Detecting Patient Information Conflicts through Conflict Reasoning in Knowledge Graphs

Enhancing Accuracy and Reliability in a Diabetes Support System

Jochem van Paridon[1]

Supervisor(s): Prof. Catholijn Jonker[1]**, J.D. Top, MSc**[2]

[1]**EEMCS, Delft University of Technology, The Netherlands**
[2]**Bernoulli Institute, University of Groningen, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Jochem van Paridon
Final project course: CSE3000 Research Project
Thesis committee: Prof. Catholijn Jonker, J.D. Top, MSc, Dr. Avishek Anand

Abstract

Lifestyle management systems aim to provide personalized health guidance by interpreting patient's self-reported data. However, these systems often overlook the temporal consistency of behavioral patterns, risking inaccurate or misleading recommendations. To address this, we present an adapted version of a diabetes lifestyle management system that integrates a conflict detection component into its knowledge graph (KG) architecture. This addition enables the system to identify temporal inconsistencies in patient data based on formally defined rules. We developed handcrafted constraints to capture clinically meaningful relationships, such as the expected timing between physical activity and glucose responses. A tunable threshold parameter, ε, is introduced to account for minor fluctuations and variability in activity intensity, enhancing the physiological realism of constraint evaluations. Handcrafted rules were favored over mined constraints due to the limited availability of real-world data, though future work may incorporate mining techniques when larger datasets become accessible. The system was tested on synthetic case studies that simulate typical lifestyle scenarios, demonstrating its potential to flag inconsistencies and improve the reliability of temporal reasoning. These results contribute to the design of more trustworthy, context-aware decision support tools for personalized health management.

1 Introduction

Automated medical systems have the potential to increase the quality of healthcare by offering accessible and rapid guidance to patients based on self-reported information [1]. Some of these systems utilize knowledge graphs (KGs) to structure and analyze medical data, enabling complex reasoning about symptoms and conditions [2, 3]. Their success depends on the consistency and accuracy of patient's self-reported data, which can be affected by inconsistencies stemming from misunderstanding, deception or non-adherence to medical guidance. These discrepancies introduce conflicts within KGs, potentially leading to misdiagnoses or inappropriate treatment suggestions [4, 5].

KGs in healthcare play a crucial role in structuring medical data, enabling advanced reasoning and improving diagnostic processes [3, 6]. While significant efforts have been made to store patient information within KGs [7], there still remains research to be done in detecting and resolving inconsistencies in KGs with patient's self-reported data. Temporal conflicts, semantic ambiguities, and factual contradictions can impact the accuracy of automated healthcare systems, leading to potential misinterpretations of symptoms and conditions.

The CHIP Modular System [8], developed by The Netherlands Organization for Applied Scientific Research (TNO) and the Hybrid Intelligence Project (HI) [8, 9], provides a structured framework for medical advice generation by processing patient input through specialized modules that extract relevant medical details. However, existing implementations do not yet incorporate mechanisms for identifying contradictions across multiple patient sessions. Strengthening conflict resolution strategies within such systems is essential for improving their reliability and ensuring more accurate healthcare recommendations.

The current research focuses on enhancing conflict resolution within knowledge graphs to improve the accuracy of lifestyle management systems, specifically the CHIP system. It examines temporal conflicts that can exist throughout the system [10, 11]. However, it is important to note that other conflicts exist as well such as: semantic conflicts [12, 13] and fact-validation or mutex conflicts (mutual exclusion) [14–16]. Addressing these challenges will refine medical knowledge graphs, to ensure more consistent and reliable lifestyle guidance.

Temporal conflicts arise when a patient's self-reported data changes over time, creating inconsistencies in their medical history. Such conflicts can impact the accuracy of automated healthcare systems by leading to outdated or misleading recommendations. For instance, a patient using a continuous glucose monitor (CGM) may initially report stable glucose levels but later experience frequent fluctuations, contradicting earlier entries. Similarly, a user of an automated insulin delivery system might alter their dosage patterns without informing the system, leading to mismatches between historical records and real-time health status. To mitigate these conflicts, knowledge graphs must incorporate longitudinal data tracking, enabling comparisons across multiple interactions to identify evolving health trends.

Semantic conflicts occur due to variations in terminology or differences in contextual understanding, leading to misinterpretations in medical assessments. Patients, healthcare providers, and automated systems may use different descriptions for the same condition or treatment, introducing ambiguity into healthcare data. For example, a patient might describe their insulin pump settings as 'low-dose basal', while medical guidelines refer to it as 'continuous subcutaneous insulin infusion', causing discrepancies in treatment plans. Additionally, a user reporting 'sugar spikes', might be referring to postprandial hyperglycemia (a spike in glucose levels after eating a meal), but if the system interprets it incorrectly, medical guidance may be misaligned. A different situation might be that one person describes their workout to the system as 'an intensive workout', while someone else might describe the same workout as 'a simple workout'. Resolving semantic conflicts requires semantic decomposition techniques to standardize input, detect variations in patient descriptions, and ensure accurate data interpretation.

Fact-validation conflicts or mutex (mutual exclusion) conflicts arise when contradictory statements exist within the knowledge graph. These conflicts occur when different

sources or data entries disagree on a direct fact (independent of time), leading to inconsistencies in medical assessments or treatment recommendations. For instance, one entry in the graph may assert that a specific type of CGM provides real-time blood sugar readings, while another entry states that the same device only offers periodic measurements. Similarly, a knowledge graph might contain conflicting records regarding the effectiveness of automated insulin delivery systems, one stating that a certain algorithm optimizes dosage based on glucose trends, while another contradicts it by claiming the algorithm functions on fixed timing alone. Resolving fact-validation conflicts requires structured fact-checking mechanisms, leveraging computational verification methods that cross-check conflicting statements, determine authoritative sources, and refine entries to preserve the integrity and accuracy of the knowledge graph.

Effectively resolving different types of conflicts is crucial to maintaining accurate and reliable medical advice in lifestyle management systems. Addressing these requires structured approaches that integrate longitudinal tracking, semantic standardization, and computational fact-checking to ensure consistency across patient's self-reported data and medical knowledge graphs. Given the impact of these conflicts on medical advice, the research aims to explore the broader question: "How can conflict resolution in knowledge graphs enhance the accuracy and reliability of lifestyle management systems?" This inquiry is further broken down into the following sub-questions:

- What types of conflicts can occur within Knowledge Graphs relevant in lifestyle management systems, and how might these conflicts affect the advice provided by such systems?
- What are the most effective methods to detect different types of conflicts within Knowledge Graphs in lifestyle management systems?
- What are the most effective methods to resolve different types of conflicts within Knowledge Graphs in lifestyle management systems?
- What metrics and validation techniques are used to assess the accuracy and reliability of lifestyle management systems?
- How can the CHIP Modular System be extended such that it detects conflicting information and possibly solves the conflicts within the Knowledge Graphs?
- To what extent does conflict resolution in CHIP's Knowledge Graphs influence the accuracy and reliability of patient advice?

By developing methods to detect temporal conflicts, this research contributes to the advancement of hybrid intelligence in healthcare. The findings will improve the consistency and trustworthiness of lifestyle management systems, ensuring patients receive more reliable guidance. The study also offers insights into refining KG-based reasoning in healthcare applications.

The structure of this paper is organized to systematically explore conflict resolution within knowledge graphs for medical advice systems. Section 2 presents the theoretical background, outlining foundational principles relevant to conflict detection. Section 3 details the methodology, including the background, literature review, and a thorough explanation of the system's design, implementation, and experimentation. Section 4 describes the experimental setup and results, covering the specific experimental settings, overall performance, and conflict detection analysis. Section 5 discusses the findings. Section 6 outlines identified limitations and areas for improvement. Section 7 concludes the study with key insights and proposes directions for future research to advance conflict resolution strategies within medical knowledge graphs. Finally, Section 8 addresses responsible research, ensuring ethical considerations are thoroughly integrated into the study.

2 Theoretical Background

Before delving into the methodology (see Section 3), it is important to first establish a solid theoretical background by examining the following concepts in more detail.

Knowledge Graphs (KGs) are structured representations of information that store data in the form of triples: (*s, p, o*). The subject *s* and the object *o* represent entities, while the predicate *p* defines their relationship. These graphs enable structured knowledge representation, supporting applications such as semantic search, recommendation systems, and automated reasoning [17]. However, they also introduce challenges such as data inconsistencies, scalability limitations, and semantic ambiguities [18]. Ensuring reliability in KGs requires addressing these drawbacks through efficient storage, validation mechanisms, and disambiguation techniques.

Temporal Knowledge Graphs (TKGs) are extensions of traditional knowledge graphs in which each triple is enriched with a temporal dimension, typically represented as a timestamp or a time interval *t*. This allows the graphs to model dynamic knowledge that evolves over time, capturing facts that are only valid during certain periods. Formally, a temporal quad can be denoted as (*s, p, o, t*), where *t* may refer to a point or interval in time. TKGs are especially useful for applications such as event tracking, temporal reasoning, historical analysis, and real-time decision-making in domains like healthcare, finance, and social networks [19] . However, the temporal aspect also introduces unique challenges, such as temporal conflicts, data sparsity, and increased computational complexity in reasoning tasks [19]. As such, effective temporal modeling, conflict detection, and resolution strategies are crucial for maintaining the integrity and usefulness of TKGs.

Building on this, various strategies have been investigated to address temporal conflicts within knowledge graphs. Two prominent approaches include the use of handcrafted

constraint rules and data-driven constraint mining. Handcrafted constraints rely on expert-defined logical structures to enforce temporal coherence between events, often reflecting medically grounded expectations, such as the anticipated timing between physical activity and glucose fluctuations. In contrast, constraint mining leverages large datasets to automatically uncover frequently occurring temporal patterns [20]. While the former offers interpretability and domain alignment, the latter provides adaptability and scalability, especially in data-rich environments.

There are several ways to formulate constraint rules, depending on the specific temporal relationships being modeled. Common types include Precedence, Overlap, Inclusion, Disjointedness, and Mutex [16].

Precedence describes the relationship in which one task must be completed before another begins. Overlap occurs when two intervals share a portion of time, mathematically expressed as $s_i < s_j < f_i < f_j$, where s and f denote start and finish times. Inclusion implies that one time interval is fully contained within another, as in $s_i < s_j < f_j < f_i$. Disjointedness, on the other hand, indicates that two intervals do not intersect at all [16].

Another important type is mutex (mutual exclusion), which refers to logical facts that are inherently contradictory and therefore cannot coexist, independent of any temporal context. In this work, such constraints are also used to model fact-validation conflicts (see Section 1), situations where the presence of one fact logically invalidates another [16]. This concept is referred to under a different name in this context, because it applies irrespective of time.

3 Methodology

The methodology section outlines the structured approach of this research, starting with the background in Section 3.1, which provides context on the CHIP system and its integration of knowledge graphs in automated medical advice. The literature review, in Section 3.2, examines existing research on conflict detection and resolution within knowledge graphs, highlighting gaps and opportunities for improvement. The design, in Section 3.3 presents tailored strategies for identifying temporal conflicts, explaining key methodological choices, including the use of handcrafted constraints over constraint mining. The implementation, in Section 3.4, details how conflict resolution mechanisms are embedded within CHIP, describing the integration of temporal data constraints, the development of new modules, and modifications to system architecture. Finally, the experimentation, in Section 3.5, defines the evaluation framework, specifying the data used and validation techniques used to assess the accuracy and reliability of the proposed conflict resolution methods through testing.

3.1 Background

The CHIP system provides a conversational user interface that allows users to engage in text-based interactions (see Figure 1). When a user inputs a query or medical information, the Text2Triple Module processes the text, extracting relevant information and converting it into structured triples. These triples are then stored in the Patient Knowledge Graph, ensuring patient-specific data is organized effectively.

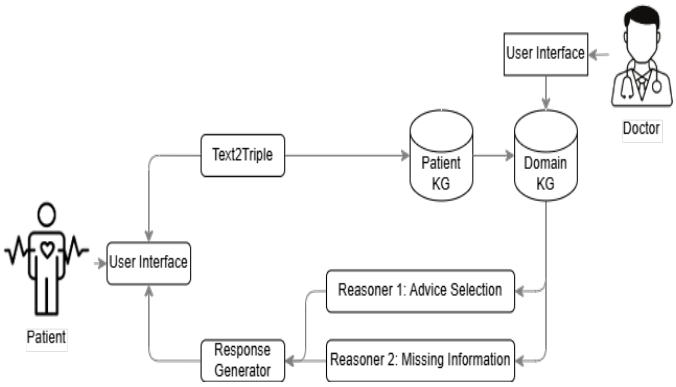

Figure 1: The CHIP system overview illustrates how patient inputs are processed through modular reasoning components and knowledge graphs to generate accurate medical advice.

On the medical side, doctors interact with their own interface, which allows them to input information directly into the Domain Knowledge Graph. This graph stores expert knowledge, clinical guidelines, and medical best practices, complementing the patient-specific data.

Decision-making within CHIP relies on two reasoning modules. The Missing Information Reasoner identifies gaps in patient data and requests additional details when necessary to ensure accurate advice. The Advice Selection Reasoner determines whether sufficient information has been gathered to provide guidance. Once a decision is made, both reasoners communicate with the Response Generator, which formulates a clear and concise message for the user. The generated response is then delivered through the user interface, completing the cycle of interaction.

The Web Ontology Language (OWL) [21] is a widely adopted standard for representing ontologies. OWL provides a formal, structured framework for defining relationships within KGs, allowing for precise reasoning and automated inference. Ontologies defined using OWL enable interoperability by ensuring that different systems share a common vocabulary, reducing ambiguity and inconsistencies in structured data.

CHIP leverages OWL to maintain logical consistency across its Patient and Domain Knowledge Graphs. By using OWL-based reasoning, CHIP ensures that medical data is properly structured, enabling precise entailment of new information based on predefined relationships.

3.2 Literature Review

Google Scholar and Scopus were used as the primary academic databases to identify relevant literature. For sources with a stronger medical and clinical focus, PubMed was additionally consulted. Together, these platforms provide comprehensive coverage of publications related to knowledge graphs (KGs) and healthcare, facilitating a systematic and well-rounded review.

The search terms included 'conflict types in knowledge graphs,' 'KG inconsistencies,' 'ontology conflicts,' and 'semantic reasoning for conflict detection.' The first phase of the literature review focused on identifying different types of conflicts present in KGs. Initial searches targeted fundamental papers discussing conflict classification in KGs [12–14], followed by an analysis of methodologies for detecting them [10, 16, 20, 22]. The prioritization of conflicts was then determined based on their impact on data integrity, reasoning validity, and the computational resources required for resolution. Higher-priority conflicts were those that had the potential to identify the largest group of conflicts in the KG in CHIP.

A second phase of the literature review was executed to explore conflict detection and resolution techniques. Terms such as 'knowledge graph conflict resolution,' 'automated inconsistency handling,' and 'temporal reasoning for conflict detection' guided the process. This phase examined tools and algorithms developed to automatically identify and resolve conflicts within KGs.

The final phase involved evaluating when and how different conflict resolution methods were chosen. The selection criteria included availability of structured data, applicability of reasoning frameworks, and computational feasibility of different approaches. If no sufficient data was available for a resolution technique, alternative inference strategies were considered, ensuring flexibility in handling KG inconsistencies.

3.3 Design

An overview of the modified CHIP system architecture is presented in Figure 2. Compared to the original framework shown in Figure 1, this adapted design incorporates an additional component responsible for conflict detection within the patient KG. This addition enables the system to identify temporal inconsistencies in the data using predefined rules. Specifically, Equations 1–3 define the constraints used to detect temporal conflicts. These serve as examples of how formal logic can be operationalized within lifestyle management systems to ensure better alignment between patient behavior and modeled expectations.

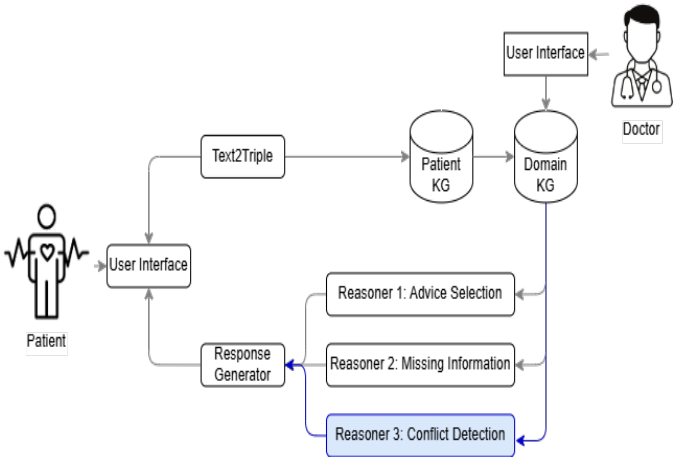

Figure 2: Schematic overview of CHIP system overview illustrates how the system is adapted using a conflict detection reasoner (in blue).

Several methods for detecting temporal conflicts were explored, as outlined in Section 2, to guide the integration of constraints into the system design.

A key design decision was to incorporate temporal information directly into the knowledge graph by associating each triple with a timestamp. This structure enables time-aware reasoning over patient data, forming the foundation for conflict detection as further detailed in Section 2.

Another of the design choices that needed to be made was whether to use handcrafted constraints or constraint mining. Handcrafted constraints were chosen due to the data requirements for effective mining. Constraint mining relies on extensive real-world datasets to identify meaningful patterns and relationships. Without sufficient high-quality data, mined constraints may be incomplete, unreliable, or unrepresentative of real-world interactions. As shown in studies such as [20, 22], successful constraint mining demands a large volume of structured data, which is often unavailable or difficult to obtain in certain domains.

However, constraint mining could yield more optimal results if sufficient data were available [22]. Unlike handcrafted constraints that rely on predefined rules and expert knowledge, mined constraints are derived directly from the data, capturing temporal relationships that emerge from patterns. This approach enhances flexibility and accuracy, making constraint mining a potentially superior method in data-rich environments. If large-scale, well-curated datasets were accessible, constraint mining could provide more precise and context-aware constraints, improving temporal reasoning in KGs.

There are multiple types of temporal constraints used in temporal reasoning [16]. Among these, precedence was specifically chosen to demonstrate causal relationships within the KG. This constraint type ensures that events follow a causal order and interact meaningfully based on

their time intervals.

Precedence constraints [16] enforce a strict sequence, guaranteeing that one fact finishes before another begins. This approach is essential for illustrating causality, ensuring that dependent events unfold in a structured manner. As shown in Figure (to be added), precedence constraints help maintain the logical flow of temporal data, preventing inconsistencies that could arise from disordered sequences.

The end times of the precedence relationship between the two components must be carefully considered, because according to Allen's interval algebra [23], the temporal relation *precedes* requires that the end of the first interval occurs strictly before the start of the second. Moreover, it is also important that there is not a gap that is biologically implausible between the end time of the first interval and the start time of the second. For instance, detecting a glucose spike 24 hours after food consumption becomes irrelevant in establishing a direct causal relationship. To define these expectations formally, a set of temporal constraints was developed, to capture causal patterns such as meal–glucose or activity–glucose relationships. These constraints are used to determine whether observed event sequences fall within biologically meaningful time windows. The implementation of this approach is detailed in Section 3.4.

Constraint 1. *The constraint represents the physiological principle that glucose levels decrease following exercise [24]. This constraint can be broken by the patient, for example when they reported they exercised, but did not see through with it and instead stayed on the couch and ate something.*

$$\begin{aligned} \text{precedes}(t_1, t_2) := \; & (x, \text{glucoseMeasurement}, a, t_1) \\ & (x, \text{activity}, \text{y}, t_2) \\ & (x, \text{glucoseMeasurement}, b, t_2) \\ & a > b + \varepsilon \end{aligned} \tag{1}$$

Constraint 2. *The constraint represents the physiological expectation that after consuming sugar, blood glucose levels should rise [25]. However, this assumption can be violated if, for example, a patient fails to report their food intake while exhibiting an unexplained spike in glucose levels. Such discrepancies highlight the importance of accurate self-reporting and continuous monitoring to ensure proper medical assessment.*

$$\begin{aligned} \text{precedes}(t_1, t_2) := \; & (x, \text{glucoseMeasurement}, a, t_1) \\ & (x, \text{consumption}, \text{y}, t_2) \\ & (x, \text{glucoseMeasurement}, b, t_2) \\ & a + \varepsilon < b \end{aligned} \tag{2}$$

Constraint 3. *The constraint illustrates the principle that insulin intake should lead to a reduction in blood glucose levels [26]. If this expected decline does not occur, possible explanations include a patient misreporting their insulin usage or experiencing complete insulin resistance. Both scenarios pose serious medical risks, requiring close observation and intervention to prevent potential complications. Detecting such inconsistencies is essential for maintaining effective glucose regulation and ensuring appropriate treatment responses.*

$$\begin{aligned} \text{precedes}(t_1, t_2) := \; & (x, \text{glucoseMeasurement}, a, t_1) \\ & (x, \text{treatment}, \text{insuline}, t_2) \\ & (x, \text{glucoseMeasurement}, b, t_2) \\ & a > b + \varepsilon \end{aligned} \tag{3}$$

As seen in Equations 1, 2 and 3, a variable ε is added. This epsilon accounts primarily for measurement errors, or acts as a threshold for the measurements in the constraint.

For example, in Equation 1, glucose levels are expected to decrease following an activity. However, if the equation does not specify the magnitude of the drop, this can result in minor fluctuations, such as from 5.4 mmol/L to 5.39 mmol/L, being treated as a valid drop, which is unlikely to be physiologically meaningful. The inclusion of ε helps to address this ambiguity.

Another justification for the presence of ε is the variability in activity intensity. For instance, revisiting Equation 1, activities can be categorized as high-intensity (e.g., running, swimming) or low-intensity (e.g., walking, yoga). High-intensity activities generally lead to a more pronounced glucose drop compared to low-intensity ones [24]. By tuning ε accordingly, larger for high-intensity and smaller for low-intensity, we can better align the model with physiological expectations.

The value of ε can potentially be optimized in future work through hyperparameter tuning using machine learning techniques, see Section 7.

3.4 Implementation

OWL-Time [27] is an ontology designed to represent a standardized vocabulary to describe the temporal property of resources. It is used for describing ordering relations among time intervals, durations, and positions using different reference systems, such as the Gregorian calendar and Unix-time. By making CHIP use OWL-Time, it can maintain logical consistency in time-based reasoning, supporting applications like event scheduling, historical data tracking, and knowledge inference.

SPARQL [28] is employed as the rule language in our system, leveraging its powerful query capabilities to define, manage, and execute reasoning rules over our ontology. By adopting SPARQL for constraint specification, we ensure seamless integration with existing semantic web standards, which in turn

enables sophisticated semantic querying and inference. This choice not only enhances the expressiveness and granularity of our constraint definitions but also facilitates interoperability and scalability within our KG management framework.
For the management and design of the knowledge graph, we employed Protégé [29]. This is a robust and user-friendly ontology editor. Protégé facilitated the efficient construction and visualization of complex relationships among data entities. Its comprehensive toolset supported effective maintenance and consistency throughout the ontology, thereby enhancing the overall clarity and robustness of our KG.

CHIP has also been adapted to expect continuous input from a continuous glucose monitor. When a patient reports an activity at a specific time, for example, as referenced in constraint 1, the system checks the corresponding glucose measurements to verify whether the constraint holds. This validation process is applied to all constraints.

A continuous glucose monitor was chosen as an additional input to the system (not only user text input), because accurate timestamps are crucial for validating constraints. If there is a delay between the reported activity and the glucose measurement, other factors may have influenced the patient's glucose levels in the meantime. Ensuring that timestamps closely match actual activities helps maintain the integrity of the analysis.

3.5 Experimentation

The evaluation is based on a series of handcrafted scenarios designed to test the system's ability to detect temporal conflicts. Each scenario is embedded within a larger patient knowledge graph to verify that the conflict detection mechanism functions reliably even as the graph scales in complexity. The scenarios simulate synthetic patient timelines composed of structured, timestamped events, such as blood glucose measurements, physical activities, meal intake and insulin intake. For each constraint type (see Section 3.3), one scenario is created, each containing a patient message that intentionally conflicts with corresponding glucose measurements. Upon processing this information, the system is expected to identify the inconsistency and return a conflict message that explains the nature of the detected violation. Subsequently, the scenario includes a corrected patient message with updated information, after which the system should confirm that the data is valid and no longer conflicting.

4 Experimental Setup and Results

The implementation supporting the findings of this study is publicly available on GitHub at https://github.com/JochemvanP/chip-conflict-detection.

This section outlines the experimental performance analysis of CHIP in identifying temporal inconsistencies within patient data. The results highlight the system's ability to enforce patient knowledge through constraint validation.

4.1 Experimental Settings

In our experiment, the tolerance parameter ε is carefully calibrated based on the expected metabolic responses to physical activity. For the first constraint, which differentiates activity intensity, ε is set to 0.14 mmol/L for high-intensity activities and 0.09 mmol/L for low-intensity activities. The second constraint employs an ε value of 2.3 mmol/L, corresponding to the expected rise in glucose levels within one hour. Similarly, for the third constraint, ε is set to 2.0 mmol/L, which reflects the anticipated drop in glucose levels over the same time period.

4.2 Overall Performance

CHIP successfully detected all designed conflicts across the three scenarios, with no false positives on valid records. The system precisely adhered to the boundary margins defined by the parameter ε, ensuring realistic modeling of glucose responses relative to activity intensity. In our experiments, data generated from a variety of patient scenarios, each as closely as possible representing realistic behavioral patterns, was rigorously processed to validate the system's reasoning performance.

The detection performance is exemplified by the results for Constraint 1. As shown in Figure 3, CHIP flagged a conflict when an insufficient glucose drop was observed following a reported high-intensity exercise session. The conflict message clearly indicated the discrepancy between the expected physiological response and the recorded data. This instance is representative of the overall performance, as similar outcomes were observed in all experimental cases.

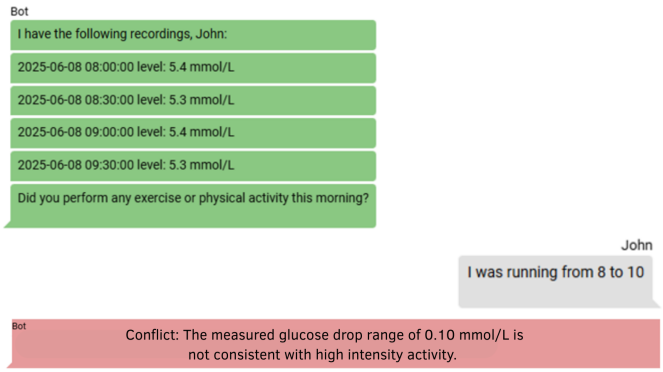

Figure 3: Example scenario with CHIP detecting a conflict due to a violation of Constraint 1.

Moreover, when a corrected input was provided, the system reliably validated the new record as conflict-free. Figure 4 illustrates this updated scenario, where the patient's revised data, now consistent with expected outcomes, was accepted without triggering any conflict.

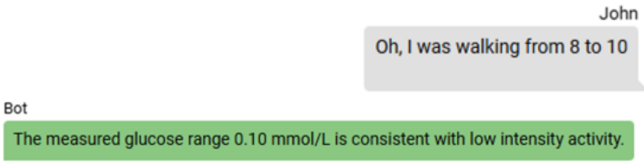

Figure 4: Example scenario with CHIP accepting a scenario conflict with regard to Constraint 1.

4.3 Conflict Detection

CHIP was methodically evaluated by deliberately injecting a single targeted conflict in each scenario. These conflicts accounted for errors such as insufficient glucose declines. This approach enabled a clear assessment of CHIP's reasoning mechanism under controlled, yet realistic conditions.

In every case, even when input values approached the preset ε boundaries, CHIP accurately identified the conflict. The logical rules embedded in the system were able to distinguish between innocuous fluctuations and medical deviations, ensuring that only genuine inconsistencies were flagged. This level of precision was maintained across all targeted tests.

5 Discussion

The experimental results validate that CHIP effectively identifies temporal inconsistencies in patient data. By reliably detecting all injected conflicts without generating misclassifications, the system demonstrates robust symbolic reasoning and confirms the correctness of the constraint-based approach.

CHIP's current design assumes inputs from continuous glucose monitoring devices, which allows for high-frequency, real-time data collection. This shift to automated, continuous measurement enhances the system's ability to generate accurate and timely advice. The process capitalizes on CGM data to accurately capture glucose variability (GV), a critical factor determining the magnitude of physiological responses, such as post-exercise glucose declines or increases in glucose after eating.

The constraint rules applied in this study serve as a proof-of-concept for using rule-based logical frameworks in healthcare. While the current rules perform well within a controlled experimental environment, questions remain about their direct applicability in more complex, real-world scenarios, where variability and noise are significantly higher.

In summary, the results presented herein illustrate the potential of constraint-based reasoning in enhancing the reliability of lifestyle management systems. Although further validation in real-world settings is necessary, especially with diverse populations and more varied data inputs, this work lays a solid foundation for the integration of symbolic logic with advanced data analytics in personalized health monitoring.

6 Limitations

Although the system demonstrates promising results in conflict detection within structured health data, several limitations must be acknowledged. These reflect the current scope of the implementation and identify opportunities for further refinement and scalability. The following subsections highlight key challenges regarding data representativeness, constraint diversity, and parameter adaptability.

Limited Number of Evaluation Scenarios. While the system performs well across the scenarios tested, the number of evaluation scenarios is currently too limited to make general claims about robustness. A broader set of test cases would be necessary to confirm performance consistency across edge cases, complex situations, and different patient profiles.

Use of Handcrafted Data. The experimental data used in this study was synthetically generated and manually tailored to reflect constraint violations. While this allows for targeted testing, it introduces a sampling bias and may not fully capture the unpredictability and noise inherent in real-world patient data. Consequently, performance on actual clinical datasets remains to be validated.

Assignment of the Epsilon Parameter. The epsilon (ε) threshold, used to define tolerances in constraint checks, was manually specified rather than learned or estimated from data. This introduces subjectivity and reduces adaptability. In practical applications, data-driven approaches, such as using machine learning to infer optimal thresholds, would provide more dynamic and patient-specific constraint tuning. Additionally, individual differences in glucose variability (GV) may significantly influence how much a patient's glucose levels fluctuate in response to various triggers such as exercise or meals. Factoring GV into the determination of ε could improve the sensitivity of conflict detection by tailoring expectations to each patient's metabolic profile.

Limited Constraint Coverage. The current implementation includes only a small subset of possible constraints. They all focus now on causal relationships. However, health data is semantically rich, encompassing a wide range of clinical behaviors, physiological trends, and treatment guidelines. Expanding the constraint set to reflect this complexity would significantly enhance the system's expressiveness and practical utility.

Restricted Vocabulary for Constraints. In the first constraint related to activity-induced glucose change, only a limited set of predefined activities (e.g., running, walking, swimming) are supported. Patients may engage in a much broader range of physical tasks, and the system's ability to generalize across unlisted or nuanced activities is currently constrained. Similarly, for the second constraint concerning food consumption, the current implementation supports only a fixed set of meal types and nutrition labels. However, real-world dietary behavior is far more diverse and context-dependent. Incorporating a richer ontology of physical activities and food intake, or integrating external classifiers could substantially improve the adaptability and expressiveness of the constraint model.

7 Conclusions and Future Work

This research demonstrates how conflict detection in knowledge graphs can enhance the accuracy and reliability of lifestyle management systems. Through the successful implementation of temporal constraints, the system was able to validate patient behavior against clinically grounded expectations. The handcrafted scenarios confirmed the correctness of the constraints and showcased the system's robustness in handling structured, timestamped data within a knowledge graph. By embedding constraints into a lifestyle management system, this work improves the accuracy and reliability of health monitoring tools and strengthens patient data validation. Although the evaluation was limited to synthetic data, the controlled design allowed for precise inspection of system behavior and offered valuable insights into constraint-driven reasoning.

Future research could focus on learning optimal values for the ε threshold using machine learning techniques. By analyzing historical patient data and outcome trends, models could adaptively calibrate ε to align with individual physiological variability and context-specific sensitivity. Another promising direction is to explore constraint mining from large-scale lifestyle or clinical datasets. This approach could uncover previously unrecognized temporal patterns and personalize constraint sets, increasing system adaptability in diverse populations. Future work could explore the use of uncertain temporal knowledge graphs [16], to account for imprecise or incomplete temporal information in patient data. This would allow the system to reason probabilistically over temporal relationships, improving flexibility and robustness when exact timestamps are unavailable, noisy, or ambiguous. Integrating uncertainty aware logic could enhance real-world applicability, especially in domains like lifestyle monitoring where data may be self-reported or sporadic.

8 Responsible Research

This research was designed with an emphasis on transparency and methodological clarity; however, several elements warrant reflection regarding the reproducibility and integrity of the work.

One component that needs to be mentioned is the use of AI tools to assist in language and structure. In this case, Microsoft Copilot (an AI language model) was used during the writing and refinement process. A typical prompt used to guide rewriting included: *"Can you improve the clarity and fluency of this academic paragraph without changing the technical content?"* This tool was leveraged primarily for stylistic enhancement and clarification, not for ideation or content creation.

A challenge lies in the assignment of hyperparameters such as the epsilon (ε) threshold, within the constraints. Since ε was manually tuned in this study, different values, particularly in systems with more granular glucose variability (GV) or alternative activity types, could yield different constraint outcomes. This raises the need for careful sensitivity analysis or automated tuning strategies in future work.

In terms of the scope of evaluation, the scenarios presented are intentionally narrow and designed to isolate specific constraint violations. While this aids clarity, it risks limiting generalizability. A different researcher, using alternative but valid scenarios, may encounter different system behaviors depending on how edge cases are represented.

From an integrity standpoint, all modeled constraints have been tailored to the domain of diabetes lifestyle management and are explicitly defined in the paper. The theoretical basis is transparent, and constraints are grounded in cited literature and physiological rationale. No results have been omitted or filtered beyond the deliberate scope of evaluation. Releasing both the engine and scenarios publicly alongside this paper (see Section 4) further supports ethical and reproducible research practices.

Moreover, this project engages in the development and enhancement of CHIP through innovative AI-driven features including adaptive user interfaces, automated decision support, and personalized dialogue systems based on user profiling and behavioral data. Our approach is built upon a strong commitment to ethical integrity and regulatory compliance, placing us in strict alignment with the EU Artificial Intelligence Act (AI Act) [30].

Chapter II, Article 5, Paragraph 1, outlines prohibited practices such as manipulative or exploitative AI, has been a guiding framework in defining our ethical boundaries and design constraints. The systems does not manipulate or exploit the patients in any way. So would the scenario evaluation, see Section 3.5, be done with real patients it would still follow the constraints, thus it respects both legal- and ethical mandates.

Furthermore, as we consider the inclusion of components like AI-based diabetes support, we recognize that such health-related applications are likely to be classified as high-risk. This classification is rooted in several factors:

- Critical Impact on Health: AI-powered diabetes support systems handle sensitive health data and directly influence treatment decisions. Erroneous recommendations, such as misinterpreted blood sugar trends or incorrect insulin dosing, could lead to severe adverse health outcomes.
- Safety Risks: In healthcare, the margin for error is extremely low. Inaccuracies may result in life-threatening situations, underscoring the need for rigorous validation and precise risk management.
- Regulatory Oversight: Given their potential impact, health-related AI systems fall under stricter regulatory oversight. This necessitates comprehensive risk management strategies, extensive testing, and real-time monitoring to ensure both patient safety and data protection.

Acknowledgements

I would like to express my sincere gratitude to the Netherlands Organisation for Applied Scientific Research (TNO) and to the team behind the Hybrid Intelligence Project (HI) for their invaluable support and insights throughout this project. With sincere appreciation towards Rineke Verbrugge and Harmen de Weerd for their expertise and generous provision, those have been crucial contributions to this work.

References

[1] I. C. Layadi, "Healthcare automation: A systematic literature review," in *Human-Automation Interaction: Manufacturing, Services and User Experience*, V. G. Duffy, M. Lehto, Y. Yih, and R. W. Proctor, Eds. Cham: Springer International Publishing, 2023, pp. 167–183, ISBN: 978-3-031-10780-1. DOI: 10.1007/978-3-031-10780-1_9.

[2] J. Wu, W. Deng, X. Li, S. Liu, T. Mi, Y. Peng, Z. Xu, Y. Liu, H. Cho, C.-I. Choi, Y. Cao, H. Ren, X. Li, X. Li, and Y. Zhou, *Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs*, 2025. arXiv: 2504.00993 [cs.CL].

[3] E. Rajabi and S. Kafaie, "Knowledge graphs and explainable ai in healthcare," *Information*, vol. 13, no. 10, 2022, ISSN: 2078-2489. DOI: 10.3390/info13100459.

[4] N. Britten, F. A. Stevenson, C. A. Barry, N. Barber, and C. P. Bradley, "Misunderstandings in prescribing decisions in general practice: Qualitative study," *BMJ*, vol. 320, no. 7233, pp. 484–488, 2000. DOI: 10.1136/bmj.320.7233.484.

[5] R. S. Mazze, H. Shamoon, R. Pasmantier, D. Lucido, J. Murphy, K. Hartmann, V. Kuykendall, and W. Lopatin, "Reliability of blood glucose monitoring by patients with diabetes mellitus," *The American Journal of Medicine*, vol. 77, no. 2, pp. 211–217, 1984. DOI: 10.1016/0002-9343(84)90082-X.

[6] B. Abu-Salih, M. AL-Qurishi, M. Alweshah, M. AL-Smadi, R. Alfayez, and H. Saadeh, "Healthcare knowledge graph construction: A systematic review of the state-of-the-art, open issues, and opportunities," *Journal of Big Data*, vol. 10, p. 81, 2023. DOI: 10.1186/s40537-023-00774-9.

[7] N. Rastogi and M. J. Zaki, "Personal health knowledge graphs for patients," *arXiv preprint arXiv:2004.00071*, 2020. arXiv: 2004.00071 [cs.AI].

[8] B. J. W. Dudzik, J. S. van der Waa, P.-Y. Chen, R. Dobbe, Í. M. de Troya, R. M. Bakker, M. H. T. de Boer, Q. T. Smit, D. Dell'Anna, E. Erdogan, P. Yolum, S. Wang, S. Baez Santamaria, L. Krause, and B. A. Kamphorst, "Viewpoint: Hybrid intelligence supports application development for diabetes lifestyle management," *J. Artif. Int. Res.*, vol. 80, Sep. 2024, ISSN: 1076-9757. DOI: 10.1613/jair.1.15916.

[9] Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen, and M. Welling, "A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence," *Computer*, vol. 53, no. 8, pp. 18–28, Aug. 2020. DOI: 10.1109/MC.2020.2996587.

[10] M. W. Chekol, G. Pirrò, J. Schoenfisch, and H. Stuckenschmidt, "Tecore: A temporal conflict resolution framework for knowledge graphs," *Proceedings of the VLDB Endowment*, vol. 10, no. 12, pp. 1929–1932, 2017. DOI: 10.14778/3137765.3137811.

[11] M. W. Chekol, G. Pirrò, J. Schoenfisch, and H. Stuckenschmidt, "Marrying uncertainty and time in knowledge graphs," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017. DOI: 10.1609/aaai.v31i1.10495.

[12] J. Fähndrich and M. Wischow, "Automated reasoning for conflict solving in knowledge graphs," in *INFORMATIK 2024, Lecture Notes in Informatics (LNI)*, Gesellschaft für Informatik, Bonn, 2024. DOI: 10.18420/inf2024_24.

[13] A. L. Opdahl, T. Al-Moslmi, D. T. Dang Nguyen, M. Gallofré Ocaña, B. Tessem, and C. Veres, "Semantic knowledge graphs for the news: A review," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–38, 2022. DOI: 10.1145/3543508.

[14] P. Lin, Q. Song, Y. Wu, and J. Pi, "Discovering patterns for fact checking in knowledge graphs," *Journal of Data and Information Quality*, vol. 11, no. 3, Article 13, 2019. DOI: 10.1145/3286488.

[15] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PLOS ONE*, vol. 10, no. 6, e0128193, 2015. DOI: 10.1371/journal.pone.0128193.

[16] L. Lu, J. Fang, P. Zhao, J. Xu, H. Yin, and L. Zhao, "Eliminating temporal conflicts in uncertain temporal knowledge graphs," in *Web Information Systems Engineering – WISE 2018*, H. Hacid, W. Cellary, H. Wang, H.-Y. Paik, and R. Zhou, Eds., Cham: Springer International Publishing, 2018, pp. 333–347, ISBN: 978-3-030-02922-7.

[17] A. Zaveri, D. Kontokostas, S. Hellmann, J. Umbrich, M. Färber, F. Bartscherer, C. Menne, A. Rettinger, A. Zaveri, D. Kontokostas, S. Hellmann, and J. Umbrich, "Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago," *Semant. Web*, vol. 9, no. 1, pp. 77–129, Jan. 2018, ISSN: 1570-0844. DOI: 10.3233/SW-170275.

[18] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge graphs: Opportunities and challenges," *Artificial Intelligence Review*, vol. 56, no. 11,

pp. 13 071–13 102, Nov. 2023, ISSN: 1573-7462. DOI: 10.1007/s10462-023-10465-9.

[19] L. Cai, X. Mao, Y. Zhou, Z. Long, C. Wu, and M. Lan, *A survey on temporal knowledge graph: Representation learning and applications*, 2024. arXiv: 2403.04782 [cs.CL].

[20] J. Chen, J. Ren, W. Ding, and Y. Qu, "Patecon: A pattern-based temporal constraint mining method for conflict detection on knowledge graphs," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, pp. 4166–4172, 2023. DOI: 10.1609/aaai.v37i4.25533.

[21] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, *Owl 2 web ontology language primer (2nd edition)*, W3C Recommendation, 2012.

[22] J. Chen, J. Ren, W. Ding, H. Ouyang, W. Hu, and Y. Qu, *Conflict detection for temporal knowledge graphs:a fast constraint mining algorithm and new benchmarks*, 2023. arXiv: 2312.11053 [cs.AI].

[23] J. F. Allen, "Maintaining knowledge about temporal intervals," *Commun. ACM*, vol. 26, no. 11, pp. 832–843, Nov. 1983, ISSN: 0001-0782. DOI: 10.1145/182.358434.

[24] E. D. R. Pruett and S. Oseid, "Effect of exercise on glucose and insulin response to glucose infusion," *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 26, no. 3, pp. 277–285, 1970. DOI: 10.3109/00365517009046234.

[25] T. M. Wolever and J. B. Miller, "Sugars and blood glucose control," *The American Journal of Clinical Nutrition*, vol. 62, no. 1, 212S–227S, 1995, ISSN: 0002-9165. DOI: 10.1093/ajcn/62.1.212S.

[26] G. Wilcox, "Insulin and insulin resistance," *Clinical Biochemistry Reviews*, vol. 26, no. 2, pp. 19–39, 2005.

[27] S. Cox and C. Little, *Time ontology in owl*, W3C Candidate Recommendation Draft, Nov. 2022.

[28] "Sparql 1.1 query language," World Wide Web Consortium (W3C), Tech. Rep., Mar. 2013, W3C Recommendation, 21 March 2013. Editors: Steve Harris, Garlik (Experian); Andy Seaborne (The Apache Software Foundation). Available at: http://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[29] M. A. Musen, "The protégé project: A look back and a look forward," *AI Matters*, vol. 1, no. 4, pp. 4–12, 2015. DOI: 10.1145/2757001.2757003.

[30] *Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act)*, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, Official Journal of the European Union, L series, European Union, 2024.