Cultivating Expertise: Unravelling Type 2 Diabetes Associations through Incremental Knowledge-Based System Development

Submission type: Research-in-progress.

Adel Omar
School of Systems, Management and Leadership
University of Technology Sydney
Sydney, Australia
Email: adel.omar@student.uts.edu.au

Ghassan Beydoun
School of Systems, Management and Leadership
University of Technology Sydney
Sydney, Australia
Email: ghassan.beydoun@uts.edu.au

Khin Win
School of Computing and Information Technology
Wollongong University
Wollongong, Australia
Email: win@uow.edu.au

Nagesh Shukla
Department of Business Strategy and Innovation
Griffith University
Griffith, Australia
Email: n.shukla@griffith.edu.au

Herbert Jelinek
Biomedical Engineering
Khalifa University of Science and Technology
Abu Dhabi, UAE
Email: Herbert.jelinek@ku.ac.ae

Hector Elias
Wollongong University College
Wollongong University
Wollongong, Australia
Email: hector@uow.edu.au

Abstract

Type 2 diabetes is a chronic disease that is caused by a combination of genetic and environmental factors, with social determinants playing a significant role. These social determinants include but not limited to factors such as education level, occupation, family history, ethnicity and place of residence etc.

The prevalence of Type 2 diabetes (T2D) necessitates inventive management strategies. This study explores the potential of utilizing Ripple Down Rules (RDR) in comparison with Machine Learning (ML) approaches for the incremental development of a Knowledge-Based System (KBS). The aim is to create a robust Decision Support System (DSS) for effective T2D management, focusing on the dynamic influence of evolving social determinants.

The research outcomes substantiate the viability of this approach. The KBS, fortified by RDRs, demonstrated notable performance metrics. Specifically, the system achieved an impressive accuracy rate of 90.2%, accompanied by a specificity of 96.9%. The sensitivity, crucial for identifying potential T2D cases, reached 73.6%, indicating the system's proficiency in recognizing diverse instances.

Keywords: Social Determinants, Type 2 Diabetes (T2D, Decision Support Systems (DSS), Knowledge-Based Systems (KBS), Ripple Down Rules (RDR), Machine Language (ML)

1 Introduction

Type 2 Diabetes (T2D) presents a complex global health challenge, exerting substantial strain on healthcare infrastructures and affecting millions worldwide. Beyond its fundamental biological framework, the prevalence and progression of T2D are intrinsically intertwined with an array of intricate social determinants, necessitating an interdisciplinary approach to its management.

The ubiquitous impact of T2D reverberates across the globe, underscoring the urgency of addressing its profound health implications. This chronic ailment, shaped by both genetic predisposition and environmental factors, is notably influenced by social determinants. These determinants encapsulate multifaceted aspects spanning socio-economic conditions, geographic variances, and lifestyle disparities, thus demanding health interventions that holistically encompass this intricate interplay.

The global surge in T2D is undeniable; with an estimated 366 million affected in 2011, this figure is projected to escalate to 552 million by 2030 (Sim et al. 2017). Illustratively, Australia witnesses a staggering economic burden due to diabetes, with an annual expenditure of approximately $14.6 billion (Lee et al. 2013). This mounting concern, evidenced by an increasing allocation of healthcare resources, underscores the exigency for refined strategies in T2D management.

The essence of our research lies in the incremental development of a KBS using RDRs (Omar et al. 2019). By structuring a comprehensive knowledge repository and applying the adaptable framework of RDR (Compton and Kang 2021), our aim is to design a DSS that accurately navigates the intricacies of T2D management. A notable emphasis is placed on integrating the ever-changing dimensions of social determinants (Chen et al. 2020), recognizing their pivotal role in shaping the disease trajectory (Unwin et al. 2010).

It is pertinent to acknowledge that this paper follows a continuum of research that debuted at the 2022 Pacific Asia Conference on Information Systems, titled "The incremental Development of a Diabetes Decision Support System using Ripple Down Rules" (Omar et al. 2022). In this paper, we delve further into our findings, building upon the insights gained since the conference.

1.1 The Socio-biological Landscape of Type 2 Diabetes

T2D transcends a mere medical diagnosis, embedding within larger socio-cultural contexts. Pivotal research by Hill, Nielsen, and Fox (2013) unveiled the 'sociobiological cycle of diabetes', positing those determinants such as economic status, educational attainment, and ethnicity mould health behaviours, subsequently shaping physiological outcomes, thereby perpetuating the disease cycle. This intricate nexus demands a nuanced KBS attentive to these determinants. Figure 1 indicates their proposed cycle of sociobiological cycle of diabetes.

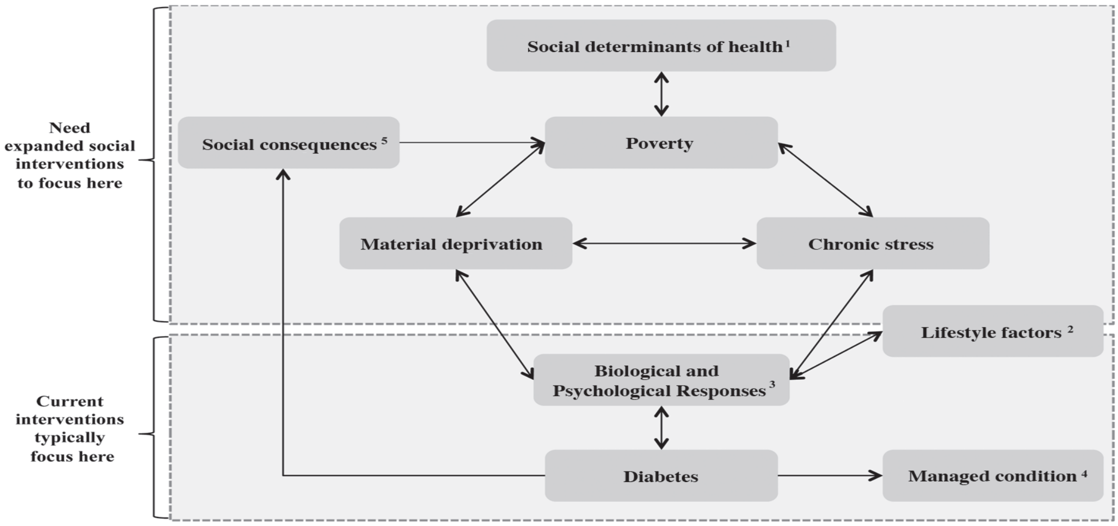

Figure 1. The sociobiological cycle of diabetes. Source: Hill, Nielsen & Fox (2013)

1.2 Knowledge-Based Systems (KBS) in T2D Management

Rooted in comprehensive data repositories, KBS deploy intricate decision-making paradigms by harnessing domain-specific knowledge. Given their adeptness at deciphering complex multivariable relationships, they present a promising tool for T2D interventions. It is worth noting that the application of KBS in healthcare management is not unprecedented, as demonstrated in the works of Boutcher et al. (2022), Al Khamisi et al. (2018), and various other scholarly contributions. However, the distinctiveness of this study lies in the progressive evolution of the KBS facilitated through the employment of Ripple Down Rules (RDRs).

1.3 Incorporating Ripple Down Rules (RDR) in T2D Management

Distinct from traditional rule-based systems, RDR deploys a hierarchical knowledge representation. It thrives in environments with incomplete information, necessitating minimal data (Compton and Kang 2021), making it apt for handling the multifaceted, often underreported, social determinants related to T2D. This makes it ideal for this research as has been confirmed by NSW Health and other organisations, a there is not enough data collected about social determinants that may have a correlation to T2D (Taylor 2019). A similar response was obtained from the head of the diabetes department at Liverpool Hospital (Wong 2016), NSW Australia.

1.4 Advantages of RDR Over Conventional Machine Learning

While machine learning holds transformative potential, its efficacy often hinges on extensive data, a criterion challenging to fulfil with intricate social determinants. Conversely, RDR's adaptive, incremental knowledge base construction makes it adept at handling emerging data on T2D's social determinants. Notably, RDRs offer the advantage of affording subject matter experts (SME) greater flexibility in infusing their personal expertise and experiential insights into the system, thereby enhancing the accuracy of the Knowledge-Based System (KBS) (Compton and Kang 2021). That is, the SME examines the knowledge during the knowledge acquisition phase (Beydoun and Hoffmann 2013). Furthermore, RDRs demonstrate a capacity to elucidate outlier instances that might be overlooked by conventional machine learning techniques.

1.5 Forecasting T2D Onset Using KBS

Harnessing the insights from Omar et al. (2019), it's discernible that the confluence of social and technical dimensions is pivotal in T2D management. An RDR-integrated KBS can illuminate potential T2D hotspots, facilitating pre-emptive interventions. This prospective ability can engender health policies targeting high-risk demographics, ensuring resource-efficient, and culturally nuanced interventions.

2 Discussion on Findings

In accordance with Omar et al. (2022), the preliminary stages of this research involved the segmentation of the dataset into three distinct subsets. This process encompassed the random selection of 500 cases to compose the Training dataset. The KB developed using these 500 records was then used on the the 1200 cases constituting the Production dataset (Omar et al. 2022). The remaining 1038 records (Development Dataset) were not used in this experiment and were left aside for future work. The primary intent of this division was to facilitate an iterative refinement of the Knowledge Base (KB) through successive stages of experimentation.

Interestingly, the Development dataset, initially designated for the purpose of KB enhancement, eventually emerged as a significant testament to the robustness of the KB itself. Notably, the KB's accuracy, as determined through meticulous analysis, demonstrated a level of proficiency that obviated the need for further augmentation before its application to the Production dataset (Omar et al. 2022).

The research outcomes outlined in Omar et al. (2022) shed light on the empirical results obtained through experimentation. As depicted in Table 1, the presented metrics encapsulate the accuracy, specificity, and sensitivity attained for both the Training and Production datasets. The experiment yielded a notable accuracy of 90.2% for the Training dataset, complemented by a specificity of 96.9% and a sensitivity of 73.6%. Correspondingly, the Production dataset demonstrated a commendable accuracy rate of 84.5%, accompanied by a specificity of 92.3% and a sensitivity of 46% (Omar et al. 2022).

	Accuracy	Specificity	Sensitivity
Training Dataset	90.2%	96.9%	73.6
Production Dataset	84.5%	92.3%	46%

Table 1. Results obtained during experiment (Source: Omar et. al (2022))

The notable disparity in test results between the "Training" dataset, indicating a sensitivity of 73.6%, and the "Production" dataset, where the sensitivity drops to 46%, underscores the presence of intricate factors influencing Type 2 Diabetes (T2D) outcomes. These factors encompass a range of variables, such as dietary patterns, cultural contexts, and familial backgrounds, among others. It is worth emphasizing that the absence of comprehensive data collection (Taylor 2019) and (Wong 2016) pertaining to these multifaceted determinants may contribute to the observed variations in sensitivity, highlighting the complexity of T2D management within a real-world context.

Furthermore, the contextual backdrop of the dataset, originating from the Albury-Wodonga region on the New South Wales (NSW) – Victoria (VIC) border of Australia, remains pivotal. As established in Omar et al. (2022), the inherent specificity of the collected social determinants to this geographical location forms a foundation that will significantly influence the ensuing phases of this research. This localized insight underscores the essentiality of contextual nuances when extrapolating findings to broader implications and applications.

3 Comparison with Machine Language (ML)

In the pursuit of a comprehensive Decision Support System (DSS) for the management of Type 2 Diabetes (T2D), a fundamental stage of research revolves around the exploration and evaluation of contrasting methodologies. As a continuation of the research, the current study endeavours to present a comparative analysis between the development of a Knowledge-Based System (KBS) using Ripple Down Rules (RDR) and Weka's J48 algorithm, utilizing the live dataset extracted from the Albury-Wodonga region. This examination is underpinned by the intent to discern the nuanced advantages offered by each technique in terms of accuracy, specificity, and sensitivity.

Upon obtaining significant results through the deployment of RDRs, a consequential phase materialized involving the replication of experiments within the realm of Machine Learning (ML) - specifically employing Weka's J48 as the chosen algorithmic framework for comparison. The selection of Weka's J48 was influenced by its prominence as a decision tree classifier, well-suited for the classification tasks inherent to the management of T2D. A prominent aspect of this comparison was the modification of the dataset to conform to Weka's input requirements.

It is noteworthy to mention that prior to running the data through Weka's J48, a fair bit of data manipulation was initially required. That is, further modifying the file. Modifications included, saving as a csv file as opposed to Excel format, removing 's from any street and/or suburb name (as Weka would not read these and hence couldn't read the file) and converting the "Target" outcomes to a "Yes" or a "No". Weka mistook a lower-case letter as a completely different conclusion, it thought that it made another classification category. This additional work would therefore add more time for the operator to develop the KBS.

Once the above-mentioned modifications were made the "Production" dataset was then run through Weka's J48. Appendix A shows Weka's J48 output.

Upon execution, Weka's J48 generated a pruned tree containing 143 rules, achieving an accuracy rate of 94.9%. However, when contrasted with RDRs, the analysis of the pruned tree structure indicated a more extensive rule set, prompting a pivotal inquiry: Does the incremental enhancement in accuracy justify the significant expansion of rule complexity?

The full results from running Weka's J48 are shown in table 2 below.

	No Rules	Accuracy	Specificity	Sensitivity
Production Dataset	143	87.6%%	98.8%	29.7%

Table 2. Results obtained by Weka's J48

Table 3 below shows the statistics obtained in developing T2D KBS using RDRs as compared to Weka's J48.

	No Rules	Accuracy	Specificity	Sensitivity
RDRs	79	84.5%	92.3%	46%
Weka's J48	143	87.6%	98.8%	29.7%

Table 3. Comparison of results obtained by RDRs as opposed to Weka's J48

3.1 Discussion on Comparison between RDRs and Machine Language

It is worth mentioning that while various Machine Learning (ML) algorithms could have been employed for this comparison, for the sake of simplicity and speed, Weka's J48 was chosen using default parameters. This choice allowed for a straightforward comparison and evaluation of the two approaches.

A crucial distinction surfaces when examining the efficiency of the rule generation process. Notably, RDRs exhibited remarkable parsimony, requiring the construction of 79 rules to attain an accuracy level of 84.5%, with specificity and sensitivity values of 92.3% and 46% respectively. In stark contrast, Weka's J48 yielded a larger rule set but with a relatively marginal improvement in accuracy, emphasizing the intricate trade-off between accuracy and complexity in decision tree construction.

It is imperative to underscore that the findings underline the pragmatic utility of RDRs as an effective means of rule acquisition in the context of T2D management. The incremental increase in accuracy achieved through Weka's J48 comes at the cost of substantial rule proliferation, potentially rendering the resultant model less interpretable. In essence, the central query surfaces: Can the quest for enhanced accuracy be balanced with the need for comprehensible and pragmatic Knowledge Base Support?

A noteworthy advantage of RDRs becomes evident when the role of subject matter experts is considered. The integration of a domain expert's nuanced understanding and experiential insights into the KBS development process is facilitated seamlessly by RDRs (Beydoun and Hoffmann 2013). This infusion of expertise elevates the accuracy of rule development, as subject matter experts can inject context-specific insights into rule formulation. This symbiotic relationship between technology and expert knowledge substantiates RDRs' capacity to effectively encapsulate the intricacies of T2D management, particularly within the realm of social determinants. The role of the subject matter expert in formulating the KBS during the knowledge acquisition phase is illustrated in Figure 3.

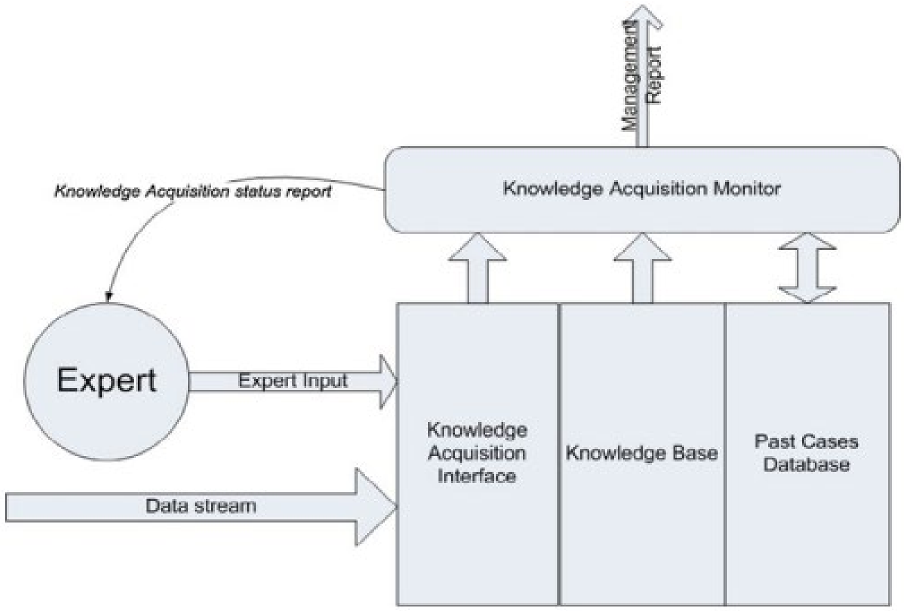

Figure 3. A dynamic monitoring process supplements the knowledge acquisition process. This process deploys a statistical monitor that reports on the quality of the last interaction to the expert

and the progress of the whole knowledge acquisition process to the management overseeing the knowledge acquisition project in the organisation. Source: Beydoun & Hoffmann, (2013)

Moreover, RDRs offer a distinct advantage in terms of rule generation economy. This facet is paramount in healthcare contexts, where interpretability and transparency are essential. RDRs' parsimonious rule generation process contributes to comprehensible models that resonate with clinicians, thereby enhancing their acceptance and utilization. This inherent comprehensibility complements the essence of a DSS, where actionable insights gleaned from the system's recommendations are pivotal.

Ultimately, this comparison accentuates the merits of RDRs as a pragmatic avenue for constructing a KBS tailored for T2D management. The succinctness of rule development coupled with commendable accuracy substantiates the assertion that the simplicity and comprehensibility of RDRs hold intrinsic value in the realm of medical Knowledge Base Systems.

In conclusion, the comparison between RDRs and Weka's J48 for the development of a KBS tailored for T2D management underscores the compelling attributes of RDRs. The scarcity of comprehensive information on T2D's intricate correlations with social determinants amplifies the advantages of RDRs' incremental development process, expert integration, and rule economy. These qualities position RDRs as a prudent choice, fostering the symbiotic relationship between data-driven insights and domain expertise required for effective T2D management.

4 Future Work

In accordance with the information provided in appendix A, a substantial portion of the rules within the Knowledge-Based System (KBS) is composed of street and suburb names. These characteristic attributes a geographic specificity to the KBS, making it tailored to the area from which the data was collected. As previously elaborated, this specific data was gathered from the Albury – Wodonga region situated along the New South Wales and Victoria border on the East Coast of Australia. However, the portability of the KBS to different regions would result in its diminished utility.

In light of this consideration, the subsequent phase of research endeavours to enhance the applicability of the KBS by substituting street and suburb names with pertinent social determinants corresponding to those localities. This modification aims to render the KBS adaptable to diverse geographic contexts. To attain this objective, the experimental procedures delineated earlier in this paper will be replicated, albeit with the utilization of social determinants in lieu of street and suburb names. Subsequently, the outcomes of these experiments will be re-evaluated and contrasted with the findings detailed in this study.

Consequently, this future work seeks to augment the versatility of the KBS, enabling its utility across various geographical regions as long as the relevant social determinants for those regions are discerned. This refinement is anticipated to enhance the KBS's practicality in aiding medical professionals and researchers in the effective management and understanding of Type 2 diabetes.

5 Conclusion

In conclusion, this research endeavour has delved into the development of a Knowledge-Based System (KBS) aimed at unravelling the intricate associations between Type 2 diabetes (T2D) and an array of social determinants. The exploration of this dynamic interface has been conducted within the purview of the Design Science Research (DSR) methodology, encompassing a comprehensive four-phase approach. The proposed KBS serves as a pivotal Decision Support System (DSS) in the domain of T2D management, bridging the gap between medical expertise and data-driven insights. This study has demonstrated the significance of integrating evolving social determinants into the KBS, augmenting its utility and relevance in healthcare contexts.

The examination of KBS development methodologies has underscored the notable advantages offered by Ripple Down Rules (RDRs) in comparison to Machine Learning (ML) algorithms. RDRs emerge as an astute choice, primarily due to their facilitation of subject matter experts' active involvement in the development process. The seamless integration of domain experts' nuanced insights enriches the rule formulation process, enhancing the accuracy and context-specificity of the KBS. Furthermore, the inherent interpretability of RDRs engenders a higher degree of transparency, a pivotal attribute in healthcare decision-making processes. This is in stark contrast to the often "black-box" nature of ML algorithms, which poses challenges in understanding the rationale behind their decisions.

The incremental acquisition of knowledge within the KBS, supported by the phased approach, bolsters its credibility and efficacy. Moreover, the proposed KBS framework presents a pathway toward extending its applicability beyond geographic constraints by incorporating social determinants instead of location-specific attributes. This visionary transformation aspires to equip the KBS with the versatility to cater to diverse healthcare contexts and regions, enhancing its practicality on a global scale.

The synergistic interplay between technology and expert knowledge, epitomized by the utilization of RDRs, substantiates the KBS's capacity to holistically encapsulate the intricate dynamics of T2D management. The choice of RDRs over ML algorithms underscores a strategic shift towards a more human-centric and interpretable approach, thereby fostering a harmonious confluence between advanced technology and informed medical practice. As this research strives to illuminate the path toward a more comprehensive understanding of T2D and its determinants, the judicious choice of RDRs attests to its commitment to advancing the realm of healthcare informatics with a discerning balance of innovation and expertise.

6 References

Beydoun, G., and Hoffmann, A. 2013. "Dynamic Evaluation of the Development Process of Knowledge-Based Information Systems," *Knowledge and Information Systems* (35:1), pp. 233-247.

Beydoun, G., Hoffmann, A., Breis, J. T. F., Bejar, R.M., Valencia-Garcia, R. 2005. "Cooperative Modelling Evaluated". International Journal of Cooperative Information Systems 14 (1), 45-71.

Beydoun, G. & Hoffmann, A. 1997, "Acquisition of search knowledge", Knowledge Acquisition, Modeling and Management, Springer, pp. 1-16.

Chen, M., Tan, X., and Padman, R. 2020. "Social Determinants of Health in Electronic Health Records and Their Impact on Analysis and Risk Prediction: A Systematic Review," *Journal of the American Medical Informatics Association* (27:11), pp. 1764-1773.

Compton, P., and Kang, B. H. 2021. *Ripple-Down Rules: The Alternative to Machine Learning*. CRC Press.

Lee, C., Cohort, R., and Magliano, D. 2013. "The Cost of Diabetes in Adults in Australia," *Diabetes res Clin*), pp. 385 - 390.

Omar, A., Beydoun, G., Win, K., Shukla, N., and Baker, G. 2019. "Socio-Technical Perspective on Managing Type Ii Diabetes,").

Omar, A., Beydoun, G., Win, K. T., and Jelinek, H. 2022. "The Incremental Development of a Diabetes 2 Knowledge Base System Using Ripple Down Rules," in: *Pacific Asia Conference on Information Systems, July 5-9 2022*. Taipei/Sydney Virtual Conference.

Sim, L. L. W., Ban, K. H. K., Tan, T. W., Sethi, S. K., and Loh, T. P. 2017. "Development of a Clinical Decision Support System for Diabetes Care: A Pilot Study," *PLoS ONE* (12:2).

Taylor, L. 2019. "What Kind of Social Determinants Data Is Available on Diabetes Patients?." Sydney, Australia: Epidemiology and Biostatistics | Centre for Epidemiology and Evidence, NSW Ministry of Health.

Unwin, N., Whiting, D., and Roglic, G. 2010. "Social Determinants of Diabetes and Challenges of Prevention," *The Lancet* (375:9733), pp. 2204-2205.

Wong, V. 2016. "Head of Diabetes (Liverpool Hospital)," A. Omar (ed.).