]KNN: k-nearest neighbors.
[k]HIPAA: Health Insurance Portability and Accountability Act.
[l]GDPR: General Data Protection Regulation.

Unlike approaches such as Hasan et al [15], Biscotti [19], and Chang et al [20], which rely on homogeneous model structures or gradient-based updates, FedEnTrust supports heterogeneity-aware model assignment. Each participant trains a locally suitable model (eg, random forest, extreme gradient boosting, decision tree, support vector machine [SVM], k-nearest neighbors [KNN]) based on its available resources, enabling participation from hospitals, clinics, kiosks, and wearable devices.

FedEnTrust also differs from blockchain-enabled systems such as Shalan et al [24] and TinyFL [25]. While these frameworks integrate blockchain for logging or access control, they do not incorporate soft-label knowledge distillation or adaptive ensemble aggregation to unify heterogeneous model outputs. FedEnTrust introduces a unique coupling of soft-label–based distillation with blockchain-enforced RBAC, enabling secure verification of participant identity and role prior to model update submission, on-chain logging of update hashes to ensure auditability, prevention of malicious or unauthorized contributions before they influence aggregation, and interoperability of predictions across diverse model architectures.

This integration ensures that only authenticated, validated soft labels contribute to the global model. This design is particularly effective for non-IID and imbalanced health care data settings, where traditional gradient-averaging approaches struggle.

Module 1: Decentralized Local Training With Heterogeneous Models

FedEnTrust begins with a network of decentralized health care participants, including large hospitals, regional clinics, kiosks, and personal health trackers, each training its own machine learning model locally. These models are tailored to each participant's computational capabilities and data volume. For example, high-resource hospitals may use deep neural networks, while low-resource settings use shallow learning such as KNN or support vector classifier (SVC) to support real-time inference with minimal memory demands.

This heterogeneity-aware model assignment ensures that all participants, regardless of scale or technical capacity, can contribute meaningfully. Local training is performed privately using internal datasets, aligning with privacy regulations such as HIPAA and GDPR.

Module 2: Knowledge Distillation via Soft Labels

To facilitate collaborative learning without exposing raw data, participants generate soft labels, probability distributions over prediction classes (eg, diabetic, nondiabetic). These soft labels encode richer information than binary outputs and are shared with a central aggregator, enabling cross-site knowledge transfer.

Soft Label Generation

Each participant generates soft labels, probability distributions reflecting its model's confidence across classes, and transmits these predictions to the aggregator. Unlike gradient-based approaches, soft labels create an interoperable representation across heterogeneous model types. Before being used for ensemble aggregation, every soft label submission is paired with metadata including local validation accuracy, model identifier, and round number. For an input instance *x*, the participant's model outputs a probability vector:

$$(1) P_i(x)=[p_1,p_2,\ldots,p_C]\in R^C, \text{ where } \sum_{c=1}^{C} p_c=1$$

These soft labels encapsulate the model's confidence across the Cclasses and support knowledge transfer without sharing raw patient data or internal model parameters.

To address differences in how heterogeneous models calibrate probability outputs, FedEnTrust applies temperature scaling, which smooths the probability distribution by dividing logits $z_i(x)$by a temperature parameter T:

$$(2) P_i^{(t)}(x)=\text{softmax}\left(\frac{z_i(x)}{T}\right),\ T=2$$

A temperature of T=2 was selected because values greater than 1 produce smoother, less overconfident probability distributions, which improves the stability of aggregation across models with different calibration characteristics. A small temperature (eg, T=1) can lead to overly sharp probabilities that amplify noise, while excessively large values dilute useful predictive signals. Empirical testing showed that T=2 offers an optimal balance.

Dynamic Weight Updates Across Federated Rounds

Once soft labels are generated by each participant model, the system proceeds to combine these distributed outputs into a unified global prediction. This ensemble consensus represents a key step in transferring collective intelligence across all nodes while respecting the constraints of data privacy and computational diversity.

The ensemble aggregation process employs adaptive weighted soft voting, where more reliable and accurate models are given stronger influence. For example, a well-resourced clinic with consistently high validation performance will contribute more to the global prediction than a basic kiosk with limited data. However, no participant is excluded; each contributes according to its validated strength, ensuring fairness and inclusivity in the learning process. FedEnTrust adaptively updates the influence of each participant during communication round t. Each participant evaluates its model using a shared public validation subset to compute Acc_i^t, which is the validation accuracy of participant i at round t. The ensemble assigns each participant a normalized contribution weight:

$$(3) W_i^{(t)}=\frac{Acc_i^{(t)}}{\sum_{j=1}^{N} Acc_j^{(t)}}$$

To prevent dominant institutions (eg, large hospitals with more data) from exerting disproportionate influence, FedEnTrust applies weight clipping, capping W_i^tat an upper bound. This ensures contribution fairness and reduces the risk of bias toward specific demographic subpopulations.

Justification for Heterogeneous Model Assignment

The model architectures listed in Table 2 were selected to reflect realistic resource constraints and deployment contexts:

- Random forest (hospitals): Hospitals possess sufficient computational capacity and large datasets; random forest models capture nonlinear relationships and perform well on tabular clinical data.
- XGB (regional clinics): XGB provides strong performance under moderate computational resources, making it suitable for mid-sized clinics.
- Decision trees and KNN (community clinics or kiosks): These models require minimal training cost and support real-time inference in low-power environments.
- Linear SVM (wearables or personal trackers): Linear SVM has a lower memory footprint than logistic regression and offers more stable performance on small, noisy physiological samples typically produced by wearables.

Table . Simulated decentralized participants and their models.

ID	Participant	Model architecture	Key parameters	Resource level	Weight	Remarks
1	Large hospital	Random forest	n_estimators=130max_depth=15max_features=0.75 data_use=50%	Very high	0.50	Trains complex models on large datasets; serves as a high-capacity node
2	Urban health kiosk	K-nearest neighbors	n_neighbors=5 algorithm='auto'data_use=5%	Low	0.05	Designed for low-resource environments using simple, efficient models
3	Regional clinic	XGBoost	learning_rate=0.01max_depth=10n_estimators=180data_use=30%	High	0.30	Supports moderately complex modeling on medium-sized datasets
4	Community health clinic	Decision tree	max_depth=Nonecriterion='gini'data_use=10%	Medium	0.10	Runs interpretable tree-based models with moderate resource needs
5	Personal health tracker	Support vector machine	kernel='linear'C=1.0data_use=5%	Very low	0.05	Uses lightweight models suitable for wearables and embedded devices

This heterogeneity-aware mapping allows each participant to train a model aligned with its resource profile while still contributing to a unified ensemble.

Enhanced Knowledge Distillation and Pseudo-Label Generation

In each communication round t, participant models generate calibrated soft probability vectors $P_i^t(x)$, which are aggregated using dynamically updated participant weights to produce a global soft prediction.

Our proposed model aggregates the calibrated soft labels using the dynamic weights to produce a global soft prediction:

$$(4)\quad P^t(x)=\sum_{i=1}^{N} W_i^t * P_i^t(x)$$

Because aggregation operates entirely on probability distributions rather than gradients or model parameters, FedEnTrust naturally supports heterogeneous machine learning architectures across hospitals, clinics, kiosks, and personal wearable devices while preserving data locality and privacy.

To improve the reliability of knowledge transfer, each participant's soft predictions undergo normalization followed by temperature scaling (with T=2) to smooth overconfident outputs. The ensemble output is then evaluated using a confidence-based filtering mechanism, where pseudo-labels are generated only if the maximum ensemble probability satisfies:

$$(5)\quad \max(P^t(x))\geq\tau$$

With τ=0.7 Predictions failing this criterion are discarded to prevent the propagation of uncertainty or noise. Accepted pseudo-labels are normalized and redistributed to participants, where they are appended to local datasets and used for continued training in the subsequent round. This feedback loop enables low-resource participants to benefit from globally distilled knowledge while retaining local autonomy.

All soft-label submissions are validated through the blockchain-based RBAC mechanism described in Module 3. Only soft labels originating from authenticated and authorized roles (eg, model-provider) are accepted. Validated submissions are incorporated into an adaptive weighted soft-voting process, where participant weights are updated based on observed local performance across rounds. The resulting global outputs are then redistributed as pseudo-labels for the next training iteration, ensuring robustness against non-IID data distributions, preventing malicious or fabricated updates, and enhancing cross-site generalization across heterogeneous health care environments.

Module 3: Blockchain-Based Secure Coordination

Overview

Module 3 employs an Ethereum-based smart contract to authenticate participants, enforce role permissions, and log immutable update metadata. When a node attempts to upload soft labels, the smart contract verifies the participant's role, identity, timestamp, and declared accuracy. The contract then generates and stores a hashed representation of the update, which validator nodes review. Only soft labels that receive approval from multivalidators are admitted to the aggregation pool. This ensures tamper resistance, prevents poisoning attacks, and provides end-to-end traceability for health care compliance requirements. When a participant attempts to contribute soft labels, the smart contract performs the following checks:

1. Identity verification: Confirms that the contributor is a registered network participant.
2. Role validation: Ensures the contributor holds a permitted role to submit model outputs.
3. Metadata verification: Confirms the integrity of reported metrics (eg, accuracy, round number).

4. Hash logging: Stores a transaction hash to provide auditability without exposing any data.

Only after passing these checks is the soft label included in the aggregation pool. This design prevents poisoned or fabricated updates from influencing the global model and eliminates single points of failure in participation management. By integrating RBAC directly with knowledge distillation, FedEnTrust establishes a secure and transparent trust layer that coordinates collaborative learning across diverse health care nodes.

Blockchain Platform Selection and Justification

FedEnTrust is implemented on an Ethereum-compatible private blockchain network. Ethereum was selected due to its deterministic smart contract execution, robust security guarantees, and mature tooling ecosystem. The platform supports Solidity-based smart contracts, Remix IDE integration, and widely adopted standards for access control and event logging. These characteristics make Ethereum well suited for privacy-preserving health care collaboration, where verifiable execution and auditability are required.

To justify this choice, we compared Ethereum with 2 commonly used permissioned blockchain platforms: Hyperledger Fabric and Corda. Table 3 presents a feature-level comparison of Ethereum, Hyperledger Fabric, and Corda across network type, decentralization, smart contract support, privacy mechanisms, ecosystem maturity, and application alignment. Given the need for flexible smart contract logic, verifiable coordination, and broad compatibility with Internet of Things (IoT) and health care prototypes, Ethereum provides the most practical platform for FedEnTrust.

Table . Comparison of blockchain platforms.

Feature	Ethereum	Hyperledger fabric	Corda
Network type	Public or private	Permissioned	Permissioned
Decentralization	Highly decentralized	Semi-decentralized	Semi-decentralized
Smart contracts	Solidity, robust tooling	Chaincode (Go/Java/Node.js)	Contract flows for financial logic
Privacy	Extensible via Layer-2/private networks	Strong privacy (channels, private collections)	Strong bilateral privacy
Ecosystem	Very large developer ecosystem	Enterprise-focused	Financial institutions
Use alignment	Decentralized coordination across heterogeneous nodes	Consortium-style enterprise networks	Regulated financial workflows

Adversarial Threat Model and Security Resilience

FL deployments in real-world health care environments may be exposed to adversarial participants attempting to manipulate the global model, disrupt training, or infer sensitive information. To address these risks, we construct a structured threat model covering three primary attack categories: (1) model poisoning; (2) collusion among compromised participants; and (3) malicious soft-label injection, where adversaries submit manipulated pseudo-probabilities to bias the aggregation process.

FedEnTrust incorporates multiple, tightly coupled defense mechanisms across its blockchain coordination and ensemble aggregation layers to provide resilience against these threats.

1. Model poisoning and malicious soft-label injection: A compromised participant may attempt to submit adversarial or fabricated soft labels to influence global predictions. FedEnTrust mitigates this risk through smart contract–enforced RBAC, which restricts update submission exclusively to authenticated participants holding an authorized model-provider role. Each submission is accompanied by metadata including round number, reported validation accuracy, and timestamp, which are verified for internal consistency before acceptance. To ensure integrity and prevent replay or tampering, all submissions are cryptographically hashed and logged on-chain. Furthermore, FedEnTrust employs validator redundancy, requiring approval from multiple trusted validator nodes (eg, lead hospitals within the consortium) before a submission is incorporated into aggregation, preventing single-node compromise.
2. Collusion and validator compromise: To reduce the impact of colluding or compromised participants, FedEnTrust adopts a consortium-style multivalidator approval mechanism. No single validator can independently approve a model update; instead, a quorum of validators must jointly authorize submissions. The validator set itself is managed through governed smart contract functions, allowing secure updates to validator membership over time and eliminating static trust assumptions.
3. Blockchain-specific threats: Public blockchain deployments may be vulnerable to front-running, transaction reordering, or gas manipulation attacks. FedEnTrust avoids these risks by operating on a private Ethereum-compatible consortium network without a public mempool, eliminating front-running opportunities. Smart contracts use fixed gas budgets and sequential transaction counters to ensure deterministic execution and prevent reordering attacks.
4. Privacy leakage through on-chain metadata: Although raw data and model parameters are never shared, metadata leakage can still pose privacy risks. FedEnTrust minimizes exposure by storing only hashed identifiers and role-verification logs on-chain. No patient-level attributes, raw predictions, or model parameters are recorded. All soft labels remain strictly off-chain and are exchanged only between authorized participants and the aggregator over secure channels.
5. Aggregation-level safeguards: Beyond blockchain enforcement, the adaptive ensemble layer further mitigates

adversarial influence by applying temperature scaling, confidence thresholds, and weight clipping. These mechanisms limit the amplification of extreme or adversarial soft-label probabilities and restrict the maximum influence any single participant can exert, even if it reports high accuracy.

Collectively, these mechanisms establish a multilayered security architecture that protects FedEnTrust against common poisoning, collusion, and manipulation attempts at the coordination and authorization layers while preserving decentralized operation and data privacy. The empirical results demonstrate that unauthorized and malicious submissions are consistently detected and rejected through blockchain-enforced RBAC and validator checks. While this study focuses on secure enforcement and system robustness rather than controlled adversarial learning simulations, the framework is explicitly designed to support future evaluation against targeted and untargeted attacks, including label-flipping, probability-shifting, and adaptive adversarial strategies.

Module 4: Adaptive Model Aggregation and Feedback Loop

After soft labels are aggregated into a global ensemble prediction, FedEnTrust redistributes this consensus to participants as pseudo-labels for retraining. This adaptive aggregation ensures that high-performing models contribute more to the global prediction, while low-resource nodes still benefit from the collective knowledge.

This module enables faster convergence across non-IID data, fair and inclusive participation, and improved generalization without data sharing.

The result is a balanced feedback loop: local models become more aligned with the ensemble, improving personalization and global performance over time.

System Implementation and Evaluation Setup

We evaluated FedEnTrust using the publicly available PIMA Indians Diabetes Dataset [26], which includes 768 records of female patients with 8 clinical attributes and a binary diabetes outcome. Data were preprocessed using the following steps:

1. Outlier detection with IQR and local outlier factor
2. Feature engineering (eg, binning glucose, insulin levels)
3. Normalization using *z* scores
4. Class balancing using the synthetic minority oversampling technique [27]

As shown in Table 1, to simulate a real-world heterogeneous environment, the dataset was split across 5 simulated participants with varying data volumes and models. Each participant's computational weight was reflected in the aggregation process, mimicking operational conditions ranging from large hospitals to low-power personal devices.

Ethical Considerations

This study exclusively used publicly available, deidentified secondary datasets. No new data were collected, and no interaction with human participants occurred. According to institutional policy and US federal regulations (45 CFR 46), research involving publicly available, deidentified data does not constitute human participant research and is therefore exempt from institutional review board review. As a result, institutional review board approval was not sought, and informed consent was not required. All datasets used in this study were fully deidentified prior to public release. The data contained no direct or indirect identifiers, and no attempt was made to reidentify individuals. Data were accessed and analyzed in accordance with the terms and conditions specified by the data providers. No participants were recruited for this study, and no compensation was provided.

Results

Model Performance

We evaluated the FedEnTrust framework across 5 heterogeneous participants over 15 communication rounds, focusing on prediction accuracy, precision, recall, and F_1-score. The results highlight how collaborative learning and adaptive aggregation significantly enhance performance, especially for participants with limited data and computational resources.

Figure 2 shows the accuracy trajectories of each participant over the FL rounds. Participant 1 (random forest), equipped with the largest dataset and the highest computational power, consistently achieved the highest accuracy, acting as a de facto "teacher" during knowledge distillation. Its influence helped guide improvements in lower-resource nodes, such as participant 5 (SVC) and participant 2 (KNN), which showed steady gains over time.

Figure 2. Global model accuracy improves over ensemble federated round. DT: decision tree; KNN: k-nearest neighbors; RF: random forest; SVC: support vector classifier; XGB: extreme gradient boosting.

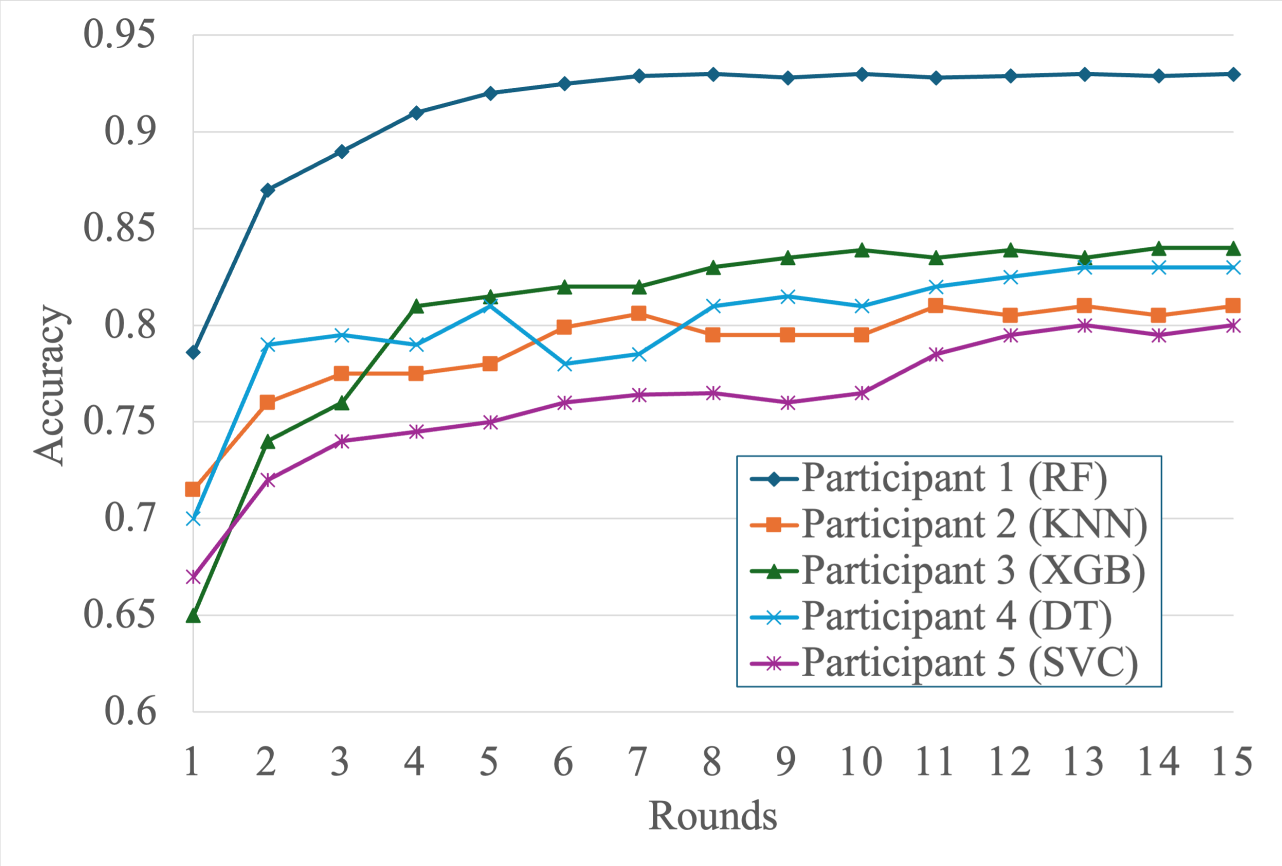


Figure 3 presents the corresponding model loss curves. All participants experienced substantial loss reduction early on, with convergence observed by round 15. Participant 1 maintained the lowest loss throughout, while participants 4 and 5 showed marked improvement from higher initial losses, demonstrating the benefit of federated collaboration.

Figure 3. Federated model losses over rounds. DT: decision tree; KNN: k-nearest neighbors; RF: random forest; SVC: support vector classifier; XGB: extreme gradient boosting.

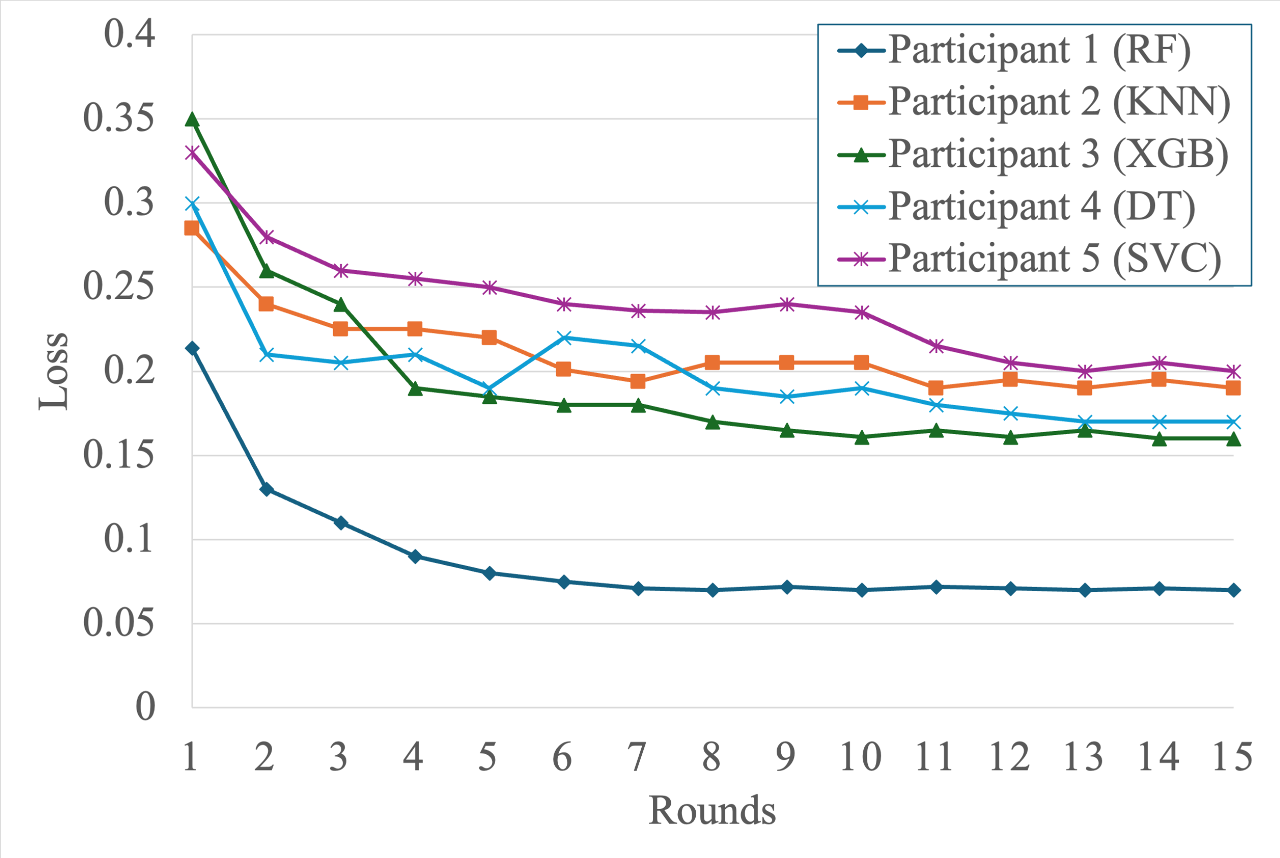


Table 5. Federated Models Performance after 15 rounds

Comparing the initial and federated performance results (Tables 4 and 5) reveals substantial gains for all participants after collaborative training. Accuracy improvements of up to 28% are observed in lower-resource participants, and F_1-scores increase consistently across all models, demonstrating the effectiveness of knowledge distillation and adaptive aggregation in heterogeneous environments. For example, participant 4 (decision tree) improves its F_1-score from 0.71 to 0.88, while participant 3 (XGBoost) improves from 0.64 to 0.85, highlighting the benefits of ensemble-driven knowledge transfer.

Table . Initial models' performance.

Participant	Accuracy	Precision	Recall	F_1-score
1	0.78	0.85	0.84	0.83
2	0.71	0.73	0.71	0.72
3	0.65	0.63	0.65	0.64
4	0.70	0.73	0.71	0.71
5	0.67	0.67	0.68	0.67

Table . Federated models' performance after 15 rounds.

Participant	Accuracy	Precision	Recall	F_1-score
1	0.93	0.92	0.94	0.93
2	0.81	0.80	0.86	0.83
3	0.84	0.85	0.86	0.85
4	0.83	0.87	0.90	0.88
5	0.80	0.79	0.87	0.83

To further characterize performance stability across communication rounds, Table 6 reports both the final accuracy at round 15 and the mean (SD) of accuracy over all 15 federated rounds. The relatively low SDs indicate stable convergence behavior for all participants, even for lightweight models such as KNN and SVC. These results confirm that FedEnTrust effectively accommodates device and data heterogeneity while maintaining strong predictive performance, privacy preservation, and decentralized operation. Tailored model architectures, aligned with participant resource constraints, ensure balanced contribution and efficient deployment across the collaborative learning process.

Table . Federated model accuracy and variability across 15 rounds.

Participant	Model	Final accuracy	Accuracy, mean (SD)
1	RF[a]	0.93	0.91 (0.04)
2	KNN[b]	0.81	0.79 (0.03)
3	XGB[c]	0.84	0.81 (0.05)
4	DT[d]	0.83	0.80 (0.03)
5	SVC[e]	0.80	0.76 (0.03)

[a]RF: random forest.

[b]KNN: k-nearest neighbors.

[c]XGB: extreme gradient boosting.

[d]DT: decision tree.

[e]SVC: support vector classifier.

To assess whether the performance differences between FedEnTrust and baseline models were statistically meaningful on the PIMA Indians Diabetes Dataset, we conducted a nonparametric bootstrap significance analysis using the same held-out test set as the main evaluation. Because accuracy, precision, recall, and F_1-score are bounded metrics that may deviate from normality, bootstrap resampling provides a distribution-free and robust alternative to parametric methods such as the *t* test. We used a 2-tailed *t* test, as no directional assumption was imposed a priori and the objective was to assess whether there was any statistically significant difference between the compared methods.

We generated B=1000 bootstrap resamples by sampling test instances with replacement from the held-out evaluation set. For each bootstrap resample, we evaluated FedEnTrust and the decentralized baseline from Blockchain-FL with Differential Privacy [20], which represents the closest methodologically comparable prior work under similar privacy and decentralization constraints. This procedure produced 1000-sample empirical distributions for both models' accuracy. To quantify comparative performance, we computed the bootstrap metric difference for each resample:

$$(6)\Delta(b)=MFedEnTrust(b)-MBaseline(b)$$

where Mb represents the accuracy, precision, recall, or F_1-score on bootstrap resample b. We then constructed 95% CIs for each metric difference using the percentile method.

The bootstrap CI analysis indicates that FedEnTrust achieves statistically significant performance improvements over the decentralized blockchain-based FL baseline [20]. Specifically, FedEnTrust attains a mean accuracy of 0.842 with a 95% bootstrap CI of 0.831-0.853, compared to 0.827 (0.814-0.839) for the decentralized baseline. The resulting accuracy difference of +0.015 yields a 95% CI of 0.004-0.027, which excludes zero, indicating statistical significance at α=.05. These results confirm that the performance gains observed for FedEnTrust are not due to random variation but rather stem from its integration of heterogeneous ensemble learning with blockchain-backed coordination under privacy constraints.

These findings validate that FedEnTrust's performance gains are not only empirical but statistically robust, reinforcing the effectiveness of combining heterogeneous ensemble learning with blockchain-backed coordination in constrained health care environments.

Blockchain Performance

We deployed the smart contract with 6 key functions and evaluated it under a realistic configuration consisting of 5 decentralized health care participants and 1 global aggregator. These components facilitated secure collaboration, access management, and federated training. The details are shown in Table 7.

Table . Blockchain system configuration.

Operation	Count	Description
Total registered participants	5	Registered using registerClient()
Federated coordination nodes	1	Global aggregator for accuracy aggregation and model ensemble
Smart contract functions deployed	6	Includes registration, role assignment, update logging, and access checks

To assess computational efficiency, we monitored key metrics such as gas consumption, data size, and latency for major smart contract operations. These measurements reflect the cost-effectiveness and responsiveness of blockchain-mediated tasks.

These operations incur gas overhead beyond Ethereum's 21,000 base gas due to additional computation, state updates, and event emissions. The *modelUpdate*() function, for example, consumes about 98,560 gas (~295 bytes of encoded parameters), balancing cost with functional depth and traceability (Table 8).

Table . Smart contract performance metrics.

Operation	Average gas cost	Data size (bytes)	Average latency (ms)
Client registration	118,073	370	220
Role assignment	109,820	345	210
Model update	98,560	295	195
Model aggregation	105,310	315	215

Despite slight delays compared to traditional systems, the observed latency (195 - 220 ms) remains acceptable for health care applications, considering the gains in trust, verifiability, and tamper resistance. To assess longer-term stability, we analyzed all 212 smart contract operations recorded during the training. All valid transactions executed successfully without anomalies, indicating stable performance across repeated interactions. The expanded evaluation in Table 9 includes average latency, latency range, and variability across extended cycles. These findings support the suitability of the blockchain layer for multiround federated training.

Table . Transaction integrity and enforcement metrics.

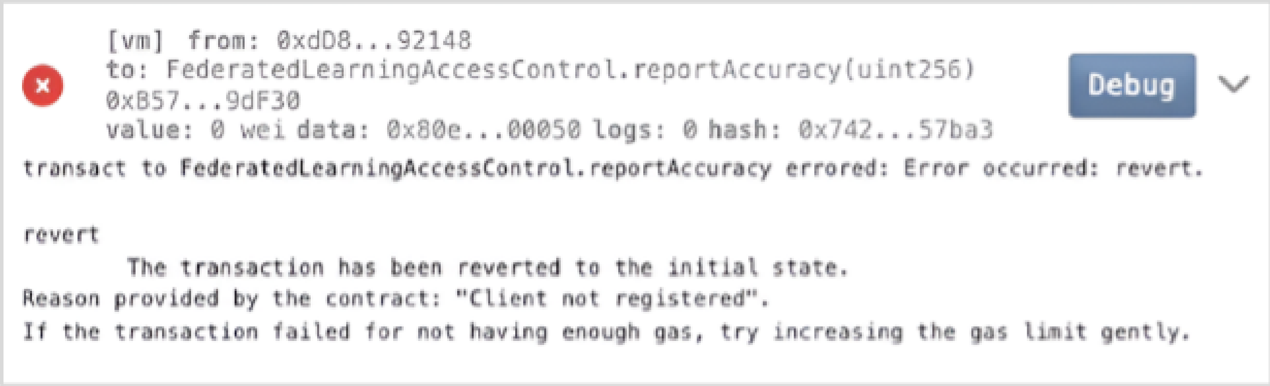
Category	Values	Description
Total transactions	212	All smart contract operations
Valid transactions	201	Successfully executed by authorized participants
Rejected transactions	11 (5.19%)	Unauthorized queries (6), malicious submissions (3), invalid role updates (2)
Success rate	100%	All valid transactions completed without error
Average latency	21.4 ms	Mean execution time for valid transaction
Latency range	14.8 - 36.2 ms	Minimum and maximum observed latency
SD	±4.7 ms	Variability in execution time
Latency over extended cycles (100 iterations)	Mean: 22.1 ms; variation:±5.3 ms	Long-term stability testing simulating multiround FL[a]
Finality time	~1 block(~1 s)	Deterministic finality in private PoA[b] Ethereum network
Estimated throughput	~47 tx/s	Consistent with private Ethereum networks

[a]FL: federated learning.

[b]PoA: proof-of-authority.

As illustrated in Figure 4, unauthorized model submissions are automatically rejected, triggering an on-chain error: "Client not registered." This ensures that only authenticated nodes contribute to the learning process, strengthening data integrity.

Figure 4. Access rejection for unauthorized participant.



Throughout 15 communication rounds, the smart contract reliably supported secure, real-time exchange of soft label predictions and model aggregation updates. For instance, participant 1 improved from 78% to 93% accuracy, while participant 4 rose from 70% to 83%, all while maintaining privacy and resisting tampering.

These results underscore the effectiveness of combining blockchain with federated ensemble learning to achieve scalable, secure, and privacy-preserving AI in health care environments.

Discussion

Principal Findings

This study presents FedEnTrust, a blockchain-enabled federated ensemble learning framework that offers a privacy-preserving and scalable solution for decentralized diabetes prediction. Our system effectively balances accuracy, privacy, and adaptability by integrating diverse machine learning models with knowledge distillation and adaptive weighted aggregation. With a predictive accuracy of 84.2%, FedEnTrust demonstrates competitive performance while maintaining strict privacy guarantees and supporting heterogeneous health care participants ranging from hospitals to wearable devices.

The framework's integration with blockchain smart contracts provides secure participant coordination, role-based access control, and transparent model validation without incurring substantial latency or resource overhead. Importantly, our results show that even low-resource participants benefit from collaboration through soft label exchange, enabling equitable participation in the learning process.

Comparison With Prior Work

Table 10 summarizes the performance of FedEnTrust against the existing centralized and decentralized methods applied to the PIMA Indians Diabetes Dataset. While centralized deep learning approaches achieve slightly higher accuracy (eg, 95.2% with light gradient boosting machine, 96.1% with convolutional neural networks), these models require full data centralization, sacrificing privacy and increasing system vulnerability.

Table . Comparative performance on the PIMA Indians Diabetes Dataset.

Model or study	Accuracy (%)	Precision (%)	Recall (%)	F_1-score (%)	Notes
FedEnTrust	84.2	84.6	88.6	86.4	Federated ensemble with adaptive weighted voting and blockchain smart contract integration
ML[a] classifiers approach [28]	95.2	N/A[b]	N/A	N/A	Centralized; evaluated multiple classifiers (LR[c], XGB[d], GB[e], DT[f], ET[g], RF[h], and LGBM[i]) on PIMA Indians dataset; best accuracy achieved by LGBM
Recursive feature elimination with a gated recurrent unit RFE-GRU[j] [29]	90.7	90.5	90.7	90.5	Centralized; utilized RFE-GRU on PIMA Dataset
Hybrid classification approach [30]	83.1	N/A	64.8	N/A	Centralized; applied SVM[k], RF, DT, naive Bayes with K-means preprocessing; best accuracy achieved by SVM
Three predictive algorithms [31]	77.1	N/A	N/A	N/A	Centralized; applied LR, RF, and ANN[l]; LR achieved the best accuracy (77.10%) with AUC[m] 0.83 over RF and ANN
Soft voting ensemble [32]	79.1	73.1	71.6	80.9	Centralized; combined RF, LR, and naive Bayes classifiers
Ensemble hierarchical model [33]	83.1	25.0 (positive)/98.6 (negative)	38.4 (positive)/90.2 (negative)	82.8	Centralized; applied DT and LR, fused by neural network
Stacking ensemble [25]	77.1	N/A	N/A	N/A	Centralized; stacking ensemble of ML models; accuracy achieved using cross-validation protocol
Deep learning pipeline [34]	92.3	N/A	N/A	N/A	Centralized; deep learning pipeline using VAE[n] for data augmentation, SAE[o] for feature augmentation, and CNN[p] for classification
Deep CNN with correlation-based features [35]	96.1	94.4	94.4	94.5	Centralized; applied deep CNN with feature selection based on correlation
Blockchain-FL with adaptive DP [20]	82.7	N/A	N/A	N/A	Decentralized; implemented federated learning with differential privacy using blockchain technology

[a]ML: machine learning.

[b]N/A: not applicable.

[c]LR: logistic regression.
[d]XGB: extreme gradient boosting.
[e]GB: gradient boosting.
[f]DT: decision tree.
[g]ET: extra tree.
[h]RF: random forest.
[i]LGBM: light gradient boosting machine.
[j]RFE-GRU: Recursive Feature Elimination with Gated Recurrent Unit.
[k]SVM: support vector machine.
[l]ANN: artificial neural network.
[m]AUC: area under the curve.
[n]VAE: variational autoencoder.
[o]SAE: stacked autoencoder.
[p]CNN: convolutional neural network.

In contrast, FedEnTrust improves over recent decentralized models, such as blockchain-integrated FL with differential privacy (accuracy≈82.7%), by incorporating ensemble learning and adaptive aggregation. Despite the constraints of data fragmentation and heterogeneity, our framework maintains robust performance across all key metrics, including precision (84.6%), recall (88.6%), and F_1-score (86.4%).

FedEnTrust achieves a favorable trade-off between privacy, generalizability, and computational practicality, making it well suited for real-world deployment in regulated health care environments.

Ethical AI Considerations: Fairness, Transparency, and Accountability

Ethical Framework

Ethical concerns are central to the deployment of AI systems in health care, where unequal access to computational resources and imbalanced data distributions may inadvertently create or reinforce model biases. FedEnTrust incorporates several design principles aligned with emerging ethical AI guidelines, including those recommended by the World Health Organization and major AI governance frameworks.

Fairness Across Heterogeneous Participants

Health care institutions vary substantially in data volume, demographic composition, and computational capacity, which can introduce systematic bias in collaborative learning systems. FedEnTrust is designed to mitigate such bias by supporting heterogeneity-aware participation, allowing low-resource nodes to contribute using models aligned with their capabilities without sacrificing predictive performance. Adaptive weight clipping is applied during aggregation to prevent high-resource institutions from disproportionately dominating the global ensemble. In addition, temperature-calibrated soft labels are used to reduce overconfidence from models trained on larger or more homogeneous datasets, while confidence thresholding ensures that noisy or low-confidence predictions are not propagated across participants. Together, these mechanisms promote more balanced influence across diverse health care contributors and support fairer model outcomes in heterogeneous federated environments.

Transparency and Auditability

Transparency in FedEnTrust is enabled through the blockchain-based coordination layer, which provides immutable audit trails for all update submissions and verifiable records of role validation events. Each model contribution is traceably logged, allowing the system to record which institutions participated in and influenced each training round. This tamper-resistant logging mechanism enhances accountability, supports post hoc auditing, and increases trust among participating health care entities without exposing sensitive data or model parameters.

Privacy and Data Minimization

FedEnTrust adheres to privacy-by-design principles:

- Raw patient data remain strictly on the device
- Only soft-label vectors and hashed metadata are transmitted
- No identifiable information is stored on-chain, supporting HIPAA, GDPR, and similar regulatory frameworks

Role-based access ensures that only authorized clinical entities may participate.

Accountability and Governance

The multivalidator consensus layer enables shared governance rather than reliance on a single coordinating institution. This creates a more accountable decision-making process and aligns with ethical expectations for distributed medical AI systems.

Blockchain Performance and Practical Considerations

Implementation Considerations

Beyond empirical accuracy and security validation, the practical deployment of blockchain-enabled FL systems in health care requires careful consideration of scalability, cost, and regulatory compliance. While the blockchain layer in FedEnTrust demonstrated stable and reliable performance under controlled experimental conditions, real-world health care environments introduce additional operational and governance challenges. This section discusses key practical considerations and outlines how FedEnTrust is designed to address them.

Scalability and Throughput

Public blockchain platforms, such as the Ethereum main net, face inherent constraints related to transaction throughput, block confirmation latency, and network congestion. These limitations can lead to unpredictable delays and may not support the repeated coordination required across multiple FL rounds. To address this, FedEnTrust is designed for deployment on private or consortium-based Ethereum networks, where consensus parameters, block times, and validator participation can be tailored to health care workflows. Such configurations enable deterministic execution and consistent performance, as observed in our evaluation. Nevertheless, large-scale deployments involving many institutions may require additional enhancements, including optimized validator load balancing, hierarchical or sharded blockchain structures, and integration with layer-2 scaling mechanisms to further increase throughput.

Cost Variability and Resource Requirements

In public blockchain environments, gas fees fluctuate dramatically based on network conditions, resulting in variable operational costs for smart contract execution. This variability is incompatible with cost-sensitive health care environments. Deploying FedEnTrust on a private Ethereum network eliminates transaction fees and allows institutions to control computational and storage overhead. However, operating such networks requires institutional commitment to maintain validator nodes, ensure uptime, and manage governance policies. Future work will investigate cost-benefit trade-offs between private, hybrid, and layer-2 blockchain configurations for FL.

Regulatory and Compliance Constraints

Health care systems must comply with strict privacy regulations such as HIPAA, GDPR, and provincial or national data-protection laws. These frameworks introduce challenges, such as prohibiting the storage of patient data or identifiers on-chain, requiring transparent audit trails for collaborative analytics, and ensuring that cross-institution coordination adheres to data-sharing agreements.

FedEnTrust addresses these concerns by storing only hashed metadata and role-verification entries on-chain, keeping soft labels and model outputs entirely off-chain. However, real-world deployment requires integration with institutional governance mechanisms to ensure compliance documentation, legal interoperability among institutions, and formal auditing procedures.

Generalizability to Multimodal and Longitudinal Health Care Data

Although the PIMA dataset provides a controlled benchmark for evaluating prediction accuracy, it does not reflect the complexity of real-world clinical environments. Modern health care systems generate multimodal data that may include structured electronic health record fields, laboratory values, medical imaging, clinician notes, and continuous wearable sensor streams. Additionally, many health conditions, including diabetes, require longitudinal modeling to capture evolving physiological states over time.

FedEnTrust is designed to naturally extend to these scenarios. The framework's heterogeneity-aware model assignment allows each participant to select model architectures aligned with its data modality and computational resources. For example, hospitals could train sequence models (eg, long short-term memories or transformers) on longitudinal EHR data, while wearable devices may contribute short-term physiological features via lightweight SVM or tree-based models. The knowledge-distillation component operates on probability distributions and is therefore agnostic to model type, enabling soft-label fusion across diverse modalities and temporal structures. This capability is particularly suitable for integrating outputs from time-series models, tabular models, and sensor analytics.

The blockchain-based coordination layer also supports generalization, as its role-based validation and update logging apply to any model output regardless of modality. Future work will apply FedEnTrust to multicenter datasets such as MIMIC-IV, NHANES, and integrated wearable–EHR cohorts to evaluate its performance under more heterogeneous and clinically realistic conditions.

Limitations

Despite promising results, several limitations remain:

- Dataset representativeness: The PIMA dataset is limited in scope and population diversity. Future work should evaluate FedEnTrust on broader, real-world datasets from varied demographics and geographies.
- Extreme client heterogeneity: Devices with ultra-low resources may still face difficulties in real-time model adaptation. Exploring ultra-lightweight architectures and communication compression techniques is a key next step.
- Controlled blockchain simulation: Our blockchain operations were simulated under stable conditions. Future deployment on public testnets or mainnets is necessary to assess real-world transaction delays, scalability, and cost variability.
- Advanced threat modeling: While the smart contract blocks unauthorized actions, adversarial behaviors such as collusion or model poisoning were not addressed. Future extensions may integrate anomaly detection and audit trails to enhance system resilience.

Although the PIMA Indians Diabetes Dataset is a well-established benchmark for evaluating diabetes prediction models, its limited demographic diversity and relatively small size restrict the generalizability of the findings. The simulated heterogeneous environment in Table 2, while constructed to reflect realistic participant variability, does not fully replicate the complexity of multi-institution health care settings, where differences in clinical practice, sensor characteristics, and patient demographics lead to substantially wider non-IID distributions. Accordingly, the results presented here should be viewed as a controlled feasibility demonstration rather than a comprehensive real-world validation.

Conclusions

This study presents FedEnTrust, a secure and intelligent federated ensemble learning framework for privacy-preserving diabetes prediction. Our approach addresses key challenges in decentralized health care AI, including data privacy, system trust, and participant heterogeneity, without requiring access to raw patient data.

By integrating knowledge distillation and adaptive ensemble aggregation, the framework enables resource-aware contributions from a diverse range of participants, from high-performance hospital systems to low-power personal devices. The experimental results demonstrate consistent improvements in predictive performance across all participants, validating both the effectiveness and inclusiveness of the design.

A central innovation is the blockchain-enabled coordination layer, which ensures secure registration, role-based access control, and verifiable model updates. Smart contract simulations confirm the system's efficiency, low latency, and robustness against unauthorized actions, supporting scalable and tamper-resistant deployment in health care environments.

In sum, FedEnTrust offers a practical, scalable solution for secure, decentralized medical AI, balancing privacy, performance, and trust. Future work will extend this framework to additional clinical domains, multisite studies, and dynamic personalization for broader impact in real-world health care.

Funding

The authors express their gratitude to the funding provided to support this study from the National Science Foundation with award number 2218046.

Conflicts of Interest

None declared.

References

1. Ong KL, Stafford LK, McLaughlin SA. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet 2023 Jul 15;402(10397):203-234. [doi: 10.1016/S0140-6736(23)01301-6] [Medline: 37356446]
2. Jaiswal V, Negi A, Pal T. A review on current advances in machine learning based diabetes prediction. Prim Care Diabetes 2021 Jun;15(3):435-443. [doi: 10.1016/j.pcd.2021.02.005] [Medline: 33642253]
3. Nomura A, Noguchi M, Kometani M, Furukawa K, Yoneda T. Artificial intelligence in current diabetes management and prediction. Curr Diab Rep 2021 Dec 13;21(12):61. [doi: 10.1007/s11892-021-01423-2] [Medline: 34902070]
4. Ahmed U, Issa GF, Khan MA, et al. Prediction of diabetes empowered with fused machine learning. IEEE Access 2022;10:8529-8538. [doi: 10.1109/ACCESS.2022.3142097]
5. Choudhur O, Gkoulalas-Divanis A, Salonidis T, et al. Differential privacy-enabled federated learning for sensitive health data. arXiv. Preprint posted online on Oct 7, 2019. [doi: 10.48550/arXiv.1910.02578]
6. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. J Healthc Inform Res 2021;5(1):1-19. [doi: 10.1007/s41666-020-00082-4] [Medline: 33204939]
7. Zisis K, Pavi E, Geitona M, Athanasakis K. Real-world data: a comprehensive literature review on the barriers, challenges, and opportunities associated with their inclusion in the health technology assessment process. J Pharm Pharm Sci 2024;27:12302. [doi: 10.3389/jpps.2024.12302] [Medline: 38481726]
8. Kairou P, McMahan HB, Avent B, et al. Advances and open problems in federated learning. arXiv. Preprint posted online on Dec 10, 2019. [doi: 10.48550/arXiv.1912.04977]
9. Islam H, Mosa A. A federated mining approach on predicting diabetes-related complications: demonstration using real-world clinical data. AMIA Annu Symp Proc 2021;2021:556-564. [Medline: 35308968]
10. Abbas SR, Abbas Z, Zahir A, Lee SW. Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with IoT integration. Healthcare (Basel) 2024 Dec 22;12(24):2587. [doi: 10.3390/healthcare12242587] [Medline: 39766014]
11. Liu J, Chen C, Li Y, et al. Enhancing trust and privacy in distributed networks: a comprehensive survey on blockchain-based federated learning. Knowl Inf Syst 2024 Aug;66(8):4377-4403. [doi: 10.1007/s10115-024-02117-3]
12. Lu Z, Pan H, Dai Y, Si X, Zhang Y. Federated learning with non-IID data: a survey. IEEE Internet Things J 2024 Jun;11(11):19188-19209. [doi: 10.1109/JIOT.2024.3376548]
13. Yang G, Edwards B, Bakas S, et al. Federated learning as a catalyst for digital healthcare innovations. Patterns (N Y) 2024 Jul 12;5(7):101026. [doi: 10.1016/j.patter.2024.101026] [Medline: 39081574]
14. Joshi M, Pal A, Sankarasubbu M. Federated learning for healthcare domain—pipeline, applications and challenges. ACM Trans Comput Healthcare 2022 Oct 31;3(4):1-36. [doi: 10.1145/3533708]
15. Hasan MR, Li Q, Saha U, Li J. Decentralized and secure collaborative framework for personalized diabetes prediction. Biomedicines 2024 Aug 21;12(8):1916. [doi: 10.3390/biomedicines12081916] [Medline: 39200380]
16. Patel VA, Bhattacharya P, Tanwar S, et al. Adoption of federated learning for healthcare informatics: emerging applications and future directions. IEEE Access 2022;10:90792-90826. [doi: 10.1109/ACCESS.2022.3201876]
17. Wulfert L, Wiede C, Grabmaier A. TinyFL: on-device training, communication and aggregation on a microcontroller for federated learning. Presented at: 2023 21st IEEE Interregional NEWCAS Conference (NEWCAS); Jun 26-28, 2023; Edinburgh, United Kingdom p. 1-5. [doi: 10.1109/NEWCAS57931.2023.10198040]

18. Niknam S, Dhillon HS, Reed JH. Federated learning for wireless communications: motivation, opportunities, and challenges. IEEE Commun Mag 2020 May;58(6):46-51. [doi: 10.1109/MCOM.001.1900461]
19. Shayan M, Fung C, Yoon CJM, Beschastnikh I. Biscotti: a blockchain system for private and secure federated learning. IEEE Trans Parallel Distrib Syst 2021 Jul;32(7):1513-1525. [doi: 10.1109/TPDS.2020.3044223]
20. Chang Y, Fang C, Sun W. A blockchain-based federated learning method for smart healthcare. Comput Intell Neurosci 2021;2021:4376418. [doi: 10.1155/2021/4376418] [Medline: 34868289]
21. Kang J, Xiong Z, Niyato D, Zou Y, Zhang Y, Guizani M. Reliable federated learning for mobile networks. IEEE Wireless Commun 2020 Apr;27(2):72-80. [doi: 10.1109/MWC.001.1900119]
22. Zhang H, Li G, Zhang Y, Gai K, Qiu M. Blockchain-based privacy-preserving medical data sharing scheme using federated learning. In: Knowledge Science, Engineering and Management: Springer; 2021:634-646. [doi: 10.1007/978-3-030-82153-1_52]
23. Tripathi G, Ahad MA, Casalino G. A comprehensive review of blockchain technology: underlying principles and historical background with future challenges. Decis Anal J 2023 Dec;9:100344. [doi: 10.1016/j.dajour.2023.100344]
24. Shalan M, Hasan MR, Bai Y, Li J. Enhancing smart home security: blockchain-enabled federated learning with knowledge distillation for intrusion detection. Smart Cities 2025 Feb;8(1):35. [doi: 10.3390/smartcities8010035]
25. Reza MS, Amin R, Yasmin R, Kulsum W, Ruhi S. Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. Heliyon 2024 Jan 30;10(2):e24536. [doi: 10.1016/j.heliyon.2024.e24536] [Medline: 38312584]
26. Chang V, Bailey J, Xu QA, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Comput Appl 2022 Mar 24:1-17. [doi: 10.1007/s00521-022-07049-z] [Medline: 35345556]
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2011;16:321-357. [doi: 10.1613/jair.953]
28. Ahamed BS, Arya MS, Nancy VAO. Prediction of type-2 diabetes mellitus disease using machine learning classifiers and techniques. Front Comput Sci 2022 May;4. [doi: 10.3389/fcomp.2022.835242]
29. Shams MY, Tarek Z, Elshewey AM. A novel RFE-GRU model for diabetes classification using PIMA Indian dataset. Sci Rep 2025 Jan 6;15(1):982. [doi: 10.1038/s41598-024-82420-9] [Medline: 39762262]
30. Edeh MO, Khalaf OI, Tavera CA, et al. A classification algorithm-based hybrid diabetes prediction model. Front Public Health 2022;10:829519. [doi: 10.3389/fpubh.2022.829519] [Medline: 35433625]
31. Darolia A, Chhillar RS. Analyzing three predictive algorithms for diabetes mellitus against the PIMA Indians Dataset. ECS Trans 2022 Apr 24;107(1):2697-2704. [doi: 10.1149/10701.2697ecst]
32. Kumari S, Kumar D, Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cogn Comput Eng 2021 Jun;2:40-46. [doi: 10.1016/j.ijcce.2021.01.001]
33. Abedini M, Bijari A, Banirostam T. Classification of Pima Indian Diabetes Dataset using ensemble of decision tree, logistic regression and neural network. Int J Adv Res Comput Commun Eng 2020 Jul 30;9(7):1-4. [doi: 10.17148/IJARCCE.2020.9701]
34. García-Ordás MT, Benavides C, Benítez-Andrades JA, Alaiz-Moretón H, García-Rodríguez I. Diabetes detection using deep learning techniques with oversampling and feature augmentation. Comput Methods Programs Biomed 2021 Apr;202:105968. [doi: 10.1016/j.cmpb.2021.105968] [Medline: 33631638]
35. Thushara A. An efficient Alzheimer's disease prediction based on MEPC-SSC segmentation and momentum geo-transient MLPs. Comput Biol Med 2022 Dec;151(Pt A):106247. [doi: 10.1016/j.compbiomed.2022.106247] [Medline: 36375415]

Abbreviations

AI: artificial intelligence
FL: federated learning
GDPR: General Data Protection Regulation
HIPPA: Health Insurance Portability and Accountability Act
IID: independent and identically distributed
KNN: k-nearest neighbors
RBAC: role-based access control
SVC: support vector classifier
SVM: support vector machine

Edited by S Li; submitted 16.Jun.2025; peer-reviewed by M Gheisari, S Patel; revised version received 13.Dec.2025; accepted 17.Dec.2025; published 26.Jan.2026.

Please cite as:
Hasan MR, Li J
Privacy-Preserving Collaborative Diabetes Prediction in Heterogeneous Health Care Systems: Algorithm Development and Validation of a Secure Federated Ensemble Framework
JMIR Diabetes 2026;11:e79166
URL: https://diabetes.jmir.org/2026/1/e79166
doi:10.2196/79166

Predictors of Glycemic Response to Sulfonylurea Therapy in Type 2 Diabetes Over 12 Months: Comparative Analysis of Linear Regression and Machine Learning Models

Shilpa Garg[1], BTech; Robert Kitchen[2], PhD; Ramneek Gupta[2,3], PhD; Emanuele Trucco[4], PhD; Ewan Pearson[1], MA, MB, BChir, PhD

[1]Diabetes Endocrinology and Reproductive Biology, School of Medicine, University of Dundee, Nethergate, Dundee, Dundee, United Kingdom

[2]Novo Nordisk, Oxford, United Kingdom

[3]Disease Intelligence Pte Ltd, 10 Anson Road, Singapore, Singapore

[4]Computing Department, School of Science and Engineering, University of Dundee, Dundee, United Kingdom

Corresponding Author:
Shilpa Garg, BTech
Diabetes Endocrinology and Reproductive Biology, School of Medicine, University of Dundee, Nethergate, Dundee, Dundee, United Kingdom

Abstract

Background: Sulfonylureas are commonly prescribed for managing type 2 diabetes, yet treatment responses vary significantly among individuals. Although advances in machine learning (ML) may enhance predictive capabilities compared to traditional statistical methods, their practical utility in real-world clinical environments remains uncertain.

Objective: This study aimed to evaluate and compare the predictive performance of linear regression models with several ML approaches for predicting glycemic response to sulfonylurea therapy using routine clinical data, and to assess model interpretability using Shapley Additive Explanations (SHAP) analysis as a secondary analysis.

Methods: A cohort of 7557 individuals with type 2 diabetes who initiated sulfonylurea therapy was analyzed, with all patients followed for 1 year. Linear and logistic regression models were used as baseline comparisons. A range of ML models was trained to predict the continuous change in hemoglobin A_{1c} (HbA_{1c}) levels and the achievement of HbA_{1c} <58 mmol/mol at follow-up. These models included random forest, extreme gradient boosting, support vector machines, a conventional feedforward neural network, and Bayesian additive regression trees. Model performance was assessed using standard metrics including R^2 and root mean squared error for regression tasks and area under the receiver operating characteristic for classification. In a subset of 2361 patients, nonfasting connecting peptide (C-peptide) was analyzed as a proxy for β-cell function. SHAP analysis was performed to identify and compare key predictors driving model performance across methods.

Results: All models exhibited similar performance, with no significant advantages of ML techniques over linear regression. For continuous outcomes, Bayesian additive regression trees demonstrated the highest R^2 (0.445) and lowest root mean squared error (0.105), though the differences among models were minimal. For the binary outcome, extreme gradient boosting achieved the highest area under the receiver operating characteristic curve (0.712), with CIs overlapping those of other models. Across all models, baseline HbA_{1c} was consistently the primary predictor, explaining the majority of the variance. SHAP analyses confirmed that baseline HbA_{1c}, age, BMI, and sex were the most influential predictors. Sensitivity analyses and hyperparameter tuning did not significantly improve model performance. In the C-peptide subset, higher C-peptide levels were associated with greater glycemic improvement ($\beta=-3.2$ mmol/mol per log(C-peptide); $P<.001$).

Conclusions: In this large, population-based cohort, ML models did not outperform traditional regression for predicting glycemic response to sulfonylureas. These findings suggest that limited gains from ML likely reflect an absence of strong nonlinear or high-order interactions in routine clinical data and that available features may not capture sufficient biological heterogeneity for complex models to confer added benefit. The inclusion of a C-peptide subset provides additional mechanistic insight by linking preserved β-cell function with treatment response.

(JMIR Diabetes 2026;11:e82635) doi:10.2196/82635

KEYWORDS

drug response; glycated hemoglobin; linear regression; machine learning models; treatment response prediction; type 2 diabetes

Introduction

Sulfonylureas are among the most commonly prescribed classes of glucose-lowering medications for individuals with type 2 diabetes. Their cost-effectiveness and accessibility make them particularly valuable in resource-constrained settings [1]. However, significant variability exists in glycemic responses among individuals. This variability is influenced by various clinical and biological factors, such as age, kidney function, and genetic predispositions [2,3]. Identifying predictors of treatment response is essential for advancing precision medicine approaches and minimizing trial-and-error prescribing practices [4].

Because sulfonylureas lower glucose primarily by stimulating insulin secretion from pancreatic β-cells, the degree of preserved β-cell function, often estimated by circulating connecting peptide (C-peptide) [5], may influence treatment response. However, such markers are rarely available in real-world datasets and are not routinely included in prediction studies.

Machine learning (ML) methods have shown promise in predicting treatment responses more accurately than traditional regression models, particularly due to their ability to handle complex, nonlinear interactions between variables without requiring prespecified assumptions [6,7]. In this context, ML approaches can capture subtle, multidimensional relationships that may be overlooked by traditional models, efficiently process large-scale longitudinal data, and generate data-driven insights that inform treatment selection. ML also offers better integration of diverse data types and improved interpretability through explainable AI, increasing clinical applicability [8]. Despite this promise, relatively few studies have focused on modeling glycemic response in diabetes using real-world data [9]. This gap in research presents a significant opportunity for further investigation.

Here, we use sulfonylurea response as an exemplar of diabetes drug response, due to its widespread use and the availability of clinical data. We evaluate and compare the efficacy of 5 ML models, including random forest, support vector machines, extreme gradient boosting (XGBoost), a conventional feedforward neural network (NN), and Bayesian additive regression trees (BART), in predicting the glycemic response to sulfonylureas in patients with type 2 diabetes. These models are compared with standard linear and logistic regression for continuous (change in hemoglobin A_{1c} [HbA_{1c}]) and binary (achievement of HbA_{1c} <58 mmol/mol) outcomes. Analyses were conducted using a large real-world cohort from the GoDARTS (Genetics of Diabetes Audit and Research in Tayside Scotland) study, including a biologically informative subset with C-peptide measurements to assess the contribution of β-cell function.

In addition to comparing predictive performance across models, we conducted a secondary analysis to examine feature contributions using Shapley Additive Explanations (SHAP) [10]. This analysis allowed us to determine whether ML-derived feature importance aligns with the predictors identified by traditional regression approaches, providing insight into clinical interpretability and the practical utility of ML for informing treatment choice.

Methods

Study Population

The data were obtained from the GoDARTS [11]. This population-based cohort links prescription, clinical, and laboratory records for individuals with diabetes in Tayside and Fife. The inclusion criteria included patients with type 2 diabetes who initiated sulfonylurea therapy (either as monotherapy or in combination), had a baseline HbA_{1c} measurement (defined as the closest value within 183 days before to 7 days after treatment initiation), and a follow-up HbA_{1c} measurement after a 1-year period. The 183-day window was selected to balance data availability and clinical relevance. For this analysis, only 2 HbA_{1c} values per patient were used, 1 at baseline and 1 at follow-up, in line with the model's aim of predicting glycemic response from initial clinical features.

Ethical Considerations

Data provision and linkage were carried out by the University of Dundee Health Informatics Centre, with analysis of anonymized data performed in an ISO27001 and Scottish Government–accredited secure safe haven. Health Informatics Centre standard operating procedures were reviewed and approved by the National Health Service (NHS) East of Scotland Research Ethics Service (22/ES/0034), and consent for this study was obtained from the NHS Fife Caldicott Guardian. Under these approvals, secondary analysis of anonymised routine healthcare data does not require additional participant consent or compensation.

Baseline Predictor Variables

Baseline clinical features included age, sex, HbA_{1c}, BMI, total cholesterol, high-density lipoprotein (HDL) cholesterol, smoking status, systolic blood pressure, alkaline phosphatase, alanine transaminase, serum potassium, serum creatinine, bilirubin, and albumin. These variables were selected based on their availability in routine care and their known or suspected relevance to glycemic outcomes [12]. Except for baseline HbA_{1c} (as defined above), all measurements were defined as the closest recorded value within 2 years before to 90 days after sulfonylurea initiation.

To estimate average daily sulfonylurea dose, prescription records were used to extract drug strength and quantity dispensed. Five sulfonylureas were included: gliclazide, glipizide, glimepiride, glibenclamide, and tolbutamide. Each prescription's dose was standardized by dividing the prescribed dose by the drug's maximum recommended daily dose (as per the British National Formulary). This yielded a standardized dose unit, which was then multiplied by the number of tablets prescribed per prescription to calculate the total standardized dose. For each patient, the total dose was summed across all prescriptions, excluding the last one, and divided by treatment duration to derive the average daily dose. This dose was then categorized into low, medium, and high using quartiles.

Outcome Definitions

The primary continuous outcome was defined as the change in glycated hemoglobin (HbA_{1c}), measured in millimoles per mole, from baseline (at the time of sulfonylurea initiation) to the follow-up measurement closest to 12 months, within a window of 6-15 months.

The binary outcome was defined as whether a patient achieved a follow-up HbA_{1c} level below 58 mmol/mol.

Data Preparation

To ensure consistency and compatibility with ML models, several preprocessing steps were applied. Continuous variables with skewed distributions underwent log transformation to approximate a normal distribution [13], enhancing model stability and reducing the influence of extreme values. Following this, all continuous predictors, including laboratory test results and physiological measurements, were scaled to a range between 0 and 1 using min-max normalization [14]. This rescaling placed variables on a uniform scale, which is particularly important for algorithms like NNs that are sensitive to variable magnitudes. Categorical variables (eg, sex, smoking status, treatment group, average daily dose) were converted using one-hot encoding to make them compatible with model inputs.

Missing Data Imputation and Collinearity Assessment

Patients missing either baseline or follow-up HbA_{1c} measurements were excluded. For remaining clinical predictors, missingness was below 10% and not clustered within specific individuals. Missing values were imputed using multiple imputation by chained equations [15] implemented in R (mice v3.18.0). Five imputed datasets were generated with 50 iterations each, using predictive mean matching for continuous variables. Full details of the imputation model are provided in Section 1 in Multimedia Appendix 1. Convergence was assessed using the mean and variance of each variable across iterations and comparing distributions of observed and imputed values. Analyses were performed on pooled estimates derived using Rubin's rules [16].

Collinearity among predictors was evaluated using variance inflation factors (VIFs) [17]. Predictors with VIF values greater than 5 were reviewed for redundancy. In our final models, VIFs ranged from 1.06 to 1.5, indicating no meaningful multicollinearity. As a sensitivity check, strongly correlated clinical variables ($r > 0.8$) were examined, and when overlap occurred (eg, estimated glomerular filtration rate vs serum creatinine), the variable more routinely and reliably measured in clinical practice (serum creatinine) was retained.

Statistical Analysis: Baseline Models

Initial statistical analyses were conducted using linear regression [18] for the continuous outcome and logistic regression for the binary outcome. These models identified baseline associations between clinical predictors and glycemic response to sulfonylurea therapy. Logistic regression estimated the probability of achieving an HbA_{1c} <58 mmol/mol. Of the 7557 individuals included, 3818 achieved the target, and 3739 did not.

Residualization of Baseline HbA_{1c}

To disentangle treatment response from baseline glycemia, change in HbA_{1c} was regressed on baseline HbA_{1c}. The residuals from this model were used as outcomes for ML analyses. This allowed the identification of predictors influencing glycemic response independent of baseline HbA_{1c} levels [19].

ML Models

Five ML models were implemented, reflecting diverse algorithmic strategies:

1. Random forest: An ensemble method that constructs multiple decision trees on bootstrapped data and aggregates their predictions [20].
2. Support vector machines: A kernel-based classifier that establishes an optimal separating hyperplane [21].
3. XGBoost: A boosting technique that sequentially builds trees to minimize residual error [22].
4. NNs: A conventional feedforward NN (multilayer perceptron) trained using resilient backpropagation. Comprising layers of interconnected neurons, NNs excel at modeling complex relationships and require larger datasets and regularization to mitigate overfitting [23].
5. BARTs: A Bayesian ensemble method that combines multiple regression trees and estimates uncertainty in predictions [24]. BART is noted for strong performance in clinical applications [25-27].

Model Implementation

For model development, a 2-stage validation framework combining cross-validation and a held-out test set was used. The data were randomly split into a 70% training set and a 30% held-out test set. Within the training set, 10-fold cross-validation [18] was used for hyperparameter tuning and model selection to enhance model stability and reduce overfitting. Final performance was evaluated on the held-out test set, which remained unseen during training.

All analyses were performed in R (version 4.3.0). A detailed description of the packages and functions used for each model is presented in Section 2 in Multimedia Appendix 1, and XGBoost and NN hyperparameters are provided in Sections 3 and 4 in Multimedia Appendix 1, respectively. All preprocessing and modeling code is publicly available on GitHub [28].

Feature Importance

To identify the clinical features most strongly influencing model predictions, we assessed feature importance using SHAP values along with the built-in variable importance metrics from each model. SHAP values quantify the contribution of individual predictors to model outputs, enabling transparent, model-agnostic interpretation. Although SHAP can be applied to multiple model types, our results focused on the XGBoost model because it showed optimal predictive performance. SHAP summary plots were generated to visualize the magnitude and direction of feature effects, ranking predictors by their mean absolute SHAP values. Comparative plots across models were generated to visualize predictor impact on treatment response. This unified approach supported consistent evaluation of feature relevance across models and enhanced clinical interpretability.

Performance Evaluation

Model performance was assessed separately for the continuous and binary outcomes. For the continuous outcome, evaluation metrics included root mean squared error (RMSE), mean absolute error, and the coefficient of determination (R^2), which indicates the proportion of variance in the outcome explained by the model [29,30].

For the binary outcome, performance was evaluated using standard classification metrics: area under the receiver operating characteristic curve (AUC), accuracy, sensitivity (recall), and specificity [31,32]. In the linear regression models, regression coefficients were interpreted to assess the direction and magnitude of each predictor's association with the outcome, while *P* values indicated the statistical significance of these associations. An R^2 value provided an overall measure of model fit, and a *P* value below .05 was considered statistically significant.

To evaluate differences in performance across models, a resampling-based approach was used to compare their predictive metrics [33]. Pairwise comparisons were conducted to assess whether any model significantly outperformed the others.

Results

Cohort Characteristics

The study included 7557 individuals with type 2 diabetes who initiated sulfonylurea therapy and had both baseline and follow-up HbA_{1c} values available. The cohort had a mean age of 63.7 (SD 11.8) years, and 57.9% (n=4377) were male. The mean baseline HbA_{1c} was 76.5 (SD 16.7) mmol/mol (Table 1).

Table . Baseline demographic and clinical characteristics of the study population.

Clinical variable	Sulfonylurea cohort (N=7557)
Age at therapy initiation (y), mean (SD)	63.7 (11.8)
Sex, n (%)	
Male	4377 (57.9)
Female	3180 (42.1)
Average daily dose, n (%)	
Low	1844 (24.4)
Medium	3822 (50.6)
High	1891 (25)
Duration of diabetes (y), mean (SD)	4.96 (4.49)
Duration of diabetes, n (%)	
0 - 1 years	1416 (18.7)
1 - 5 years	3099 (41)
>5 years	3042 (40.3)
Time of treatment (mo), mean (SD)	11.4 (2.2)
Time from baseline HbA_{1c}[a] measurement to treatment start (d), mean (SD)	21.3 (29.1)
Year of drug start, mean (SD)	2010 (6.12)
BMI (kg/m^2), mean (SD)	31.3 (6.3)
Total cholesterol (mmol/L), mean (SD)	4.5 (1.2)
HDL[b] cholesterol (mmol/L), mean (SD)	1.2 (0.3)
Serum creatinine (µmol/L), mean (SD)	80.3 (27.4)
Albumin (g/L), mean (SD)	42.2 (4.0)
Bilirubin (µmol/L), mean (SD)	9.9 (5.2)
Alkaline phosphatase (U/L), mean (SD)	89.5 (42.1)
ALT/SGPT[c] (U/L), mean (SD)	34.2 (24.4)
Potassium (mmol/L), mean (SD)	4.4 (0.4)
Systolic blood pressure, mean (SD)	137 (17.4)
Smoking status, n (%)	
Ever smoked—yes	5628 (74.5)
Ever smoked—no	1870 (24.7)
Ever smoked—unknown	59 (0.8)
Therapy group, n (%)	
Mono	2508 (33.2)
Dual	4251 (56.3)
Triple	798 (10.6)
Index of multiple deprivation quintile, n (%)	
1 (most deprived)	1583 (16.5)
2	1609 (21.3)
3	1497 (19.8)
4	1396 (18.5)
5 (least deprived)	1246 (16.5)

Clinical variable	Sulfonylurea cohort (N=7557)
Unknown	226 (3.0)
Ethnicity, n (%)	
White	5696 (75.4)
Others/mixed	259 (3.4)
Missing	1602 (21.2)
Region, n (%)	
Tayside	5965 (78.9)
Fife	1592 (21.1)
Baseline HbA_{1c} (mmol/mol), mean (SD)	76.5 (16.7)
HbA_{1c} outcome (mmol/mol) (treatment HbA_{1c}), mean (SD)	61.1 (15.4)
HbA_{1c} response (change from baseline; mmol/mol), mean (SD)	–15.4 (18)

[a]HbA_{1c}: hemoglobin A_{1c}.

[b]HDL: high-density lipoprotein.

[c]ALT/SGPT: alanine aminotransferase/serum glutamate pyruvate transaminase.

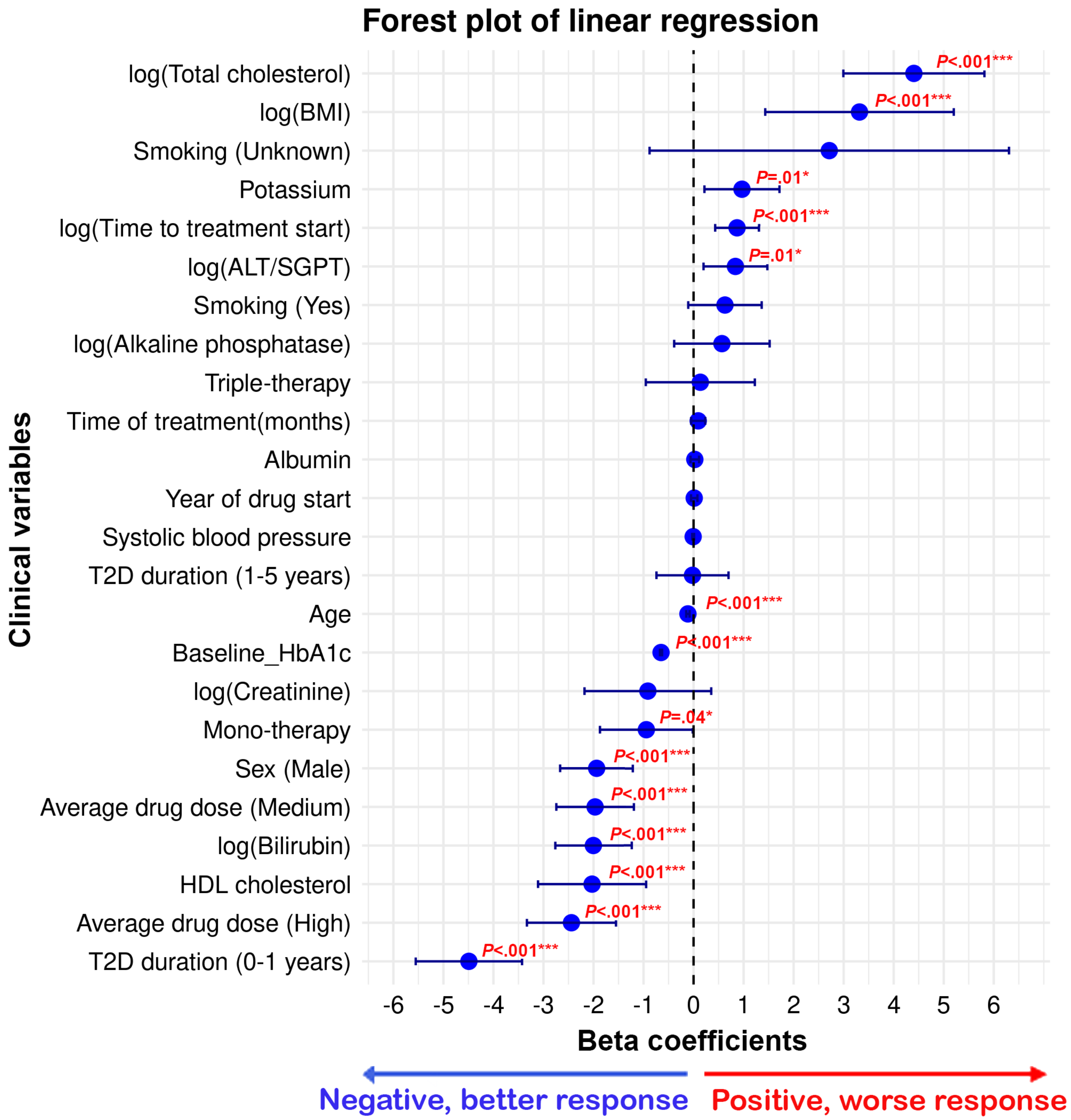
Associations Between Clinical Covariates and Treatment Response

A linear regression model was fit to assess the relationship between clinical variables and the change in HbA_{1c}, defined as the difference between baseline HbA_{1c} and follow-up values (ie, change=treatment–baseline HbA_{1c}). A negative change in HbA_{1c} indicates a better treatment response to sulfonylureas, while a positive change signifies a worse response. No feature scaling (min-max normalization) was applied to the variables in this model. The model demonstrated satisfactory fit, with an R^2 value of 0.41 and a significant F-statistic (222.2; P<.001).

The linear regression analysis identified several clinical variables significantly associated with the change in HbA_{1c} among individuals treated with sulfonylureas. Older age and higher baseline HbA_{1c} were associated with greater HbA_{1c} reductions (negative coefficients), indicating more favorable responses. In contrast, a higher BMI was associated with smaller reductions, suggesting that individuals with higher BMI may struggle to achieve desired glycemic control. Additionally, sex differences indicated that male participants demonstrated slightly better response to sulfonylurea treatment than female participants.

These associations are illustrated in Figure 1, which presents a forest plot of the linear regression coefficients and CIs, highlighting the magnitude and direction of each predictor’s effect on HbA_{1c} change.

Figure 1. Forest plot showing regression coefficients and 95% CIs for predictors of change in hemoglobin A_{1c} (HbA_{1c}). Points represent model estimates and horizontal lines indicate 95% CIs. ALT: alanine aminotransferase; SGPT: alanine aminotransferase/serum glutamate pyruvate transaminase; T2D: type 2 diabetes. **P*<.05; ***P*<.01; ****P*<.001.



C-Peptide Analysis

In a subset of 2361 individuals with nonfasting C-peptide data, higher C-peptide levels were strongly associated with greater reductions in HbA_{1c} at 12 months (linear regression: β=−3.2 mmol/mol per log(C-peptide); *P*<.001). This finding suggests that preserved endogenous insulin secretion contributes to more favorable treatment outcomes.

ML Model Performance for Continuous Outcome

For the continuous outcome of predicting changes in HbA_{1c}, several models were evaluated. The results indicate that the BART model exhibited the lowest RMSE of 0.105 (21%) and the lowest mean absolute error of 0.079 (16.1%), highlighting its effective performance in estimating continuous changes. XGBoost and NNs also performed comparably, with RMSE values of 0.106.

On the original HbA_{1c} scale, this corresponds to an approximate prediction error of 13.8 mmol/mol, comparable to the residual standard error from the linear regression model. Thus, the clinical prediction error was approximately ~14 mmol/mol.

However, the differences in RMSE and R^2 across all models were minimal, indicating that while BART performed slightly better, the performance of all models was relatively comparable

(Table 2). The similar R^2 and RMSE values further suggest that no single model stands out significantly.

Table . Regression model performance metrics (root mean squared error [RMSE], mean absolute error [MAE], and R^2) for continuous outcome prediction across all 6 models[a].

Models	RMSE	MAE	R^2
Linear regression	0.106	0.08	0.434
RF[b]	0.108	0.082	0.424
SVM[c]	0.106	0.079	0.438
XGBoost[d]	0.106	0.08	0.433
NN[e]	0.106	0.081	0.427
BART[f]	0.105	0.079	0.445

[a]RMSE is shown as the normalized values.
[b]RF: random forest.
[c]SVM: support vector machine.
[d]XGBoost: extreme gradient boosting.
[e]NN: neural network.
[f]BART: Bayesian additive regression trees.

Statistical Comparison of Model Performance

In addition to reporting standard performance metrics, statistical comparisons were performed using resampling-based techniques. Pairwise comparisons of RMSE and R^2 values across all models showed no statistically significant differences in performance; all ML models, including the linear regression baseline, performed similarly on this dataset.

Sensitivity Analysis: Residualized HbA_{1c} Change

A sensitivity analysis was performed to evaluate predictors of HbA_{1c} change independent of baseline glycemia. Across all models, the maximum R^2 value decreased to 0.05, indicating that only ~5% of the residual variance in 12-month HbA_{1c} response was explained by routine clinical features after removing the effect of baseline HbA_{1c}.

The performance metrics from this analysis further indicated that the RMSE and R^2 values remained consistent across most models (Table 3). However, XGBoost and BART showed poorer performance, with high RMSE and lower R^2 values. This likely reflects the fact that these algorithms are better suited for large, high-dimensional, or highly nonlinear datasets, whereas the present dataset may not contain sufficient complexity. This consistency across most models suggests that while some approaches explain marginally more variance in the sensitivity analysis, their predictive accuracy in terms of mean squared error remains stable.

Table . Model performance after adjustment for baseline hemoglobin A_{1c} (HbA_{1c}) across all 6 models.

Models	RMSE[a]	MAE[b]	R^2
Linear regression	0.127	0.095	0.054
RF[c]	0.126	0.095	0.056
SVM[d]	0.128	0.094	0.051
XGBoost[e]	0.230	0.183	0.01
NN[f]	0.127	0.095	0.053
BART[g]	0.214	0.191	0.021

[a]RMSE: root mean squared error.
[b]MAE: mean absolute error.
[c]RF: random forest.
[d]SVM: support vector machine.
[e]XGBoost: extreme gradient boosting.
[f]NN: neural network.
[g]BART: Bayesian additive regression trees.

ML Model Performance for Binary Outcome

For the binary outcome of predicting achievement of HbA_{1c} <58 mmol/mol, model performance was assessed using the AUC and accuracy. The XGBoost model achieved the highest AUC (0.712), followed closely by BART (0.710), indicating modest discriminatory ability. Logistic regression performed similarly, with an AUC of 0.702.

The CIs for all models showed substantial overlap (ranging from 0.681 to 0.724 for logistic regression and 0.692 to 0.733 for XGBoost), indicating that no model demonstrated statistically superior discrimination. Overall, the models were broadly comparable in their ability to distinguish responders from nonresponders.

Model-level classification metrics are summarized in Table 4, and the corresponding ROC curves for all models are shown in Figure 2, illustrating their similar performance profiles. In Figure 2, colored curves represent the individual models, visually reinforcing the overlapping AUCs and the absence of meaningful differences in classification performance.

Table . Discrimination and classification performance of binary outcome models.

Models	AUC[a] (95% CI)	Accuracy	Precision	Recall
Logistic regression	0.702 (0.681 - 0.724)	0.654	0.657	0.628
RF[b]	0.708 (0.687 - 0.729)	0.652	0.656	0.628
SVM[c]	0.705 (0.684 - 0.727)	0.65	0.656	0.618
XGBoost[d]	0.712 (0.692 - 0.733)	0.646	0.65	0.625
NN[e]	0.699 (0.678 - 0.72)	0.645	0.645	0.636
BART[f]	0.71 (0.689 - 0.731)	0.651	0.652	636

[a]AUC: area under the receiver operating characteristic curve.

[b]RF: random forest.

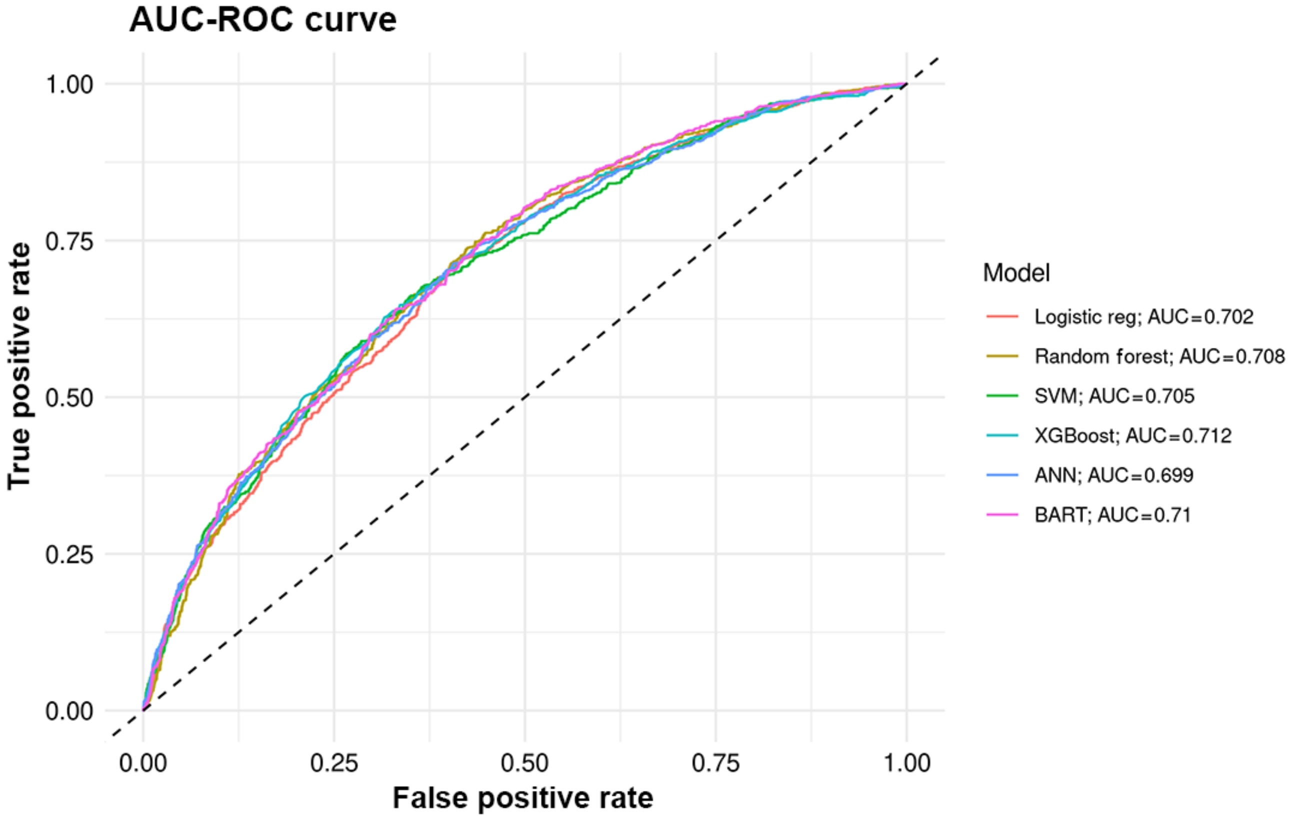
[c]SVM: support vector machine.

[d]XGBoost: extreme gradient boosting.

[e]NN: neural network.

[f]BART: Bayesian additive regression trees.

Figure 2. Receiver operating characteristic (ROC) curves for binary outcome prediction models. Colors correspond to individual models as shown in the legend. ANN: artificial neural network; AUC: area under the receiver operating characteristic curve; BART: Bayesian additive regression trees; SVM: support vector machine; XGBoost: extreme gradient boosting.

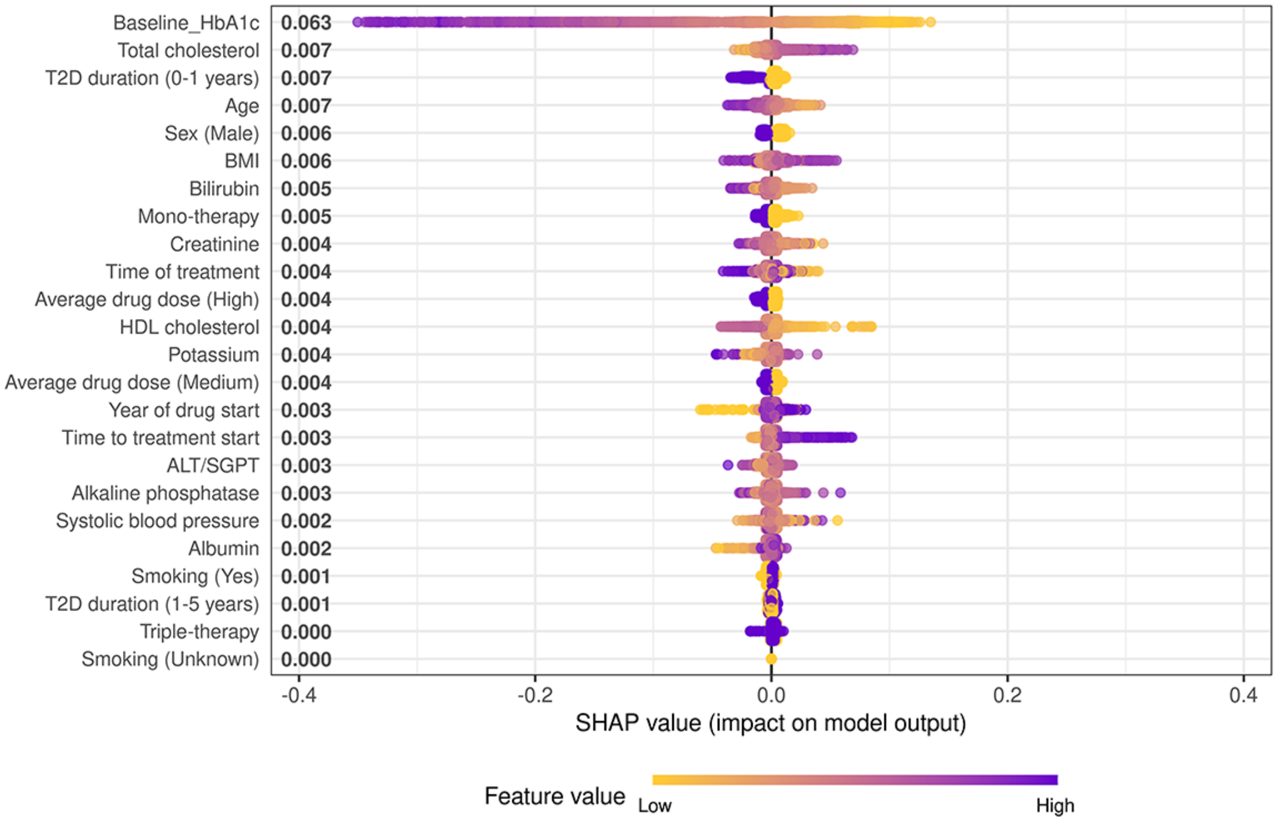


Feature Importance and SHAP Interpretability

Feature importance analyses consistently identified baseline HbA_{1c} as the most significant predictor across all models. Other variables such as BMI, alanine transaminase, total cholesterol, and systolic blood pressure were also found to be significant, though their rankings varied slightly between algorithms. Across all 5 models, the rankings of key predictors remained largely consistent.

To further explain feature contributions, a SHAP summary plot derived from the XGBoost model is presented (Figure 3), offering a more granular view of individual feature effects on model predictions. Baseline HbA_{1c} had the highest mean SHAP value (0.063), followed by total cholesterol, duration of diabetes, and age. Higher baseline HbA_{1c} values were associated with larger predicted reductions in HbA_{1c} (ie, more negative SHAP values), indicating greater expected treatment benefit.

Figure 3. Shapley Additive Explanations (SHAP) summary plot of feature importance in predicting glycemic response. Each dot represents 1 patient. The x-axis indicates the SHAP value (impact on model output). The color gradient reflects feature values (blue=higher values; yellow=lower values). Features are ordered by mean absolute SHAP values, indicating overall contribution to model predictions. ALT/SGPT: alanine aminotransferase/serum glutamate pyruvate transaminase; HbA_{1c}: hemoglobin A_{1c}; HDL: high-density lipoprotein; T2D: type 2 diabetes.



The SHAP plot further shows that lower total cholesterol values corresponded to more negative SHAP values, suggesting better predicted outcomes, whereas higher cholesterol was linked to reduced response. Similarly, shorter diabetes duration and younger age were associated with more favorable predictions. For the variable sex, blue points represent male participants and yellow points represent female participants; male participants were associated with more negative SHAP values, indicating a better predicted response, compared to female participants, whose SHAP values clustered closer to or above zero.

Discussion

Principal Findings

This study compared the predictive performance of traditional regression models and a range of ML algorithms for predicting glycemic response to sulfonylurea therapy in individuals with type 2 diabetes. The primary finding is that, with the dataset used, all models demonstrated comparable predictive performance. No ML approach significantly outperformed standard regression for either the continuous outcome or the binary outcome. These results indicate that, within routinely collected clinical data, the additional algorithmic complexity of ML methods does not necessarily yield superior predictive accuracy. Regression therefore remains a robust and interpretable option for predicting drug response in this context.

Linear regression analysis revealed that the model explained approximately 43% of the variance in changes to HbA_{1c}. In the sensitivity analysis, after adjusting for baseline HbA_{1c}, the maximum R^2 across all models dropped to 0.05, indicating that only a small proportion of outcome variability was captured by the remaining routine clinical features. This highlights the need for additional or more informative biomarkers to improve prediction.

Additionally, only about 50% (n=3818) of the participants achieved glycemic control after 1 year of sulfonylurea therapy, despite the relatively homogeneous clinical characteristics of the cohort. This finding highlights considerable interindividual variability in treatment response, suggesting that additional biological and behavioral factors may shape drug efficacy. Such heterogeneity may reflect differences in pharmacodynamic sensitivity, medication adherence, β-cell reserve, and underlying insulin resistance. BMI and HDL were considered indirect proxies of insulin resistance, as higher BMI is typically associated with greater insulin resistance, whereas higher HDL levels are generally linked to improved insulin sensitivity. Consistent with this, participants with higher BMI had poorer glycemic response, while those with higher HDL tended to show more favorable outcomes. However, direct measures of insulin resistance were not available in this dataset.

No Added Value From ML Methods

While multiple ML algorithms were evaluated in parallel with traditional regression models, none demonstrated superior predictive performance. Across both continuous and binary outcomes (Tables 2 and 4), differences in metrics such as RMSE, R^2, and AUC were small and not statistically significant, with overlapping CIs for all models. Even after hyperparameter tuning, predictive metrics remained modest, suggesting that ML methods did not uncover hidden patterns or interactions that traditional models missed.

This limited gain in predictive accuracy likely reflects the nature of routinely collected clinical data, which may lack sufficient biological complexity for ML algorithms to exploit. When input variables do not encompass detailed mechanistic information, even advanced algorithms cannot extract additional predictive signal. Consequently, transparent and easily interpretable models, such as linear or mixed-effects regression, may remain preferable, particularly when predictive performance is comparable. These models allow clinicians to understand feature contributions directly and translate findings into actionable treatment decisions.

Although complex ML models theoretically enable the capture of nonlinear relationships, their greater computational burden and reduced interpretability may limit their clinical utility unless they provide meaningful improvements in accuracy. The consistency of results across all modeling strategies, ranging from simple linear regression to ensemble and NN approaches, suggests that the available clinical features may not contain enough biological heterogeneity for ML methods to offer an advantage.

By intentionally comparing models of differing complexity, this study demonstrates that when data lack substantial nonlinearity or high-dimensional interactions, regression-based methods may remain more appropriate and efficient. This finding supports the continued reliance on interpretable models in routine clinical prediction tasks, where model parsimony and interpretability remain more valuable than algorithmic complexity for precision-medicine applications.

Features That Inform Drug Response Prediction

Feature importance and SHAP analyses consistently identified baseline HbA_{1c}, age, BMI, and sex as key predictors across regression and ML models. Baseline HbA_{1c} was the strongest predictor, reflecting both regression to the mean and true physiological responsiveness [34]. Older age and male sex were associated with greater HbA_{1c} reduction, whereas higher BMI predicted poorer response, aligning with evidence that adiposity may reduce sulfonylurea effectiveness [35]. These findings parallel results from the 5-drug predictive model developed by Dennis et al [36], suggesting that these core predictors generalize across therapeutic classes.

C-peptide, available for a subset of participants, showed a strong positive association with glycemic improvement, consistent with the insulin-secretagogue mechanism of sulfonylureas. This highlights the contribution of β-cell reserve to treatment heterogeneity and underscores the value of including mechanistic biomarkers to enhance model interpretability and predictive accuracy.

Limitations

This study has several limitations. First, routinely collected clinical data omit key determinants of treatment response such as adherence, diet, physical activity, genetics, and social factors. The limited availability of C-peptide prevented fuller assessment of β-cell function, and its strong association with response suggests that the inclusion of mechanistic biomarkers would likely improve predictive accuracy.

Second, the study population was limited to patients from Tayside and Fife in Scotland, which may reduce the generalizability of the findings to other regions or health care systems with different population characteristics or clinical practices.

Third, treatment response was assessed using a single HbA_{1c} value taken between 6 and 15 months after treatment initiation. Although data closest to 12 months post-initiation were used, variability in the follow-up period (6 - 15 mo) may introduce measurement variability and ought to be considered.

Additionally, the relatively low R^2 values across all models, even after applying rigorous methods such as a 70/30 train-test split and 10-fold cross-validation, suggest that the available clinical features alone do not explain sufficient variation in treatment response to support strong predictive performance.

Future Directions

Future research should focus on improving prediction models by incorporating richer and more diverse data sources, including genetic, metabolomic, and continuous glucose monitoring data, as well as direct measures of insulin resistance and β-cell function. Integrating these modalities could improve model accuracy and help explain why individuals respond differently to sulfonylurea treatment. In addition, future studies could explore advanced modeling approaches, such as deep learning or hybrid models that balance predictive power with ease of interpretation for clinical use. The increasing availability of real-world, longitudinal clinical data also supports the use of time-dependent models, such as recurrent neural networks or transformer models, to track how treatment response evolves over time. Finally, testing these models in independent and ethnically diverse populations will be important to assess their generalizability and real-world applicability.

In conclusion, this study shows that the traditional regression models remain robust, clinically interpretable, and sufficient for predicting glycemic response to sulfonylurea therapy using routine data. The comparable performance of ML methods suggests that model transparency and accessibility may currently outweigh the small gains offered by algorithmic complexity in this context.

Funding

This work was supported by funding from Novo Nordisk, which provided a PhD studentship to SG.

Authors' Contributions

EP, RK, and RG contributed to conceptualization and supervision. SG conducted the data analysis and wrote the original manuscript. SG, EP, RK, and RG contributed to the reviewing and editing. ET reviewed the manuscript.

Conflicts of Interest

RK is an employee of Novo Nordisk and serves as a director in Research and Development. RG is employed by Novo Nordisk and Disease Intelligence Pte Ltd and holds stock in both companies. The remaining authors declare no conflicts of interest.

Multimedia Appendix 1
Missing data imputation.
[DOCX File, 23 KB - diabetes_v11i1e82635_app1.docx]

References

1. Diabetes. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/conditions-and-diseases/diabetes-and-other-endocrinal--nutritional-and-metabolic-conditions/diabetes [accessed 2026-01-19]
2. Pearson ER, Donnelly LA, Kimber C, et al. Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study. Diabetes 2007 Aug;56(8):2178-2182. [doi: 10.2337/db07-0440] [Medline: 17519421]
3. Uemura F, Okada Y, Torimoto K, Tanaka Y. Enlarged glycemic variability in sulfonylurea-treated well-controlled type 2 diabetics identified using continuous glucose monitoring. Sci Rep 2021 Mar 1;11(1):4875. [doi: 10.1038/s41598-021-83999-z] [Medline: 33649395]
4. Chung WK, Erion K, Florez JC, et al. Precision medicine in diabetes: a consensus report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care 2020 Jul;43(7):1617-1635. [doi: 10.2337/dci20-0022] [Medline: 32561617]
5. Maddaloni E, Bolli GB, Frier BM, et al. C-peptide determination in the diagnosis of type of diabetes and its management: a clinical perspective. Diabetes Obes Metab 2022 Oct;24(10):1912-1926. [doi: 10.1111/dom.14785] [Medline: 35676794]
6. Garg S, Kitchen R, Gupta R, Pearson E. Applications of AI in predicting drug responses for type 2 diabetes. JMIR Diabetes 2025 Mar 27;10:e66831. [doi: 10.2196/66831] [Medline: 40146874]
7. Zou X, Liu Y, Ji L. Review: Machine learning in precision pharmacotherapy of type 2 diabetes—a promising future or a glimpse of hope? Digit Health 2023;9:20552076231203879. [doi: 10.1177/20552076231203879] [Medline: 37786401]
8. Hao Y, Cheng C, Li J, et al. Multimodal integration in health care: development with applications in disease management. J Med Internet Res 2025 Aug 21;27:e76557. [doi: 10.2196/76557] [Medline: 40840463]
9. Ren Q, Xiao D, Han X, et al. Genetic and clinical predictive factors of sulfonylurea failure in patients with type 2 diabetes. Diabetes Technol Ther 2016 Sep;18(9):586-593. [doi: 10.1089/dia.2015.0427] [Medline: 27403931]
10. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. Preprint posted online on May 22, 2017. [doi: 10.48550/arXiv.1705.07874]
11. Hébert HL, Shepherd B, Milburn K, et al. Cohort profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). Int J Epidemiol 2018 Apr 1;47(2):380-381j. [doi: 10.1093/ije/dyx140] [Medline: 29025058]
12. Martono DP, Lub R, Lambers Heerspink HJ, Hak E, Wilffert B, Denig P. Predictors of response in initial users of metformin and sulphonylurea derivatives: a systematic review. Diabet Med 2015 Jul;32(7):853-864. [doi: 10.1111/dme.12688] [Medline: 25582542]
13. Feng C, Wang H, Lu N, Tu XM. Log transformation: application and interpretation in biomedical research. Stat Med 2013 Jan 30;32(2):230-239. [doi: 10.1002/sim.5486] [Medline: 22806695]
14. Cao XH, Stojkovic I, Obradovic Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC Bioinformatics 2016 Sep 9;17(1):359. [doi: 10.1186/s12859-016-1236-x] [Medline: 27612635]
15. Van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw 2011 Dec;45(3):67. [doi: 10.18637/jss.v045.i03]
16. Rubin DB. Multiple Imputation for Nonresponse in Surveys: John Wiley & Sons; 1987.
17. O'brien RM. A caution regarding rules of thumb for variance inflation factors. Qual Quant 2007 Sep 11;41(5):673-690. [doi: 10.1007/s11135-006-9018-6]
18. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition: Springer Series in Statistics; 2009. [doi: 10.1007/978-0-387-84858-7]
19. García CB, Salmerón R, García C, García J. Residualization: justification, properties and application. J Appl Stat 2020;47(11):1990-2010. [doi: 10.1080/02664763.2019.1701638] [Medline: 35707574]
20. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]

21. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/BF00994018]
22. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. arXiv. Preprint posted online on Mar 9, 2016. [doi: 10.48550/arXiv.1603.02754]
23. Rumelhart D, Hinton G, Williams R. Learning representations by back-propagating errors. Nature New Biol 2002;323:533-536. [doi: 10.1038/323533a0]
24. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat 2010 Mar;4(1):266-298. [doi: 10.1214/09-AOAS285]
25. Kapelner A, Bleich J. bartMachine: machine learning with Bayesian additive regression trees. J Stat Soft 2016 Apr;70(4):40. [doi: 10.18637/jss.v070.i04]
26. Cardoso P, Young KG, Hopkins R, et al. Evaluating prediction of short-term tolerability of five type 2 diabetes drug classes using routine clinical features: UK population-based study. Diabetes Obes Metab 2025 Aug;27(8):4320-4329. [doi: 10.1111/dom.16470] [Medline: 40375735]
27. Cardoso P, Young KG, Nair ATN, et al. Phenotype-based targeted treatment of SGLT2 inhibitors and GLP-1 receptor agonists in type 2 diabetes. Diabetologia 2024 May;67(5):822-836. [doi: 10.1007/s00125-024-06099-3] [Medline: 38388753]
28. Drug response prediction. GitHub. 2025. URL: https://github.com/Pearson-Research-Lab/Drug-Response-Prediction [accessed 2026-01-13]
29. Willmott C, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 2005;30:79-82. [doi: 10.3354/cr030079]
30. Freedman D. Statistical Models: Theory and Practice: Cambridge University Press; 2005. [doi: 10.1017/CBO9781139165495]
31. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006 Jun;27(8):861-874. [doi: 10.1016/j.patrec.2005.10.010]
32. Powers DMW. Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. Mach Learn Technol 2008 Jan [FREE Full text]
33. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008 Nov;28(5):26. [doi: 10.18637/jss.v028.i05]
34. Jones AG, Lonergan M, Henley WE, Pearson ER, Hattersley AT, Shields BM. Should studies of diabetes treatment stratification correct for baseline HbA1c? PLoS ONE 2016;11(4):e0152428. [doi: 10.1371/journal.pone.0152428] [Medline: 27050911]
35. Dennis JM, Henley WE, Weedon MN, et al. Sex and BMI alter the benefits and risks of sulfonylureas and thiazolidinediones in type 2 diabetes: a framework for evaluating stratification using routine clinical and individual trial data. Diabetes Care 2018 Sep;41(9):1844-1853. [doi: 10.2337/dc18-0344] [Medline: 30072404]
36. Dennis JM, Young KG, Cardoso P, et al. A five-drug class model using routinely available clinical features to optimise prescribing in type 2 diabetes: a prediction model development and validation study. Lancet 2025 Mar 1;405(10480):701-714. [doi: 10.1016/S0140-6736(24)02617-5] [Medline: 40020703]

Abbreviations

AUC: area under the receiver operating characteristic curve
BART: Bayesian additive regression trees
C-peptide: connecting peptide
GoDARTS: Genetics of Diabetes Audit and Research in Tayside Scotland
HbA$_{1c}$: hemoglobin A$_{1c}$
HDL: high-density lipoprotein
ML: machine learning
NHS: National Health Service
NN: neural network
RMSE: root mean square error
SHAP: Shapley Additive Explanations
VIF: variance inflation factor
XGBoost: extreme gradient boosting

Edited by R Zhao; submitted 19.Aug.2025; peer-reviewed by J Low, M Yagcioglu, P Taiwo; revised version received 08.Dec.2025; accepted 11.Dec.2025; published 06.Feb.2026.

Please cite as:
Garg S, Kitchen R, Gupta R, Trucco E, Pearson E
Predictors of Glycemic Response to Sulfonylurea Therapy in Type 2 Diabetes Over 12 Months: Comparative Analysis of Linear Regression and Machine Learning Models
JMIR Diabetes 2026;11:e82635
URL: https://diabetes.jmir.org/2026/1/e82635
doi:10.2196/82635

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org