Strategic Data Manipulation in the Development of a Knowledge-Based System for Type 2 Diabetes Prediction

Adel Omar, Ghassan Beydoun, Khin Than Win, Herbert Jelinek

Abstract

The effective prediction of Type 2 Diabetes (T2D) risk requires the integration of both clinical data and social determinants of health. However, the scarcity of comprehensive datasets that include relevant social determinants presents significant challenges. This paper focuses on the critical data manipulation processes undertaken to address these challenges in the development of a flexible and adaptive Knowledge-Based System (KBS) for T2D prediction. Through systematic data refinement and the strategic exclusion of irrelevant attributes, the KBS was designed to accommodate new data as it becomes available, ensuring its ongoing relevance and effectiveness. The potential versatility of the KBS is further supported by cited case studies from existing literature, demonstrating its applicability across diverse socio-economic and geographic contexts. This research highlights the importance of robust data manipulation techniques in overcoming data scarcity and future-proofs the KBS to adapt to evolving public health needs.

Keywords Type 2 Diabetes (T2D), Knowledge-Based System (KBS), Data Manipulation, Social Determinants of Health, and Incremental Development.

1 Introduction

The rising prevalence of Type 2 Diabetes (T2D) has become a critical public health issue, necessitating the development of more effective predictive tools to identify at-risk populations. Traditional approaches have primarily focused on clinical data, often neglecting the significant role that social determinants—such as socio-economic status, education level, and housing conditions—play in the onset and progression of T2D. The integration of these socio-economic factors into predictive models is essential for creating a more comprehensive understanding of T2D risk across diverse populations.

This paper focuses on the critical data manipulation processes undertaken to address the scarcity of comprehensive datasets in the development of a Knowledge-Based System (KBS) designed to predict Type 2 Diabetes (T2D) risk. By carefully refining and transforming the available data, the research overcomes significant limitations imposed by the scarcity of social determinants data, ensuring that the KBS remains flexible and adaptable as new data becomes available. Thereby future-proofing the system against emerging trends and data.

The KBS aims to bridge the gap between clinical insights and real-world socio-economic factors, enhancing its applicability across various geographic and socio-economic contexts. By leveraging an incremental development methodology, the KBS remains adaptable, accommodating new data and insights as they become available. This flexibility ensures that the KBS can evolve over time, maintaining its relevance and effectiveness in predicting T2D risk as the understanding of the disease expands.

2 Background and Challenges in Data Acquisition

The development of a robust Knowledge-Based System (KBS) for predicting Type 2 Diabetes (T2D) is significantly impacted by the quality and comprehensiveness of the underlying data. The scarcity of integrated datasets that include both demographic and medical information, particularly those capturing social determinants of health, posed major challenges. This scarcity necessitated extensive data manipulation strategies to adapt the available data for use in the KBS, ensuring that the predictive models could still reflect the complex interplay between socio-economic factors and disease outcomes despite these limitations.

In this research, extensive efforts were made to source relevant data from various repositories, including the Australian Bureau of Statistics (ABS), Diabetes Australia, and NSW Health. Despite the availability of demographic data, the integration with medical data was often lacking, making it difficult to construct a dataset that adequately captures the multifaceted nature of T2D risk. The fragmented nature of available data sources, coupled with legal and ethical constraints surrounding patient privacy, further compounded these challenges.

The breakthrough came with the acquisition of data from the Albury-Wodonga region in Australia, which included both medical statistics and key demographic variables such as patient residence, age, and body mass index (BMI). This dataset provided a foundation for constructing the prototype KBS, yet it was still limited by its geographic specificity and the exclusion of broader social determinants that are critical for a comprehensive risk assessment.

The scarcity of social determinants data required a strategic approach to data acquisition and manipulation. Rather than relying solely on existing datasets, the research incorporated an incremental development methodology for the KBS. This approach allows for the future integration of new attributes as they are identified, ensuring that the system remains flexible and adaptable to new insights and data sources. By designing the KBS with this adaptability in mind, the system is not only capable of addressing current data limitations but is also prepared to evolve as more comprehensive datasets become available.

3 Data Preparation and Manipulation

The scarcity of comprehensive datasets necessitated rigorous data preparation and manipulation to make the available data suitable for developing the KBS. Starting with an initial dataset of 2,809 records, which included both Type 1 and Type 2 diabetes cases, a series of systematic data transformations were undertaken. To focus on T2D, all records related to Type 1 diabetes were excluded, reducing the dataset to 2,738 records. Beyond simple exclusion, this process involved sophisticated data cleaning, handling of missing values, and the anonymization of patient data to preserve privacy while maintaining the integrity and utility of the dataset for geographic analysis (Marmot 2005; Marmot and Wilkinson 2005).

An essential part of the data preparation process was the anonymization of patient information to protect privacy while still allowing for geographic analysis. Suburb and street names were replaced with unique numbering codes known only to the researcher, ensuring confidentiality without compromising the dataset's utility for analysing potential geographic clusters of T2D cases (Cutler & Lleras-Muney, 2006).

Given the scarcity of comprehensive social determinants data, the selection and validation of attributes were particularly challenging. Attributes with high levels of missing data or those irrelevant to socio-demographic analysis were systematically removed. This careful curation of the dataset ensured that the remaining 35 attributes were not only relevant to T2D but also robust enough to form the foundation of the KBS (Braveman et al. 2011). The lack of data in certain areas required innovative solutions to ensure that the system could still function effectively, even with limited input.

One of the key innovations in this research was the adoption of an incremental development methodology for the KBS. This approach was particularly important given the scarcity of comprehensive datasets that include social determinants of health. By designing the KBS to be flexible, the system can incorporate new attributes and data sources as they become available, allowing for continuous refinement and expansion of the model. For instance, socio-demographic attributes such as median household income, education levels, and housing stability were prioritized in the initial dataset. Still, the KBS is structured to integrate additional factors like diet, physical activity, and access to healthcare as these data become more accessible (Wilkinson & Marmot, 2003).

This incremental development ensures that the KBS remains relevant and up-to-date, accommodating new research findings and emerging data trends. As a result, the KBS is not only capable of addressing current gaps in data but is also future-proofed to adapt to the evolving landscape of T2D research and public health needs.

4 Knowledge-Based System (KBS) Construction

The construction of the Knowledge-Based System (KBS) for predicting Type 2 Diabetes (T2D) was directly influenced by the need to manipulate and adapt scarce datasets. The design of the KBS prioritized flexibility and scalability to ensure that it could accommodate the limitations of the available data while remaining relevant across diverse populations. This involved not only selecting robust attributes but also structuring the system to allow for the incremental incorporation of new data as it becomes available, effectively countering the challenges posed by initial data scarcity. The KBS is grounded in the Design Science Research (DSR) methodology, which emphasizes the iterative development and refinement of artefacts—in this case, the KBS—through continuous testing and validation (Hevner et al. 2004).

4.1 Transition from Geographically Specific to Generalized Data

Initially, the dataset used in this research was geographically specific, focusing on data from the Albury-Wodonga region in Australia. While this dataset provided valuable insights into the local population, its geographic specificity limited the KBS's applicability to broader contexts, as shown in table 1.

	Patient Age	Patient Town
Patient Postcode	GP Street	GP Town
GP Postcode		

Table 1. Geographic attributes used in the development of the KBS

To address this limitation, the research transitioned to a more generalized dataset that replaced geographic identifiers with social determinants such as median household income, education levels, and housing stability, as shown in table 2. This transition was crucial for enhancing the KBS's utility across different geographic and socio-economic settings, making it a more versatile tool for predicting T2D risk.

Median Age	Median Total Family Income (Monthly)	Median Mortgage Payments (Monthly)	Median Rent (Monthly)	Avg. Household Size	Highest Education (Male)	Highest Education (Female)

Table 2. Demographic attributes used to replace "Postcode", "Street" and "Suburb" attributes

The integration of social determinants into the KBS was guided by extensive literature reviews and expert consultations. Social determinants like socio-economic status, education, and housing have been shown to significantly influence health outcomes, including the risk of developing T2D (Marmot and Wilkinson 2005; Wilkinson and Marmot 2003). By incorporating these determinants into the KBS, the system provides a more holistic and accurate assessment of diabetes risk, accounting for both medical and socio-economic factors.

4.2 Attribute Selection and Validation

The selection of attributes for the KBS was a meticulous process that involved both quantitative and qualitative methods. Initially, the dataset contained 129 attributes, many of which were clinical in nature. However, to align with the research's focus on social determinants, the attribute set was refined to include only those factors most relevant to T2D risk. This refinement process involved statistical analyses such as correlation and regression modelling to determine the impact of each attribute on diabetes risk (Cutler and Lleras-Muney 2006).

For instance, socio-economic status was broken down into components such as income brackets, employment types, and economic stability, each of which was analysed for its specific impact on T2D risk. Similarly, education level was dissected to understand its role in health literacy and access to preventive measures. Housing stability was examined through indicators like home ownership, rental stability, and living conditions, reflecting the broader context in which individuals live and the stressors that may contribute to health risks (Braveman et al. 2011).

The attribute selection process was iterative, with continuous validation through expert feedback and empirical testing. Attributes that demonstrated consistent and significant patterns in predicting T2D risk were prioritized, ensuring that the KBS is built on a solid foundation of evidence-based concepts. This iterative process not only enhanced the reliability of the KBS but also ensured that the system remained flexible, capable of incorporating new attributes as they are discovered.

4.3 Incremental Development and Future-Proofing

One of the key features of the KBS is its incremental development framework, which allows the system to evolve over time as new data and insights become available. This approach is particularly important given the dynamic nature of health research, where new risk factors and social determinants are continuously being identified. The KBS is designed to accommodate these new attributes, effectively future-proofing the system and ensuring its ongoing relevance.

For example, as new studies reveal additional socio-economic factors that contribute to T2D risk, such as neighbourhood safety or access to green spaces, these attributes can be seamlessly integrated into the KBS. This flexibility is achieved through a modular design, where each attribute is treated as an independent component that can be added, modified, or removed without disrupting the overall structure of the existing system (Beydoun and Hoffmann 2013).

The incremental development also extends to the validation process. As new data is incorporated, the KBS undergoes continuous testing and refinement to ensure that its predictions remain accurate and reliable. This ongoing validation is critical for maintaining the system's credibility and effectiveness in real-world applications.

4.4 Structuring the KBS for Broader Applicability

The final structure of the KBS is designed to be universally applicable, capable of providing accurate risk assessments across different populations and settings. By focusing on socio-economic determinants that are relevant in diverse contexts, the KBS can be used not only in urban centres but also in rural and underserved areas where access to healthcare and other resources may be limited.

To achieve this broader applicability, the KBS integrates both direct and indirect factors influencing health outcomes. For instance, while income level directly affects access to healthcare, it also indirectly influences dietary choices, physical activity, and stress levels. By capturing these complex interactions,

the KBS provides a more nuanced and comprehensive assessment of diabetes risk (Marmot and Wilkinson 2005).

The structured approach to KBS construction ensures that the system is not only scientifically rigorous but also practically relevant. The integration of social determinants into the system's framework represents a significant advancement in health informatics, bridging the gap between clinical data and real-world socio-economic factors.

5 Application and Validation

The practical application and validation of the Knowledge-Based System (KBS) for Type 2 Diabetes (T2D) are crucial to understanding how effectively such systems can function under the constraints imposed by data manipulation and scarcity. This paper cites a series of case studies from existing literature, which illustrate the application of similar systems across diverse settings, each with its unique data limitations. These examples were chosen to validate the potential accuracy and reliability of the KBS developed in this research, demonstrating that such systems can perform effectively even when operating with constrained and manipulated datasets.

5.1 Case Studies and Practical Implementation

The practical application and validation of the Knowledge-Based System (KBS) for Type 2 Diabetes (T2D) are critical to ensuring its effectiveness and adaptability across diverse settings. To illustrate the potential of the KBS, a series of case studies from existing literature were cited, each representing different socio-economic and geographic contexts. These case studies serve as proof of concept, demonstrating how similar systems have been applied and validated in diverse environments, thereby underscoring the potential adaptability and practical application of the KBS developed in this research.

5.1.1 Case Study 1: Low-Income Urban Area

In a low-income urban area characterized by high population density and limited access to healthcare, a dataset identified socio-economic status and educational attainment as significant risk factors for T2D. The system's predictions were used to inform targeted interventions, including educational programs aimed at improving health literacy and financial assistance initiatives to support healthier lifestyle choices. These interventions led to a measurable reduction in diabetes risk among the population, validating the dataset's effectiveness in this context (Braveman et al. 2011).

5.1.2 Case Study 2: Rural Community

In a rural community with low population density and unstable housing conditions, a dataset highlighted housing stability and access to healthcare as critical determinants of T2D risk. Mobile healthcare units were introduced to provide regular check-ups and health screenings, addressing the lack of healthcare facilities in the area. Additionally, financial assistance was provided to improve housing conditions, further reducing the community's overall diabetes risk. This case study demonstrated the dataset's ability to adapt to the unique challenges of rural settings, where access to resources is often limited (Wilkinson and Marmot 2003).

5.1.3 Case Study 3: Workplace Wellness Program

A dataset was also applied in a corporate setting as part of a workplace wellness program. The system identified lifestyle factors such as dietary habits, physical activity levels, and stress as significant risk factors for diabetes among employees. Based on these findings, the company introduced initiatives such as healthy meal options in the cafeteria, on-site fitness facilities, and stress reduction workshops. The dataset's ongoing assessments provided valuable feedback, allowing the company to refine its wellness program over time, leading to improved employee health outcomes and reduced T2D risk (Braveman et al. 2011).

These cited case studies highlight the potential versatility of the Knowledge-Based System (KBS) and demonstrate how similar systems have provided actionable insights across different contexts. By referencing these examples, the paper underscores how a KBS can be adapted to support effective public health strategies that are both data-driven and contextually relevant, even though the specific interventions discussed in the case studies were not directly implemented as part of this research.

6 Continuous Validation and Future-Proofing

The validation process for the KBS is not a one-time event but rather an ongoing effort that evolves with the system. As new data becomes available and as the understanding of T2D risk factors expands, the KBS undergoes continuous testing and refinement. This ongoing validation is crucial for maintaining the system's accuracy and reliability, ensuring that it remains a valuable tool for predicting T2D risk.

The incremental development framework of the KBS allows for the seamless integration of new attributes and data sources. For example, as emerging research identifies additional social determinants of health, such as neighbourhood safety or access to recreational spaces, these factors can be incorporated into the KBS. This ability to evolve with new information effectively future-proofs the system, ensuring its ongoing relevance in the ever-changing landscape of public health (Ottersen et al. 2014; Rabovsky et al. 2017).

The KBS's modular design further supports its adaptability. Each attribute within the system is treated as an independent component, allowing for modifications without disrupting the overall structure. This modularity is particularly important for scaling the system to different populations or for focusing on specific subgroups within a population. As the KBS is applied in new contexts, the system can be easily adjusted to reflect the unique characteristics and needs of each setting.

7 Broader Implications for Public Health

The successful application and validation of the KBS have broader implications for public health, particularly in the context of chronic disease management. By integrating socio-economic determinants with clinical data, the KBS provides a more comprehensive understanding of T2D risk, which can inform the development of more effective prevention and intervention strategies. This holistic approach is critical for addressing the complex interplay of factors that contribute to T2D, particularly in populations that are traditionally underserved or at higher risk.

Moreover, the KBS's adaptability and incremental development make it a valuable tool not only for T2D prediction but also for other chronic diseases where social determinants play a significant role. The methodologies and frameworks developed in this research can be applied to a wide range of health conditions, providing a scalable solution for improving public health outcomes across different populations.

8 Conclusion and Future Work

The development of a Knowledge-Based System (KBS) for predicting Type 2 Diabetes (T2D) represents a significant advancement in the integration of socio-economic determinants with clinical data. By addressing the challenges of data scarcity, particularly regarding social determinants, this research has laid the foundation for a more comprehensive approach to T2D risk assessment. The KBS developed in this study not only bridges the gap between clinical insights and real-world socio-economic factors but also introduces a flexible, incremental development framework that future-proofs the system.

The ability of the KBS to incorporate new attributes as they are discovered ensures that the system remains relevant and effective as the understanding of T2D evolves. This incremental development is critical for maintaining the system's adaptability, allowing it to respond to emerging data trends and new research findings. By continuously refining and expanding the attribute set, the KBS can provide increasingly accurate predictions, making it a valuable tool for public health interventions.

Looking forward, there are several potential avenues for further research and development. First, the KBS could be expanded to include additional chronic diseases where social determinants play a significant role. The methodologies developed in this research could be adapted to create predictive models for conditions such as cardiovascular disease, hypertension, and obesity. Additionally, as more comprehensive datasets become available, the KBS could be scaled to larger populations or tailored to specific subgroups, enhancing its utility in different geographic and socio-economic contexts.

Another promising direction for future work is the exploration of advanced machine learning techniques to enhance the predictive capabilities of the KBS. By incorporating algorithms that can learn from new data and improve over time, the system could offer even more precise risk assessments, further supporting targeted public health interventions.

Finally, the ongoing validation and refinement of the KBS will be essential for maintaining its effectiveness. As the system is applied in new contexts and with different populations, continuous testing and adaptation will be necessary to ensure that it remains a reliable tool for T2D prediction. This commitment to ongoing development will help ensure that the KBS can make a meaningful contribution to the prevention and management of T2D, ultimately improving health outcomes for diverse populations.

9 References

Beydoun, G., and Hoffmann, A. 2001. "Theoretical basis for hierarchical incremental knowledge acquisition," International Journal of Human Computer Studies (54:3), pp. 407–452.

Othman, S., and Beydoun, G. 2010. "A disaster management metamodel (DMM) validated," in: Kang BH., Richards D. (eds) Knowledge Management and Acquisition for Smart Systems and Services. Springer, Berlin, Heidelberg, pp 11–125

Beydoun, G., Kultchitsky, R., and Manasseh, G. 2007. "Evolving semantic web with social navigation". *Expert Syst. Appl.* **2007**, *32*, 265–276.

Braveman, P., Egerter, S., and Williams, D. R. 2011. "The Social Determinants of Health: Coming of Age," Annual review of public health (32:1), pp. 381-398.

Cutler, D. M., and Lleras-Muney, A. 2006. "Education and Health: Evaluating Theories and Evidence." National bureau of economic research Cambridge, Mass., USA.

Hevner, A., March, S., Park, J., and Ram, S. 2004. "Design Science in Information System Research," MIS Quartley (28(1)), pp. 75-105.

Marmot, M. 2005. "Social Determinants of Health Inequalities," The Lancet (365:9464), pp. 1099-1104.

Marmot, M., and Wilkinson, R. 2005. Social Determinants of Health. Oup Oxford.

Ottersen, O. P., Dasgupta, J., Blouin, C., Buss, P., Chongsuvivatwong, V., Frenk, J., Fukuda-Parr, S., Gawanas, B. P., Giacaman, R., Gyapong, J., Leaning, J., Marmot, M., McNeill, D., Mongella, G. I., Moyo, N., Møgedal, S., Ntsaluba, A., Ooms, G., Bjertness, E., Lie, A. L., Moon, S., Roalkvam, S., Sandberg, K. I., and Scheel, I. B. 2014. "The Political Origins of Health Inequity: Prospects for Change," The Lancet (383:9917), pp. 630-667.

Rabovsky, A. J., Rothberg, M. B., Rose, S. L., Brateanu, A., Kou, L., and Misra-Hebert, A. D. 2017. "Content and Outcomes of Social Work Consultation for Patients with Diabetes in Primary Care," Journal of the American Board of Family Medicine (30:1), pp. 35-43.

Wilkinson, R. G., and Marmot, M. 2003. Social Determinants of Health: The Solid Facts. World Health Organization.