

AI-assisted analysis of public data to identify diabetes stakeholder priorities in Singapore: A proof-of-concept study

Mabel Qi He Leow^{*1,2,3} PhD, Sonia Gupta^{*4} PGDM, Denise Lin^{*4} MSc, Ngiap Chuan Tan^{1,2,3} FAMS

Dear Editor,

The exponential growth of publicly available healthcare data from open-access journals, news articles, and online forums has created an extensive repository for data mining and knowledge generation.¹ The wealth of information presents unprecedented opportunities for understanding stakeholder needs within healthcare systems. Data from online forums are increasingly recognised as a legitimate qualitative research methodology, offering access to large amounts of spontaneous and unfiltered perspectives that all stakeholders can share naturally in public domains.²

Artificial intelligence (AI) technologies in healthcare represent one of the most promising frontiers for improving health services and health outcomes. Natural language processing and machine learning in AI offer promising capabilities to systematically collect and analyse large-scale public domain data.³

Diabetes is a common chronic condition that affects approximately 400,000 people in Singapore, with a prevalence of 8.5%.⁴ The condition's management involves multiple stakeholders: patients requiring daily self-management, caregivers providing support, healthcare providers (HCPs) delivering care, and governments formulating policies and managing resources.⁵ This complexity makes diabetes particularly suitable for demonstrating how AI can capture diverse stakeholder perspectives from public domain data.

The study was conducted in 3 main phases using Accenture's UberSight platform (Accenture SG Services Pte Ltd, Singapore), a downstream analytics and forecasting layer that ingests licensed or otherwise permitted datasets, applies analytical processing, and generates forecast-oriented insights. It does not independently crawl, scrape, or extract data from external platforms. The 3 phases were: (1) data collection and preparation, (2) holistic AI application using generative AI and machine learning models, and (3) visualisation and insights development. "Data collection" here refers to the study methodology for sourcing and preparing input datasets, rather than a direct platform capability.

In phase 1, researchers collected 132,708 artefacts over 12 months (between September 2023 and August 2024) from diverse diabetes-related sources, including academic databases (e.g. PubMed) and social media platforms (e.g. YouTube, Reddit, X, Instagram, TikTok). The proportions of the data sources were: PubMed (67.6%), X (8.0%), online news (12.6%), blogs (4.2%), forums/message boards (3.4%), YouTube (1.1%), newspapers (0.8%), magazines (0.7%), Reddit (0.7%), TikTok (0.6%), Instagram (0.3%), and television/radio (0.03%). Keywords and topic clusters specific to Singapore's diabetes landscape were curated, with human supervision ensuring contextual validity and relevance. The list of search terms is in Supplementary Table S1.

In phase 2, the sources obtained were mapped systematically into a robust master data model using a defined data ingestion schema that tagged content types, date of publication, and source credibility, so that AI analysis could be conducted effectively. The data ingestion schema served as a standardised blueprint for how data was collected, structured, and stored. Each incoming data point was automatically tagged with key attributes such as content type (e.g. article, report, image, transcript) to differentiate modalities; date of publication to enable trend analysis; and source credibility to assess reliability and prioritise authoritative information. AI algorithms were assigned weightages based on relevance to diabetes topics. The study was conducted under applicable usage-rights and data governance restrictions. Access was limited to authorised personnel and use restricted to the approved analytical purpose. Redistribution of vendor data was prohibited, and project-specific data were deleted upon conclusion of the use case. Accenture's responsible AI governance framework ensured compliance with data privacy requirements, including Singapore's Personal Data Protection Act and the European Union's General Data Protection Regulation. Only anonymised, aggregated data were processed.

The Annals is an open access journal, allowing non-commercial use under CC BY-NC-SA 4.0.

¹ SingHealth Polyclinics, Singapore

² SingHealth-Duke NUS Family Medicine Academic Clinical Programme, Singapore

³ Primary Care Research Institute, SingHealth Polyclinics, Singapore

⁴ Accenture Pte Ltd, Singapore

* Joint first authors

Correspondence: Dr Mabel Leow, SingHealth Polyclinics, 167 Jalan Bukit Merah Connection One (Tower 5), #15-10, Singapore 150167.

Email: mabelleowqjhe@yahoo.com

Accepted: 7 September 2026

Published online first: 11 September 2026

The comprehensive discovery corpus incorporated contextual understanding, linking content to social, cultural, policy, and behavioural frameworks. For instance, “insulin” references were analysed beyond mere mentions to understand specific contexts regarding cost, accessibility, or medical effects. The system addressed missing values, outliers, and format inconsistencies, using natural language processing models and metadata extraction techniques to standardise diverse multimodal data formats (e.g. PDF, HTML, .mp4, .jpg, .txt, video, text, images) into unified structures optimised for generative AI models.

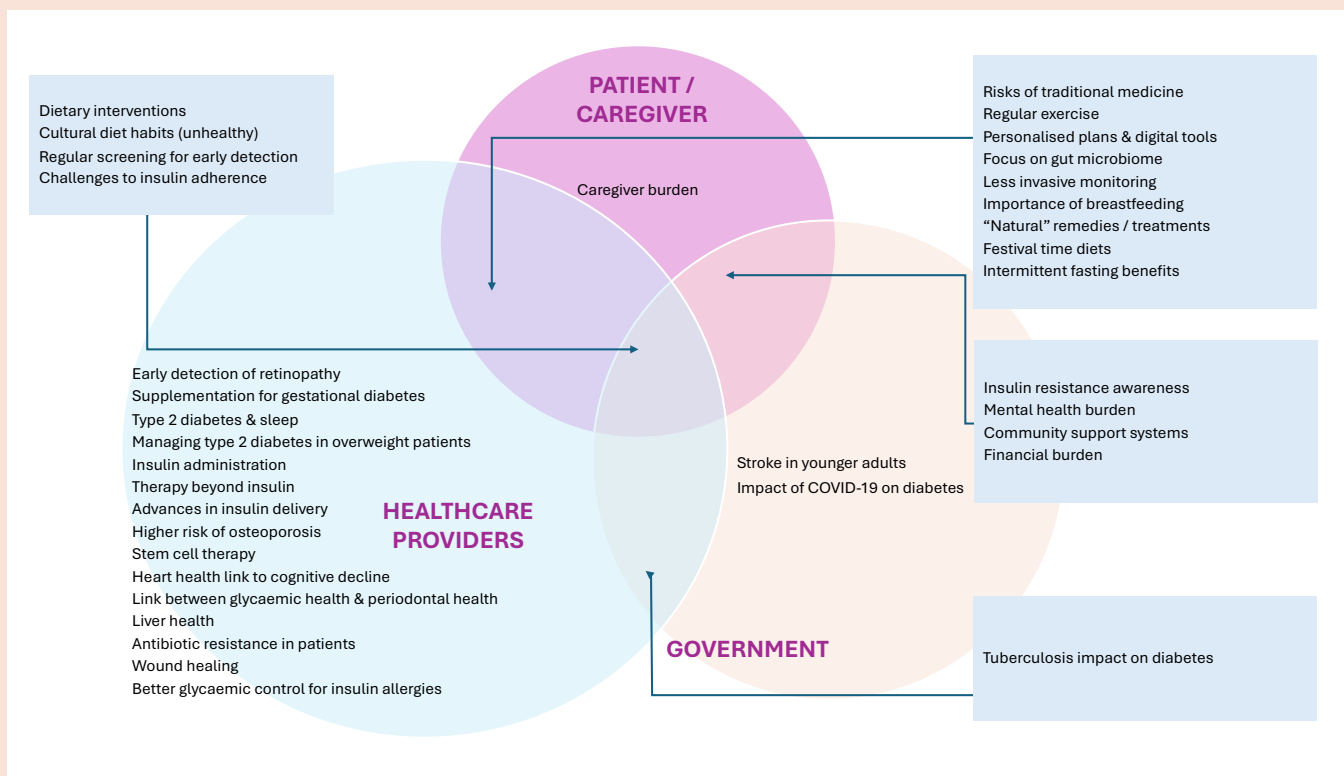
Using large language model-assisted topic generation, the researchers extracted unique topics from the master dataset to identify microtrends. Generative AI algorithms grouped semantically similar discussions. UberSight’s Trend Intensity model used its Power and Momentum dimensions to distil information into 52 unsupervised themes through AI-led topic modelling. The Power vector included topic volume, referrer volume, and sentiment, while Momentum measured pace, recency, and directional trajectory over time. Themes scoring high on both vectors emerged as significant trends. Healthcare specialists, clinicians, and researchers manually reduced the 52 themes to 36 supervised themes of

importance to Singapore’s diabetes management. These were analysed across 3 stakeholder groups: patients and caregivers, HCPs, and government.

Phase 3 involved data designers creating interactive presentations and dashboards. Themes most salient to patients and caregivers were presented through detailed vignettes and segment-specific implications. Sensitivity analysis validated final model outputs through topic coherence scores (interquartile range 0.25–0.69, median 0.35), followed by manual cross-checks to refine clustering assignments. When the sensitivity analysis was run by individual data type, a slightly higher median coherence of 0.49 was found. The low coherence scores can be attributed to the dataset comprising heterogeneous content types, including biomedical literature, transcripts, and online discussions. As topic coherence metrics are sensitive to linguistic variation across content types, the observed distribution of scores is consistent with the dataset’s composition.

From 130,000 initial artefacts, the system identified 278 microtrends across the 3 stakeholder groups, consolidated into 52 unsupervised themes, and finally refined to 36 supervised themes by clinical experts (Fig. 1). Three key topics emerged as common priorities across all 3 stakeholders: dietary interventions and

Fig. 1. Categorisation of themes.



cultural diet habits, regular screening, and insulin adherence challenges. Patients/caregivers and HCPs intersected on diabetes management methods including traditional medicine, exercise, intermittent fasting, postpartum breastfeeding, and digital tools. Patients/caregivers and government overlapped in support services, particularly psychological, social, and financial support.

Patients and caregivers were most concerned about the caregiver burden of providing care for a person with diabetes. HCPs were focused on understanding how diabetes interacted with other health conditions, such as sleep, heart, and periodontal health, and developing new therapies for patients with diabetes. The government was investigating increased stroke risk in younger adults with diabetes, and the impact of COVID-19 on diabetes.

This study demonstrates AI-driven methodologies' feasibility for analysing large-scale public domain healthcare data. The approach successfully processed 132,708 artefacts, providing comprehensive insights into Singapore's diabetes care landscape. The Power and Momentum vectors enabled identification of both current priorities and emerging concerns, supporting proactive healthcare planning. Traditionally, data from public domains were collected and analysed manually, which limited both the number of forums that could be accessed and the findings obtained.⁶ However, a systematic review between 2023 and 2024 found 130 articles that used AI for qualitative data analysis, highlighting the rising trend of using AI in this field.⁷

The distinct theme clusters that emerged for each stakeholder group reveal where perspectives align and where gaps may exist. The difference between clinical priorities and public awareness represents an opportunity for targeted intervention. For example, caregiver burden remains a major concern for caregivers and patients with diabetes, but has been overlooked by HCPs and the government, suggesting insufficient translation of research into practice.^{8,9}

The findings also highlighted HCPs' large focus on comorbidities and therapeutic developments, which are crucial for long-term health outcomes. However, these were less frequently discussed amongst patients and caregivers, possibly reflecting gaps in the group's understanding of diabetes. This is supported by a systematic review which found that diabetes knowledge among type 2 diabetes patients in Southeast Asia was unsatisfactory, particularly amongst older patients with lower education levels and poor glycaemic control.¹⁰ This suggests a need for improved patient education and communication about the broader health implications of diabetes.

Although the methodology effectively processed large volumes of data, it might not capture the depth of individual experiences that qualitative research methods provide. Therefore, this approach should complement rather than replace traditional in-depth qualitative interview methodologies. While the authors maintained diversity in the data sources, they cannot guarantee that this translates to demographic diversity amongst contributors. Given that data were collected from public forums where participants remained anonymous, demographic representation could not be verified.

In summary, this AI-driven framework represents a significant advancement in identifying the priorities of various stakeholders in diabetes management. The data have potential for use in healthcare planning and policy development, with future applications extending to other chronic conditions and identifying cross-disease patterns and systemic healthcare challenges.

Supplementary material

Table S1. List of search terms used.

Acknowledgements

The authors would like to thank Eileen Koh and Ng Ding Xuan from SingHealth Polyclinics and Damian Frankcom, Tejas Krishna, Kelly Quek, and Lee Naylor from Accenture for their contributions to this article.

Ethics statement

Not applicable, as no patient data were used in this study.

Declaration

The authors declare that they have no affiliations or financial involvement with any commercial organisation with a direct financial interest in the subject or materials discussed in the manuscript.

Keywords: access to information, artificial intelligence, data analysis, data collection, diabetes mellitus, public health informatics

REFERENCES

1. Miller BH. Open source intelligence (OSINT): An oxymoron? *Int J Intelligence Counter Intelligence* 2018;31:702-19.
2. Shaw EK. The use of online discussion forums and communities for health research. *Fam Prac* 2020;37:574-7.
3. Wang M, Sushil M, Miao BY, et al. Bottom-up and top-down paradigms of artificial intelligence research approaches to healthcare data science using growing real-world big data. *J Am Med Inform Assoc* 2023;30:1323-32.
4. Diabetes Singapore. Facts and figures. Updated 2025. www.diabetes.org.sg/diabetes-in-singapore/. Accessed 1 September 2025.
5. Ow Yong LM, Koe LW. War on diabetes in Singapore: a policy analysis. *Health Res Policy Syst* 2021;19:15.

6. Bergen N, Labonté R. "Everything Is Perfect, and We Have No Problems": Detecting and Limiting Social Desirability Bias in Qualitative Research. *Qual Health Res* 2020;30:783-92.
7. Cook DA, Ginsburg S, Sawatsky AP, et al. Artificial Intelligence to Support Qualitative Data Analysis: Promises, Approaches, Pitfalls. *Acad Med* 2025;100:1134-49.
8. Morgan DL. Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *Int J Qual Methods* 2023;22:16094069231211248.
9. Lau JH, Abidin E, Jeyagurunathan A, et al. The association between caregiver burden, distress, psychiatric morbidity and healthcare utilization among persons with dementia in Singapore. *BMC Geriatr* 2021;21:67.
10. Lee VY, Seah WY, Kang AW, et al. Managing multiple chronic conditions in Singapore - Exploring the perspectives and experiences of family caregivers of patients with diabetes and end stage renal disease on haemodialysis. *Psychol Health* 2016;31:1220-36.