



**Utrecht  
University**

Department of Information and Computing Science

---

**Master Business Informatics**

## **Modelling The Ad Hoc Nature of Nursing Interventions**

**First Supervisor**

Dr. Iris Beerepoot

**Second Supervisor**

Prof. dr. Sjaak Brinkkemper

**Student**

Steven van den Berg

May 3, 2024

## Abstract

District nurses perform numerous ad hoc activities, meaning the sequence of their activities depends on the particular situation they are faced with at that moment. As a result, the processes these nurses are involved in are largely unstructured and that leaves a challenge to understand the performed work. To bridge this gap, transcripts of client visits served as input to create different types of process models using event logs generated by Azure OpenAI. Several process models were created, and each model is a unique combination of process modelling notation and granularity level as described in a standardized nursing taxonomy. Next, these models were evaluated and standardized to investigate which model can encapsulate these processes given three quality metrics: Fitness, Precision, and Simplicity. This thesis illustrates the complexity of nursing activities, and describes how AI can play a role in discovering these processes. Additionally, the varying granularity levels of the event log serve as an example of how processes can be discovered on different levels of abstraction, facilitating communication with stakeholders on other levels of involvement. Based on the results, BPMN-D slightly outperforms BPMN on simplicity at the most detailed level, while BPMN performance increases as the abstraction level of activities increases. DECLARE was unable to express rich information about the process based on the event log used in this case study.

**Keywords:** Process Modelling, Nursing, BPMN-D, Process Mining, GPT-4, Prompt Engineering

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Research Questions . . . . .	5
1.2	Contribution . . . . .	6
<b>2</b>	<b>Research Method</b>	<b>7</b>
2.1	Case Study Context . . . . .	7
2.2	Data Collection . . . . .	8
2.3	Data Analysis . . . . .	10
<b>3</b>	<b>Related Literature</b>	<b>13</b>
3.1	Process Mining in Healthcare . . . . .	13
3.2	Imperative Process Modelling . . . . .	14
3.3	Declarative Process Modelling . . . . .	15
3.4	Hybrid Process Modelling . . . . .	16
3.5	Predictive Process Monitoring . . . . .	17
<b>4</b>	<b>Results</b>	<b>19</b>
4.1	Generating Event Logs Using GPT . . . . .	19
4.2	DECLARE . . . . .	39
4.3	BPMN . . . . .	43
4.4	BPMN-D . . . . .	47
4.5	SRQ1: What are the benefits and drawbacks of using Declarative Process Modelling? . . . . .	50
4.6	SRQ2: What are the benefits and drawbacks of using Imperative Process Modelling? . . . . .	52
4.7	SRQ3: What are the benefits and drawbacks of using Hybrid Process Modelling? . . . . .	53
4.8	SRQ4: What are important quality aspects for a process modelling technique? . . . . .	53
4.9	MRQ: How do different modelling techniques compare as a means to capture nursing processes? . . . . .	54
<b>5</b>	<b>Discussion</b>	<b>55</b>
5.1	Limitations . . . . .	57
5.2	Future Works . . . . .	58
<b>6</b>	<b>Conclusion</b>	<b>60</b>
	<b>Bibliography</b>	<b>64</b>

<b>A Appendices</b>	<b>65</b>
A.1 Appendix A . . . . .	65
A.2 Appendix B . . . . .	66

# 1. Introduction

Healthcare is a domain that is subject to numerous guidelines. These guidelines are a means to limit ineffective or potentially harmful medical interventions [1]. However, to deliver quality healthcare, organizations are imposed with increasingly heavier workloads related to secondary activities, especially in the Western world [2]. Medical procedures are well-established as universally accepted clinical guidelines constrain them, but they operate under the assumption of ideal circumstances. In reality, clinical interventions are known to have numerous variations as these procedures tend to be highly reliant on the patient. For example, they may be impacted by personal circumstances such as background, experiences, and logistics [3].

From a process perspective, preferably one (or multiple) model(s) should be constructed to represent each sequence of activities that can be executed accurately. Process standardization leads to similar input, which in turn leads to (more) similar output, and in a healthcare setting this means the crucial activities are performed with the highest quality more consistently [4]. In reality, it may prove extremely challenging to capture every single scenario. Ideally, each patient fits the exact same model and thus procedure, but this is not feasible when there are many variables affecting the process which in this case are the patient's personal circumstances [5].

On the other hand, having a process model for each unique variation would lead to thousands of different process models, leading to an overload of documentation and potentially harming operations due to the sheer number of models. Furthermore, obtaining these countless models would likely confuse readers. Imagine documenting each variation in a complex context, such as healthcare, it would prove difficult to find the corresponding process model in any given scenario due to the sheer amount of documentation. Specifically, documentation tends to be a tedious task that interrupts their core activities [6]. This thesis will also investigate how to push the limit of current technology to improve this aspect.

To further illustrate the overload of the diversity of nurses' responsibilities, tasks that nurses perform tend to go well beyond the scope of direct patient care.

Up to 24.1% of a nurse's time spent on activities is related to managerial activities, meaning they spend a large amount of time on activities that do not directly relate to patient care. Moreover, Lavander et al. found that less than half of the time spent by nurses was related to direct care, a pattern that emerged in each distinct participated group [7].

Nurse activities have been increasing in complexity for decades. Partially due to this complexity burn-outs are a common occurrence for nurses, as well as the large number of patients relative to the number of nurses [8]. Previous research suggested the need for reliable workload indicators for nurses [9], [10]. Subsequently, Levenstam and Engberg proposed a system to close this gap [11]. As a result of these high number of variations, the exact clinical procedure to follow is complex when compared to many other domains. As such it might prove beneficial to consider different approaches for healthcare. One such example is adopting a declarative modelling language instead of an imperative (also called procedural) modelling language, facilitating a more flexible approach to process modelling [12]. This flexible style of modelling tends to be a better match for ad hoc environments, as procedural languages generally harm models in complex environments concerning the readability [3].

This thesis will focus on evaluating 7 process models in healthcare to deal with the domain's inherent complexity. More specifically, the emphasis will be on district nurses in cooperation with UMC Utrecht in The Netherlands, contributing to the Care2Report (C2R) platform [13]. The goal beyond these automated reports is to predict future sequences of ongoing process executions, which is called predictive process monitoring [14]. To enable these predictions, a process model that reflects historical process executions is necessary to make future predictions.

## 1.1 Research Questions

Following the introduction and the context of this thesis, the next step is to formalize the focus by stating the relevant research questions. This leads to the main research question of this thesis:

*MRQ: How do different modelling techniques compare as a means to capture the nursing processes?*

To provide a satisfying answer, the main research question will be supported

by several sub-research questions.

*SRQ1: What are the benefits and drawbacks of using Declarative Process Modelling?*

*SRQ2: What are the benefits and drawbacks of using Imperative Process Modelling?*

*SRQ3: What are the benefits and drawbacks of using Hybrid Process Modelling?*

*SRQ4: What are important quality aspects for a process modelling technique?*

Considering the scope of this thesis, all of these research questions will be answered within the healthcare domain and focus on district nurses.

## **1.2 Contribution**

The distinguishing feature of this thesis is the comparison of Declarative Process Modelling and Imperative Process Modelling. While both of these modelling techniques have been investigated in the research literature at large, there seems to be a lack in the context of healthcare even though in theory it appears to be well-matched. In particular, declarative and hybrid modelling languages have a distinct lack of research which this thesis aims to contribute in bridging. In addition, we explore the potential of an event log where its activities can be investigated on multiple levels of granularity, allowing process discovery and communication on different levels of abstraction. Lastly, we apply generative AI techniques to event log generation, and identify its performance and reliability regarding the output as this could prove a useful tool for reducing the administrative burden of nurses.

## 2. Research Method

This thesis is a continuation within the scope of the C2R project, and aims to discover the workflow processes of nurses. The subsequent subsections will demonstrate the exact methods employed. There are many types of nurses performing numerous different activities. While a nurse has a myriad of responsibilities and executes many different activities in any given scenario, this cannot be captured within a single thesis. Therefore, this thesis focuses on one specific type of nurse: the district nurse. In particular, we scope down on the sequence of activities performed while visiting a client's house. Using transcripts from audio recordings of several such visits, we will design three process models and evaluate which model represents reality most accurately. To operationalize this accurately, within the context of this thesis this means that the sequences of activities existing in the model should also be reflected in real scenarios. Conversely, activity sequences that do not explicitly occur in the traces, should also not be included in the process models.

### 2.1 Case Study Context

This thesis will act as a case study and closely analyze the process where a district nurse visits a client to provide the care they need at the patient's home. District nurses perform a wide range of activities, helping the patient to the best of their ability. However, there is an increasing trend of people wanting to die at home, which tends to be the place where district nurses take care of their clients [15]. Moreover, due to the ad hoc nature of healthcare services, the exact sequence of activities tends to vary with every visit. As a result, the district nurse knows how to manoeuvre the hurdles they face, but there is no documented information on this knowledge. By inspecting this case study we aim to find a process model that accurately encapsulates each sequence of activities and captures this implicit knowledge.



## 2.2 Data Collection

All participants of the study reside in the central area of The Netherlands. The original audio recordings were recorded by the relevant district nurse and with explicit consent from the patient. Furthermore, transcripts of audio-recorded visits from the district nurses will serve as input for the process models. A total of three district nurses have contributed to these recordings, performing a combined total of 27 visits. The transcripts amounted to 100739 words in total, leading to an average of 3731 words per transcript. The longest transcript consists of 7352 words, while the shortest transcript is 650 words long.

Additionally, using the current ChatGPT model (GPT-4) an event log will be generated using the transcripts as input and the NIC taxonomy as a classification system. The model will be used to extract all nursing activities from the transcripts and classify them according to the NIC taxonomy. The NIC taxonomy is a standardized classification system for nursing activities [16], and it consists of three levels in a tree structure to tackle varying levels of granularity, with "Domain" at the highest level, "Classes" on the second level, and "Interventions" on the third and most fine-grained level of granularity. Due to privacy reasons, this will be conducted in a protected environment by using a GPT model that is hosted by the University. To support the prompt engineering we employ the CLEAR framework for creating prompts, providing us with a standardized framework to optimize the results [17]. One such example following the framework is the following prompt which has been used as a proof of concept for generating event logs based on transcripts using the proposed method:

*"Extract the nursing activities from the following transcript and map each activity to a level 3 Intervention of the NIC taxonomy. Respond by returning a csv formatted file containing the following columns of the respective activity: Patient number starting with P, district nurse number starting with DN, the exact quote in the transcript, the performed activity, NIC level 3 classification, Start timestamp of the activity, End timestamp of the activity. Exclude all comments or remarks about the content."*

This prompt reveals a preliminary result in figure 2.1. The figure contains only visual post-processing for readability of the figure, no content of the data has been altered.

Patient	District Nurse	Exact quote in the transcript	Performed activity	NIC level 3 classification	Start timestamp	End timestamp
P03	DN01	Zal ik de rug doen?	Bathing	Skin Care: Topical Treatments	09:58:15	09:58:17
P03	DN01	Ja, zal ik een voet doen?	Foot Care	Skin Care: Topical Treatments	09:58:34	09:58:35
P03	DN01	Onder de oksels en onder de t	Bathing	Skin Care: Topical Treatments	09:59:28	09:59:28
P03	DN01	En onderbroek.	Assisting with dressing	Self-Care Assistance: ADLs	10:02:22	10:02:24
P03	DN01	Zo, voelt u droog?	Drying the patient	Skin Care: Topical Treatments	10:02:16	10:02:18
P03	DN01	Benen ook of alleen de voeter	Applying lotion to legs	Skin Care: Topical Treatments	10:05:22	10:05:27
P03	DN01	Zo, we zijn weer klaar.	Concluding care session	Physical Comfort Promotion	10:06:54	10:06:54

**Figure 2.1:** Preliminary results of using GPT to generate event logs.

To validate the generated event log a domain expert will evaluate the results of the event log to check whether the classification and extraction of activities is correct. This will provide a more reliable base for the resulting process models will be helpful in two ways. First, it will improve the final process models. Second, it will elevate the quality of the event log, which is essential for quality models in this thesis, but also a stable foundation for future work or extension of the event logs. On top of this, the event log will be manually checked with the necessary corrections and extensions where necessary.

The collected event log will be imported into ProM 6.13 and subsequently translated into process models using the following notations:

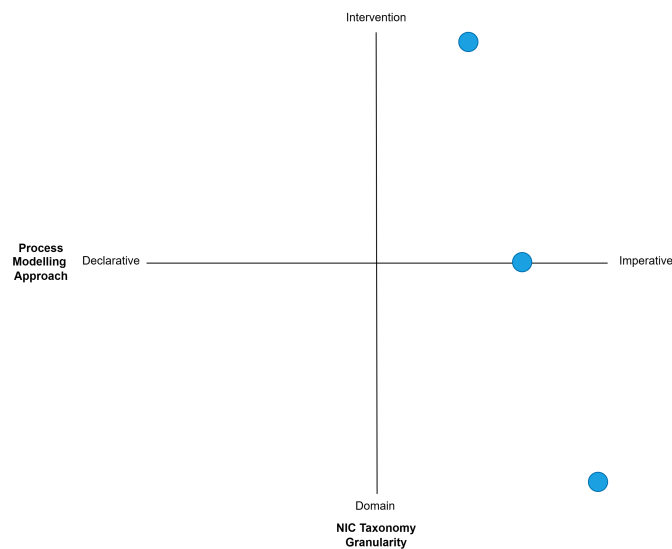
In particular, three unique process modelling notations will be used, each from a different branch of process modelling. This thesis will investigate the efficacy of the following notations:

1. Imperative notation: Business Process Management Notation (BPMN);
2. Declarative notation: DECLARE;
3. Hybrid notation: BPMN-D.

ProM is an open-source process mining framework that facilitates countless extensions called plug-ins, supporting numerous modelling languages [18]. ProM has two plug-ins available that fit and are useful considering this thesis. First, the BPMN miner plug-in allows the creation of BPMN models based on the BPMN 2.0 standard [19]. Second, DECLARE is supported by ProM using the Declare Miner plug-in [20]. The BPMN-D model has to be created manually, as there is currently a lack of support for this process modelling language. However, one additional limitation of ProM is the restricted capabilities to edit generated models. Consequently, the models will be recreated in Draw.io to allow easier and more elaborate modifiability of the models.

Finally, the event log will be expanded upon by adding the other two gran-

ularity levels of the NIC taxonomy. As the taxonomy is a tree-like structure, the domain and class can be deduced based on the intervention classification. To illustrate the consequences of these layers, we will outline our expectations regarding the relationship between the granularity level of the event log, and the approach of process modelling notation. Declarative process modelling tends to favor complex, unstructured processes. When looking at the NIC taxonomy, this complexity and unstructured nature is expected to largely reveal itself at the most detailed level, i.e. when inspecting the process on the intervention level as defined in the NIC taxonomy. From here, the expectation is that the more abstract the activities are, the fewer unique activities are performed, thus leading to a generally less complex model. As the activities become more coarse-grained, the model tends to shift to the imperative side of the spectrum as the declarative model's complex and restrictive nature will overcomplicate the model. This notion is portrayed in figure 2.2.



**Figure 2.2:** Taxonomy granularity and expected position on the process modelling approach spectrum.

### 2.3 Data Analysis

To decide which process model is considered best given a context, an evaluation method with a specific set of metrics is required as a means to measure their performance in a standardized manner. One well-known set of metrics is the combination of Recall and Precision [21], [22]. This set of metrics has been extended to include two more quality dimensions [23]. Each of the metrics receives a score between 0 and 1, where 1 is a perfect score. The following metrics will be used:

1. Recall (or replay fitness), representing the capability of the model to represent the behavior as it is present in the event log;
2. Precision, indicating whether the possible traces are only limited to the traces present in the event log. If the model can represent traces that are not in the provided event log, the precision score goes down;
3. Simplicity (or complexity), representing the readability aspect of the model. This quality dimension emphasizes minimalism: the model should be as simple as possible, only containing that which is required to convey its purpose.

These three metrics can be calculated mathematically, and the output of these calculations can be quantified objectively. Fitness and precision will be calculated using the formula defined in [24], but [25] introduced a method to objectively quantify simplicity. The proposed method was to calculate the simplicity looking only at the transitions in a petri net, and compare it to the number of unique activities in the event log. Seeing as the transitions are similar to the activities used in the models for this thesis, the same logic will be applied by using the activities from the respective models. The first evaluation consists of a domain expert evaluating the quality of the classification as part of the generated event log by the GPT-4 model. The second evaluation consists of judging the quality of the process models based on these three quality metrics, which will be calculated using the respective formulas shown below.

$$Fitness = \frac{Log \cap Model}{Log}$$

$$Precision = \frac{Log \cap Model}{Model}$$

$$Simplicity = 1 - \frac{\#ModelActivities}{\#ModelActivities + \#UniqueLogActivites}$$

Important to note is that these metrics can only be calculated for the BPMN and BPMN-D models, as DECLARE models are typically evaluated using support and confidence [26]. Instead, the DECLARE models will be evaluated using ProM for model generation with set parameters and analyzing constraints that fit these parameters per granularity level.

## 3. Related Literature

This chapter will start with a subsection describing the used method to find relevant literature to illustrate the current body of literature. The subsequent subsections will go into more detail on the specific literature, highlighting literature that best reflects the contemporary state of literature.

### 3.1 Process Mining in Healthcare

Process mining is an amalgamation of multiple prominent disciplines such as data mining and process modelling [27]. Process mining techniques rely on activities that tend to be registered in modern IT systems as they are executed. These so-called event logs may be extracted, pre-processed, and finally act as the foundation for a process model. Event logs are a collection of events, which is generally a record with the following specifications: case id, name of executed activity, timestamp, and resource. Case ID is the unique identifier for one case, referring to a single process execution. Furthermore, the event log may be enriched by other domain-relevant information. One such example within healthcare would be the age of the patient.

Process models can be used to gain valuable insight into the process execution with the aim of understanding and improving them [28]. Van der Aalst defined three predominant types of process mining, namely: discovery, conformance, and enhancement [27]. A general overview of process mining is illustrated in figure 3.1, portraying how each component in the larger process contributes to each iteration of process mining.

Process mining overlaps with the fields of Data Science and Process Science and is considered a combination of the two [29]. The number of event logs generated from healthcare processes has been increasing rapidly [28].

Due to the sheer amount of healthcare data and the complex processes recorded in Healthcare Information Systems (HIS), process mining has a clear use case within this domain. Process mining has already found its way to the healthcare domain a long time ago, showing its potential by improving process execution and reducing

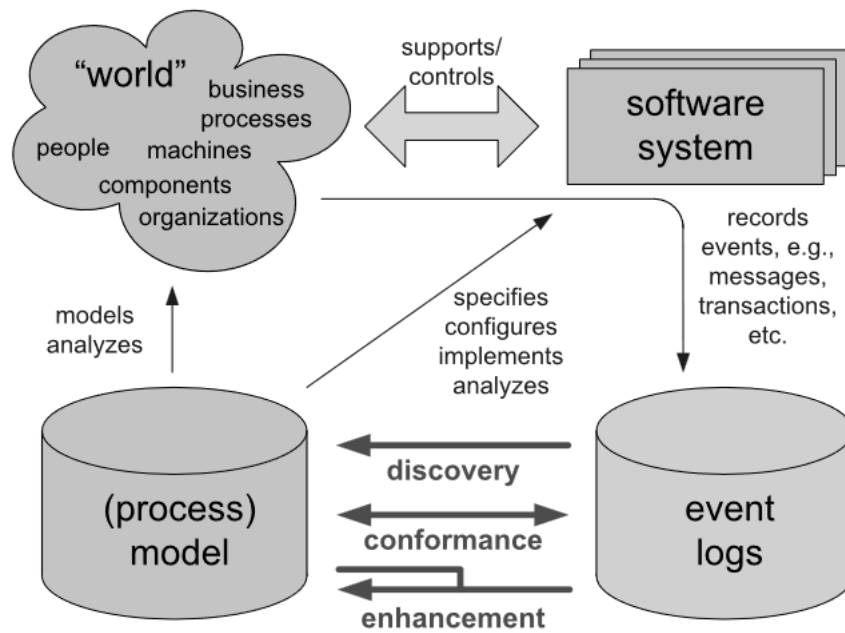


Figure 3.1: Process Mining, from [27].

costs simultaneously [30]. An early, prominent paper emphasizing the potential of process mining was published in 2009, offering the value proposition of automating many manual tasks and comprehensively portraying them [31].

To process this giant pool of data, there are multiple viable algorithms to distil the desired information. However, according to a literature review conducted in 2021, some algorithms are employed more frequently than others. More specifically, Fuzzy Miner algorithm and Heuristic Miner algorithm combined represent 56% of the total included process samples [32]. Furthermore, the same research indicated that these algorithms were primarily used for process discovery [32]. The predictive capabilities of the model represent 6.3%, revealing an interest in this goal but not quite as widely adopted as the other purposes.

### 3.2 Imperative Process Modelling

The most prominent paradigm of process modelling is imperative process modelling, also known as procedural process modelling. These models tend to be conceptually intuitive [33], and facilitate the execution of explicit sequential activities [34], [35]. As a result, each possible sequence needs to be mapped out in advance, which tends to only be feasible in trivial situations [36]. For this reason, imperative process models tend to be easy to follow. Common examples of this type of process model languages are petri nets, and BPMN. Each sequence of events is explicitly

modelled, meaning that generally, each subsequent step is predictable. Similarly, the history of a sequence may also be deduced using the same logic but backwards. Given a random activity in an imperative process model, it may be possible to find the executed sequence that led to that specific activity or narrow down the potential traces due to the sequential nature of these modelling techniques.

Visual representation supports users in learning and grasping new information [37]. Process modelling techniques apply the same general concept to process understanding, i.e. using imagery to further detail and explain a process. Research has indicated that, while there is an initial learning curve, imperative process modelling aids in understanding a process over written text [38], [39]. However, this does not mean that process modelling can replace written documentation. Instead, both text and graphical models act as a supplement to each other, enriching the presented information [38]. Furthermore, in the context of understanding a process, models tended to outperform textual variations [40].

### **3.3 Declarative Process Modelling**

Declarative process modelling techniques are fundamentally different from imperative process modelling techniques. Whereas imperative process modelling techniques aim to explore each possible sequence of steps, declarative process modelling techniques instead emphasize constraints [35], [41], [42]. This means that instead of explicitly creating a process model for each situation, only the constraints are defined. Furthermore, a single model may hold much more implicit knowledge when modelled using declarative modelling techniques [43]. Considering the dynamic, complex, and ad-hoc nature of healthcare, this matches well with the intended purpose of declarative modelling [44]–[46].

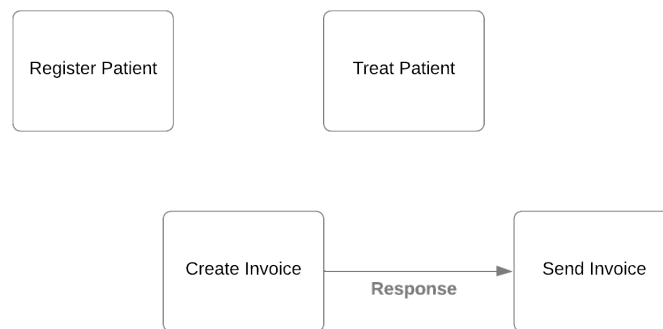
Declarative process modelling is not new in the healthcare domain. For example, Comma is a language that has been applied in the past to deal with the high process flexibility that is often required [47]. Additionally, Declare has shown similar promise as a declarative modelling technique, highlighting the potential of these types of models [3]. Rovani et al. stated that although Declare had a noticeable learning curve, it revealed insightful information for the process analysts investigating how patients are treated. Another instance of a declarative modelling technique is Dynamic Condition Response (DCR) Graphs [48]. Although the exact modelling language employed varies, each research has similar conclusions stat-



ing that declarative modelling has several clear benefits over imperative modelling languages. These advantages primarily relate to flexibility in an ad-hoc environment in richer models. For example, [48] explains how DCR Graphs facilitate the need to reconsider previously executed activities.

### 3.3.1 Declare

One prominent example of a declarative modelling language is Declare, of which a simple example is portrayed in figure 3.2 exemplifying how the "Create Invoice" activity has to always be eventually followed by the "Send Invoice" activity. The other two activities "Register Patient" and "Treat Patient" can be executed multiple times and at any time in the trace. Note that this is not a realistic example, but rather a simplified view to showcase the basic concepts of Declare. Declare is a language that has been used frequently in the existing body of literature for over a decade, even within the context of healthcare institutions [3], [49]. While Declare has its clear strengths being a declarative process modelling language, it also inherits its weaknesses. More specifically, it contains a substantial amount of implicit information and therefore requires the reader to have the capabilities to extract this knowledge [33]. So while the model may be richer compared to a BPMN model, the contained information may be too well hidden for inexperienced consumers of the model.



**Figure 3.2:** Simple example of a process model based on Declare.

## 3.4 Hybrid Process Modelling

Relatively new to the literature are hybrid process models or hybrid languages: process models combining characteristics of two unique process modelling techniques and combining them into a single process model [50], [51]. The rationale

behind this merge is, ideally, to get the best of both worlds while eliminating the worst. Hybrid process models seemingly perform well as far as the current body of literature has stated. However, this approach does not only include upsides. Adopting the upsides of a feature means accepting the downsides as well. An exhaustive subset of these downsides is as follows [51]:

1. Understandability, as there has been a lack of this property for hybrid process models presently. Depending on the specific notation used, the body of research may show slight improvements;
2. Tools and maturity, as there are very few, if any, actively maintained tools that support the implementation of hybrid process models;
3. Input data, referring to the event log (i.e. the foundation of the model) only contains successful executions. Since hybrid models are partially of declarative nature, this means that constraints need to be established. However, constraints cannot be adequately identified if the event log only contains cases where the end was reached. There are implicit constraints in an event log, but those are only a subset of the complete set of constraints.

#### **3.4.1 BPMN-D**

There are a plethora of hybrid modelling languages available, with room for countless potential extensions and possible amalgamations. However, as we have established BPMN and Declare as modelling languages previously, the most fitting hybrid notation within the scope of this thesis would be a mix of these two languages. Such a language has been introduced and appropriately named BPMN-D, which has already been used sparsely in healthcare [52]. In essence, BPMN-D uses all standard features from BPMN and cherry-picks select features from Declare as an add-on. By this logic "*... any BPMN model is also a BPMN-D model*" [52].

### **3.5 Predictive Process Monitoring**

Accurately portraying the activity sequence of nurses is the first step to implementing a system capable of predicting ongoing process instances called Predictive Process Monitoring (PPM) [53], [54]. PPM is a subcategory of process mining that attempts to predict the outcome of an ongoing process instance [14]. As stated previously, PPM has made its way to the healthcare domain, although complete adoption of this technology has yet to be embraced [32]. However, the purpose of

PPM is evident: knowing the developments of a process execution allows one to anticipate and either prepare for that outcome or change its course to alter the outcome. Being able to apply this technology to healthcare seamlessly would prove a milestone in healthcare quality improvement. Part of the reason PPM has not seen wide adoption yet is due to the difficulty of its implementation, particularly pertaining to which method or framework to employ [54]. The same paper outlines several types of PPM, depending on what kind of feature they are attempting to predict:

1. Numeric, referring to time and cost-related predictions;
2. Categorical, referring to risk assessment and categorical outcome predictions;
3. Activity sequence, referring to predicting the next sequence of activities in an ongoing process instance.

Each type has its use, and aiming to predict one feature does not necessarily lead to more value or interesting information than the other, rather each type emphasizes different aspects. Therefore, one type may be more suitable depending on the goal.

## 4. Results

This chapter presents the process of generating an event log by giving transcripts as input for the GPT-4 model, and an evaluation of its performance and the encountered obstacles. Subsequently, we present the resulting models based on this event log, starting with DECLARE, followed by BPMN, and ending with BPMN-D.

### 4.1 Generating Event Logs Using GPT

This section will be dedicated to portraying the results from the event log generation through Azure OpenAI, showcasing the iterative process of improving a prompt and how it affects the output. Code snippets and output excerpts will be revealed throughout the section, but the elaborate output can be found in the appendices.

#### 4.1.1 Output Results

To build a robust solution to generate event logs based on transcripts, we need to construct a prompt that instructs GPT to produce the desired output. This means carefully instructing the model to provide both the information that forms the foundation of the event log, and also information that will enrich the event log. Specifically, the GPT model used in this thesis is the GPT-4 model with a temperature parameter of 0. We interacted with the model through the C# API using the Azure OpenAI package version 1.0.0-beta.7.

The foundational information is rather straightforward, consisting of the patient, the nurse, the activity, the start time, and the end time of the activity. Additionally, we ideally want more detailed information regarding the activity. More specifically, we instructed the GPT model to classify the extracted activities based on the NIC taxonomy as defined in [16]. This taxonomy allows the event log to be inspected on different levels of granularity, having several benefits. Firstly, it allows us to reveal patterns that are not (as) visible on the more detailed granularity levels due to the sheer overload of information. Choosing the right level of granularity of the model will facilitate the discovery of such patterns. Secondly,

the levels of granularity may serve as a form of communication to stakeholders who are interested in different levels of the process. For example, a nurse would know the specific interventions performed, and could properly understand what the process looks like on the most detailed level. Furthermore, the nurse could use this information to develop their understanding of the process even further, as they are the ones operating on that level of granularity. However, a manager who does not operate on the intervention level could be more interested in a higher level of abstraction, gaining an understanding of the process on a less fine-grained level. In short, the event log having multiple of these levels allows communication and improves the understanding of the process on multiple organizational levels. Lastly, for the event log, we want a format that is compatible with event log generation. The one used for this specific prompting is the CSV format.

Given these specifications for the desired output, this section will outline the used prompts and their respective results, illustrating how prompt engineering and the ability to express the desired output affect the outcome. With these boundaries, the initial C# is as follows:

```
return new List<ChatMessage>
{
    new ChatMessage(ChatRole.User,
        "Extract the nursing activities
        from the following transcript and map each activity
        to a level 3 Intervention of the NIC taxonomy. " +
        "Respond by returning a csv formatted file containing
        the following columns of the respective activity: " +
        "Patient number starting with P, " +
        "district nurse number starting with DN, " +
        "the exact quote in the transcript" +
        "the performed activity," +
        "NIC level 3 classification, " +
        "Start timestamp of the activity, " +
        "End timestamp of the activity." +
        "Exclude all comments or remarks about the content."),
    new ChatMessage(ChatRole.User, transcription),
};
```

This code block serves as the setup for the request sent to the model and it consists of two messages. The first message is explaining the task to the model, while the second message is feeding the transcripts to the model.

1	Patient	District nurse	Transcript Quote	Activity	NIC level 3 classification	Start timestamp	End timestamp
2	P01	DN01	Nou ga ik alvast naar bov	Ascending stairs	Activity and Exercise Management: Indirect Activity Facilitation	8:48:11	8:48:11
3	P01	DN01	Bent u nog naar buiten ge	Assessing patient's mol	Functional Assessment	8:49:17	8:49:23
4	P01	DN01	Ja, hij hangt weer ja, me	Assessing patient's adh	Exercise Therapy: Ambulation	8:49:56	8:50:02
5	P01	DN01	Moeten we de haren was	Inquiring about hair wa	Hygiene	8:50:52	8:50:53
6	P01	DN01	Niet te veel bewegen. Go	Advising on physical ac	Fluid Management	8:51:16	8:51:22
7	P01	DN01	Daar komt die, het is wel	Assessing healing of sk	Wound Care	8:55:04	8:55:07
8	P01	DN01	Eerst even de hoeven sch	Washing patient's feet	Hygiene: Foot Care	8:55:31	8:55:33
9	P01	DN01	Ja goed schoonmaken da	Cleaning between toes	Hygiene: Foot Care	8:56:32	8:56:42
10	P01	DN01	Oh nee, de rug en de ond	Washing patient's back	Bathing	8:57:40	8:57:40
11	P01	DN01	Zal ik beginnen met de s	Starting to wash the ps	Skin Surveillance	8:54:49	8:54:49
12	P01	DN01	Even dat muts af.	Removing patient's hea	Hygiene	8:58:41	8:58:41
13	P01	DN01	Gelijk even naar de nave	Inspecting patient's na	Wound Care	8:59:58	9:00:18
14	P01	DN01	Zo zal ik verder met de a	Drying patient's back	Hygiene	9:00:50	9:00:50
15	P01	DN01	Dan gaan we even de bod	Applying body lotion	Skin Care	9:01:58	9:02:09
16	P01	DN01	Ja, ik zal wel even de ple	Removing bandages	Wound Care: Dressing Change	9:05:14	9:05:14
17	P01	DN01	Wilt u met een rollator ne	Offering mobility assis	Assistive Device Management	9:04:30	9:04:33
18	P01	DN01	Ja, steeds moeilijker me	Assisting with underwe	Self-Care Assistance: ADLs	9:03:49	9:03:55
19	P01	DN01	Nou ja, u kunt ze nu wel i	Inquiring about hearing	Communication Enhancement: Hearing Deficit	9:59:21	9:59:27
20	P01	DN01	Nou misschien net nog wa	Discussing exercise tim	Exercise Promotion: Strength and Flexibility	9:14:57	9:14:59
21	P01	DN01	Oh die moet u wel even c	Prompting to answer th	Communication: Telephone	9:15:04	9:15:16
22	P01	DN01	Nou gefeliciteerd.	Congratulating on fami	Emotional Support	9:17:43	9:17:45
23	P01	DN01	Nou dan hijsen we nog ev	Assisting with dressing	Self-Care Assistance: Dressing/Grooming	9:19:49	9:19:55
24	P01	DN01	Nog even de gulp dicht dk	Assisting with clothing	Self-Care Assistance: Dressing/Grooming	9:20:33	9:20:33
25	P01	DN01	En tot een volgende keer.	Ending the visit	Professional Practice	9:22:40	9:22:43
26	P02	DN01	Ja hartstikke goed ga ik i	Medication Manager		2310	9:35:12
27	P02	DN01	Nou, ik zal ze zo meteen	Medication Manager		2310	9:35:26
28	P02	DN01	Even kijken hoeveel bloe	Medication Manager		2310	9:36:45
29	P02	DN01	Haal ik ze ook voor u uit	Medication Manager		2310	9:38:01

Figure 4.1: An excerpt of the initial generated event log.

The result of the prompt is shown in figure 4.1. Based on the output, several aspects are returned accurately, and some aspects we can improve upon in the prompt. On the positive side, it extracts the patient, district nurse, transcript quote, the timestamps, and the activity accurately. There is no misquoted text, and the model accurately extracts the activity and places it in context correctly. For example, the extracted transcript quote includes "drying the backside", which the model rightfully classifies as "Drying patient's back", even though the exact quote is missing this context.

However, some aspects can be yet improved through better prompting. GPT is particularly inconsistent in following the NIC taxonomy. For example, it classifies an activity as "Hygiene: Foot Care", which simply does not exist within the taxonomy. Hygiene exists, and Foot Care too, but not this specific combination. There are multiple instances where GPT is only half correct in the classification, such as "Hygiene" and "Skin Care". Furthermore, the full classification consists of both a number, and a name. The current output shows three types of results: only the name, only the code, or nothing.

To tackle these challenges, we need to improve the prompt on the following aspects: first, we need to explicitly specify the desired response from GPT on the

NIC taxonomy. The current output is inconsistent, and this is one solution. Similarly, when it cannot find an accurate classification, it should return a statement indicating a match was not found instead of simply excluding it from the output. Second, to prevent hallucination we need to include the taxonomy as input for the model so the model's options are known and more consistent. In addition, to align the classification language with the transcript language the provided classification will be in Dutch, based on the Dutch translation of the aforementioned book. The Dutch translation was published in 2010, which is the third revision of the book.

### 4.1.2 Initial Output

The required improvements as identified in the previous output lead to the following C# code with a revised prompt:

```
return new List<ChatMessage>
{
    new ChatMessage(ChatRole.User, dutchNicInterventionsString),
    new ChatMessage(ChatRole.User, "
    Extract the nursing activities from the following transcript
    and map each activity to one of the previously defined
    Dutch NIC taxonomy interventions."
    + "If no fitting match can be found between the
    extracted activity and the taxonomy, set the classification as "No match"
    + "Respond in Dutch by returning a csv formatted file
    containing the following columns of the respective
    activity: " +
    "Patient number starting with P, " +
    "district nurse number starting with DN, " +
    "the exact quote in the transcript" +
    "the performed activity," +
    "the full NIC level 3 classification consisting of
    a number and a name," +
    "Start timestamp, " +
    "End timestamp" +
    "Exclude all comments or remarks about the content."),
```

```
new ChatMessage(ChatRole.User, transcription),
};
```

This prompt included the suggested improvements from the previous paragraph, and some additional statements to support these actions. More specifically, it sets the classification as "No match" when no appropriate match was found with the NIC interventions. We chose to include the result in the event log instead of excluding it to make the results more transparent, allowing us to improve the prompt more accurately. If these results were excluded, it would be impossible to trace the source, and thus the reason, for these missing classifications. Furthermore, it allows us to understand what the model considers an activity, which is an aspect of the prompt that is yet to be explicitly defined. The most noteworthy change in the prompt is explicitly defining what a classification consists of, namely a number and the name.

### 4.1.3 Revised Output

Extending the prompt with this snippet leads to the output shown in figure 4.2.

1	Patient	District nurse	Transcript Quote	Activity	NIC level 3 classification	Start timestamp	End timestamp
2	P01	DN01	Ja, je merkt meteen dat	Bevordering van licha	0201 Bevordering van lichamelijke oefening: krachttraining	8:50:09	8:50:11
3	P01	DN01	En dan een beetje de oef	Bevordering van licha	0201 Bevordering van lichamelijke oefening: krachttraining	8:51:29	8:51:42
4	P01	DN01	Moeten we de haren wass	Persoonlijke hygiëne	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	8:50:52	8:50:53
5	P01	DN01	Zal ik beginnen met de sp	Wassen	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	8:54:49	8:54:53
6	P01	DN01	Daar komt die, het is wel	Wondverzorging	3660 Wondverzorging	8:55:04	8:55:07
7	P01	DN01	Even dat muts af.	Haarverzorging	1670 Haarverzorging	8:58:41	8:58:41
8	P01	DN01	Dan gaan we even de bod	Huidverzorging	3580 Beleid bij pruritus E	9:01:58	9:02:09
9	P01	DN01	Ja, steeds moeilijker met	Kleden	1802 Ondersteuning bij persoonlijke zorg: kleden/ uiterlijke ver	9:03:49	9:03:55
10	P01	DN01	Wilt u met een rollator ne	Mobiliteitsondersteunir	0970 Verplaatsing bij beperkte mobiliteit	9:04:30	9:04:33
11	P01	DN01	Ja, ik zal wel even de plei	Verzorging van een wor	3660 Wondverzorging	9:05:14	9:05:14
12	P01	DN01	Nou, deze is altijd duideli	Aanbrengen van een pl	3662 Wondverzorging: gesloten drainage	9:13:08	9:13:08
13	P01	DN01	Nog even de gulp dicht dc	Kleden	1802 Ondersteuning bij persoonlijke zorg: kleden/ uiterlijke ver	9:20:33	9:20:33
14	P02	DN01	Ja, voor de medicijnen.	Toediening van genees	2300 Toediening van geneesmiddelen	9:35:10	9:35:10
15	P02	DN01	Ja hartstikke goed ga ik	Toediening van genees	2300 Toediening van geneesmiddelen	9:35:12	9:35:12
16	P02	DN01	Nou, ik zal ze zo meteen	Toediening van genees	2300 Toediening van geneesmiddelen	9:35:26	9:35:26
17	P02	DN01	Ja en bloedverdunner.	Toediening van genees	2300 Toediening van geneesmiddelen	9:36:35	9:36:35
18	P02	DN01	Even kijken hoeveel bloe	Toediening van genees	2300 Toediening van geneesmiddelen	9:36:45	9:36:45
19	P02	DN01	Zo. Haal ik ze ook voor u	Toediening van genees	2300 Toediening van geneesmiddelen	9:38:01	9:38:01
20	P02	DN01	Er ligt nog. Hoeveel tema	Toediening van genees	2300 Toediening van geneesmiddelen	9:38:34	9:38:34
21	P02	DN01	Dan heeft het lichaam he	Huidverzorging	3584 Huidverzorging: lokale behandeling	9:40:10	9:40:10
22	P02	DN01	Ik krijg gewoon een belle	Stomazorg	No match	9:37:30	9:37:30
23	P03	DN01	In de wasmachine of in di	Ondersteuning bij pers	1804	9:56:20	9:56:36
24	P03	DN01	Maar waar merkt u dat de	Bewaking van de voed	1160	9:58:10	9:58:15
25	P03	DN01	Zal ik de rug doen?	Ondersteuning bij pers	1801	9:58:15	9:58:17
26	P03	DN01	Ja, zal ik een voet doen?	Ondersteuning bij pers	1801	9:58:34	9:58:35
27	P03	DN01	Onder de oksels en onder	Ondersteuning bij pers	1801	9:59:28	9:59:35
28	P03	DN01	En onderbroek.	Ondersteuning bij pers	1802	10:02:22	10:02:24
29	P03	DN01	Kijken de blauwe broek.	Ondersteuning bij pers	1802	10:04:29	10:04:29

Figure 4.2: Revised generated event logs.

Even though the prompt included the definition for the level 3 classification, the output does not match the defined format. The yellow-marked cells are still missing the name of the classification, even though the number is consistently present in the output. As the output is still inconsistent, the prompt can be further improved



by more accurately phrasing the desired output, and additionally emphasizing the importance of the format.

#### 4.1.4 Domain Extended Output

To continue enriching the event log and to identify the activities on a lower granularity level, the desire to include the respective level 1 domain classification of the NIC taxonomy is added to the prompt. The updated prompt is shown in the code block below.

```
return new List<ChatMessage>
{
    new ChatMessage(ChatRole.User, dutchNicInterventionsString),
    new ChatMessage(ChatRole.User, dutchNicDomainString),
    new ChatMessage(ChatRole.User,
        "Extract the nursing activities from the following
        transcript and map each activity to one of the
        previously defined Dutch NIC taxonomy interventions."
        + "If no fitting match can be found between the
        extracted activity and the taxonomy, set the
        classification as \"No match\""
        + "Respond in Dutch by returning a csv formatted
        file containing the following columns of the
        respective activity: "
        + "Patient number starting with P, " +
        "district nurse number starting with DN, " +
        "the exact quote in the transcript" +
        "the performed activity," +
        "the full NIC classification consisting of a number
        and a name, " +
        "The respective level 1 domain from the
        Dutch NIC taxonomy as previously defined" +
        "Start timestamp, " +
        "End timestamp" +

        "Exclude all comments or remarks about the content."),
```

```

new ChatMessage(ChatRole.User, "Ensure the output
classifications are included in the predefined taxonomy"),

new ChatMessage(ChatRole.User, transcript),
};

```

The updated prompt contains a component to list the level 1 domain of the NIC taxonomy in addition to the level 3 intervention. The domain is classified based on the extraction, aiming to reproduce the tree-like structure. Similar to the interventions, all seven domains are appended to the model as context for the prompt. To emphasize the importance of correct and complete classification of the activities, a new message has been added to ensure the output classifications match the predefined taxonomy as provided in the initial context. The results of the output are illustrated in figure 4.3.

1	Patient	District nurse	Transcript quote	Activity	NIC Intervention	NIC Domain	Start timestamp	End timestamp
36	P03	DN01	In de vasmachine of in d	Ondersteuning bij pers	1804 Ondersteuning bij persoonlijke zorg: toiletgang F	Veiligheid	9:56:20	9:56:36
37	P03	DN01	En hoe is het verder?	Voorlichting: ziektepro	5602 Voorlichting: ziekteproces	Gedrag	9:57:31	9:57:31
38	P03	DN01	En wat heeft u dan precies	Voorlichting: ziektepro	5602 Voorlichting: ziekteproces	Gedrag	9:57:42	9:57:45
39	P03	DN01	En merkt u dat ook lichar	Voorlichting: ziektepro	5602 Voorlichting: ziekteproces	Gedrag	9:57:59	9:58:02
40	P03	DN01	Zal ik de rug doen?	Ondersteuning bij pers	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Veiligheid	9:58:15	9:58:17
41	P03	DN01	Ja, zal ik een voet doen?	Ondersteuning bij pers	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Veiligheid	9:58:24	9:58:35
42	P03	DN01	Hoe is het met de sullen?	Voorlichting: ziektepro	5602 Voorlichting: ziekteproces	Gedrag	9:58:52	9:58:55
43	P03	DN01	Onder de oksels en onder	Ondersteuning bij pers	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Veiligheid	9:59:28	9:59:28
44	P03	DN01	En ve zijn zo smerig he.	Ondersteuning bij pers	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Veiligheid	9:59:47	9:59:47
45	P03	DN01	En dit, zit dat er al lange	Bevaking van de huid:	3590 Bevaking van de huidconditie	Veiligheid	10:01:51	10:01:53
46	P03	DN01	Is wel goed om even in d	Bevaking van de huid:	3590 Bevaking van de huidconditie	Veiligheid	10:02:00	10:02:03
47	P03	DN01	Zo, voelt u droog?	Ondersteuning bij pers	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Veiligheid	10:02:16	10:02:18
48	P03	DN01	En onderbroek.	Ondersteuning bij pers	1802 Ondersteuning bij persoonlijke zorg: kleden/ uiterlijke ver	Veiligheid	10:02:22	10:02:24
49	P03	DN01	Heeft u nog onderzoeken	Voorlichting: ziektepro	5602 Voorlichting: ziekteproces	Gedrag	10:03:46	10:03:57
50	P03	DN01	Rijken de blauwe broek.	Ondersteuning bij pers	1802 Ondersteuning bij persoonlijke zorg: kleden/ uiterlijke ver	Veiligheid	10:04:29	10:04:29
51	P03	DN01	Zo nou, mag ik je echt ve	Ondersteuning bij pers	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Veiligheid	10:05:13	10:05:19
52	P03	DN01	Benen ook of alleen de vc	Ondersteuning bij pers	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Veiligheid	10:05:22	10:05:27
53	P03	DN01	Het is een lekker plekje l	No match	No match	No match	10:06:40	10:06:41
54	P03	DN01	Zo, ve zijn veer klaar.	Afsluten van de zorg	6492 Afsluten van de zorg	Gezondheidszorg	10:06:54	10:06:54
55	P04	DN01	Dan ga ik het even droog	Bevordering van de ve	6490 Veiligheid (NON-EXISTENT INTERVENTION)	Veiligheid	10:34:23	10:34:23
56	P04	DN01	Heb je diarree had? Ja? V	Diarree Beleid	0460 Diarree Beleid	Gedrag	10:35:05	10:35:15
57	P04	DN01	Nu alweer rustig met de c	Beleid bij diarree	0460 Diarree Beleid	Gedrag	10:35:19	10:35:19
58	P04	DN01	Gaan ve je haren wassen	Helpen met wassen	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Elementair Fysio	10:36:13	10:36:17
59	P04	DN01	Hoe lang heb je nu last v	Oogverzorging	1650 Oogverzorging	Elementair Fysio	10:36:45	10:37:04
60	P04	DN01	Heb je veel al ontbeten?	Bevaking van de voed	1160 Bevaking van de voedingsstoestand	Elementair Fysio	10:38:09	10:38:16
61	P04	DN01	Komt de diëtist nog bij je	Voedingsbeleid	1100 Voedingsbeleid	Elementair Fysio	10:38:30	10:38:33
62	P04	DN01	Ik ga het uitspelden, mag	Helpen met wassen	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Elementair Fysio	10:38:40	10:38:40
63	P04	DN01	Welke douchegel is van j	Helpen met wassen	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Elementair Fysio	10:39:03	10:39:15
64	P04	DN01	Ik zag op je rug ook al ve	Huidverzorging	3580 Beleid bij pruritus	Elementair Fysio	10:39:44	10:39:50
65	P04	DN01	Zal ik je achterkant doen	Helpen met wassen	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Elementair Fysio	10:40:28	10:40:28
66	P04	DN01	Ik ga nog even goed je b	Helpen met wassen	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Elementair Fysio	10:41:00	10:41:00
67	P04	DN01	Ja, dan moet je je buiks	Spiroefeningen: bekkn	0560 Spiroefeningen: bekkn	Elementair Fysio	10:41:19	10:41:19
68	P04	DN01	Zal ik je haar ook nog ev	Helpen met wassen	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Elementair Fysio	10:44:15	10:44:15
69	P04	DN01	Dan ga ik even je rug tro	Huidverzorging	3580 Beleid bij pruritus	Elementair Fysio	10:46:36	10:46:36
70	P04	DN01	Ik heb van de veek niet g	Gewichtsregulering	1200 Gewichtsregulering	Elementair Fysio	10:49:33	10:49:33
71	P04	DN01	En dan mag je die bestel	Verzorging van een aar	1870 Verzorging van een aan/afvoerbus	No match	10:52:20	10:52:20
72	P04	DN01	Met de assistentie.	Verzorging van een aar	1870 Verzorging van een aan/afvoerbus	No match	10:52:30	10:52:30
73	P04	DN01	Die heb ik al een poosje.	Aanbrengen van steunl	1800 Ondersteuning bij persoonlijke zorg	Elementair Fysio	10:53:04	10:53:10

Figure 4.3: Generated event logs including Domain classification.

Note that the output contains multiple wrong classifications as shown through the yellow marked cells. Additionally, it contains hallucinated intervention classifications, such as 6490 and 6492, which simply do not exist in the taxonomy. To improve the prompt, we need to reduce the hallucination capabilities of the model to a minimum.

#### 4.1.5 Domain Revised Output

To tackle the hallucination from the previous prompt, an additional message is included to make sure the options in the model are exclusively limited to the pre-defined taxonomy. This goes for both the intervention-level classifications and the domain-level classifications.

```
return new List<ChatMessage>
{
    new ChatMessage(ChatRole.User, "These are the options for
    level 3 interventions: " + dutchNicInterventionsString),
    new ChatMessage(ChatRole.User, "These are the options for
    level 1 domains: " + dutchNicDomainString),
    new ChatMessage(ChatRole.User,
    "Extract the nursing activities from the following transcript
    and map each activity to one of the previously defined
    Dutch NIC taxonomy interventions."
    + "If no fitting match can be found between the
    extracted activity and the taxonomy, set the
    activity and all classifications as \"No match\"
    while keeping the patient, district nurse,
    and referenced transcript excerpt in the output."
    + "Respond in Dutch by returning a csv formatted file
    containing the following columns of the respective activity:"
    + "Patient number starting with P, " +
    "district nurse number starting with DN," +
    "the exact quote in the transcript" +
    "the performed activity, " +
    "the level 3 intervention from the Dutch NIC taxonomy as previously defined," +
    "Infer the level 1 domain from the specified level 3 intervention.
    Limit the options to the predefined domains. " +
    "Start timestamp, " +
    "End timestamp " +

    "Exclude all comments or remarks about the content."),
    new ChatMessage(ChatRole.User, "Ensure both the
```

```

level 1 and level 3 classifications are included in
the output and are exclusively limited to the
predefined taxonomy."),
new ChatMessage(ChatRole.User, transcript),
};

```

Compared to the previous prompt, this prompt has several minor changes aimed at removing the faulty classifications, in addition to the inconsistency of missing data. This is visible mostly through the addition of "Limit the options to the predefined domains", as well as the final line of the prompt noting "Ensure both the level 1 and level 3 classifications are included...". The rationale was to explicitly limit the classification options, reducing the amount of hallucination. For a snippet of the output, see figure 4.4.

1	Patient	District nurse	Exact quote in the transcript	Activity	Level 3 Intervention	Level 1 Domain
40	P03	DN01	En merkt u dat ook lichamelijk?	Monitoring van de gezondheidstoestand	1160 Bewaking van de voedingstoestand	Complex Fysiologische Functies
41	P03	DN01	Zal ik de rug doen?	Wassen van de patiënt	1610 Baden/douchen	Elementair Fysiologische Functies
42	P03	DN01	Ja, zal ik een voet doen?	Wassen van de patiënt	1610 Baden/douchen	Elementair Fysiologische Functies
43	P03	DN01	Hoe is het met de suiker?	Monitoring van de gezondheidstoestand	1160 Bewaking van de voedingstoestand	Complex Fysiologische Functies
44	P03	DN01	Onder de oksels en onder de borst?	Wassen van de patiënt	1610 Baden/douchen	Elementair Fysiologische Functies
45	P03	DN01	En dit, zit dat er al langer?	Inspectie van de huid	3590 Bewaking van de huidconditie	Complex Fysiologische Functies
46	P03	DN01	Is wel goed om even in de gaten te	Monitoring van de gezondheidstoestand	1160 Bewaking van de voedingstoestand	Complex Fysiologische Functies
47	P03	DN01	Zo, voelt u droog?	Ondersteuning bij persoonlijke zorg: wassen/hygiëne	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Elementair Fysiologische Functies
48	P03	DN01	En onderbroek.	Ondersteuning bij persoonlijke zorg: kleden/ uiterlijke v	1802 Ondersteuning bij persoonlijke zorg: kleden/ uiterlijke v	Elementair Fysiologische Functies
49	P03	DN01	Benen ook of alleen de voeten?	Aanbrengen van lotion	3584 Huidverzorging: lokale behandeling	Complex Fysiologische Functies
50	P03	DN01	Het is een lekker plekje hierbuiten.	Ondersteuning bij persoonlijke zorg: verplaatsing	1806 Ondersteuning bij persoonlijke zorg: verplaatsing F	Elementair Fysiologische Functies
51	P04	DN01	Dan ga ik het even droog maken v	Voorkomen van uitglijden	6482 Regulering van de omgeving: comfort	Veiligheid
52	P04	DN01	Heb je diarree had? Ja? Vanochten	Beoordelen van diarree	0460 Diarree Beleid	Elementair Fysiologische Functies
53	P04	DN01	Nu alweer rustig met de darmen?	Beoordelen van darmfunctie	0430 Defecatiebeleid	Elementair Fysiologische Functies
54	P04	DN01	Ga je nog een keer een douche ma	Adviseren over veiligheid in huis	6482 Regulering van de omgeving: comfort	Veiligheid
55	P04	DN01	Gaan we je haren wassen?	Haarverzorging	1670 Haarverzorging	Persoonlijke verzorging en hygiëne
56	P04	DN01	Hoe lang heb je nu last van je ogen	Beoordelen van oogproblemen	1650 Oogverzorging	Complex Fysiologische Functies
57	P04	DN01	Dan ga ik de washandjes pakken al	Oogverzorging	1650 Oogverzorging	Complex Fysiologische Functies
58	P04	DN01	Komt de dietist nog bij je langs?	Beoordelen van voedingstoestand	1160 Bewaking van de voedingstoestand	Elementair Fysiologische Functies
59	P04	DN01	Ik ga het uitspoelen, mag jouw hoc	Haarverzorging	1670 Haarverzorging	Persoonlijke verzorging en hygiëne
60	P04	DN01	Eerst even je gezicht lekker doen zc	Wassen van het gezicht	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Persoonlijke verzorging en hygiëne
61	P04	DN01	Veel leuk. Ik zag op je rug ook al w	Beoordelen van huidconditie	3590 Bewaking van de huidconditie	Complex Fysiologische Functies
62	P04	DN01	Zal ik je achterkant doen?	Wassen van de rug	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Persoonlijke verzorging en hygiëne
63	P04	DN01	Ik ga nog even goed je billen doen.	Wassen van de billen	1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	Persoonlijke verzorging en hygiëne

Figure 4.4: Revised generated event logs including Domain classification.

As the picture illustrates, the output still contains hallucinations. For example, "Persoonlijke verzorging en hygiëne" is not a valid domain according to the Dutch NIC taxonomy. Yet, it still appears throughout the results related to several distinct interventions. Furthermore, the classification was rather inconsistent in quality. Domain classification accuracy would vary depending on the transcript. Therefore, the prompt needs to be extended to take this into account.

#### 4.1.6 Final Output

As the generation of event logs using an LLM is largely exploratory, this will be the final iteration of prompt revisions. This prompt will be used as the base for the event log that will in turn serve as input for the process models. At this stage, we accept that GPT-4 is unable to reach the stage of perfection regarding activity classification. Therefore, this is where we draw the line and will manually clean the data to make sure the process models have a solid foundation.

The final prompt is shown in the following code snippet:

```
return new List<ChatMessage>
{
    new ChatMessage(ChatRole.User, "These are the options
    for level 3 interventions: " + dutchNicInterventionsString),
    new ChatMessage(ChatRole.User, "These are the options
    for level 1 domains: " + dutchNicDomainString),
    new ChatMessage(ChatRole.User,
    "Extract the nursing activities from the following
    transcript and map each activity to one of the previously
    defined Dutch NIC taxonomy interventions."
    + "If no fitting match can be found between the extracted
    activity and the predefined interventions,
    set the activity and both classifications as \"No match\"
    while keeping the patient, district nurse, and referenced
    transcript excerpt in the output."
    + "Respond in Dutch by returning a csv formatted event log
    containing the following columns of the respective activity:"
    + "Patient number starting with P, " +
    "district nurse number starting with DN," +
    "the exact quote in the transcript excluding the timestamp
    and person saying it" +
    "the performed activity, " +
    "the level 3 intervention from the Dutch NIC taxonomy as
    previously defined, " +
    "Infer the level 1 domain from the specified level 3
    intervention within the constraints of the NIC taxonomy.
```

```

Limit the options to the predefined domains. " +
"Start timestamp, " +
"End timestamp " +

"Exclude all comments or remarks about the content."),
new ChatMessage(ChatRole.User, transcript),
new ChatMessage(ChatRole.User, "Ensure all the outputted
level 1 and level 3 classifications are included in
the response and are limited to the predefined taxonomy."),
new ChatMessage(ChatRole.User, "Check all outputted
classifications again and make sure each row explicitly
included patient, district nurse, quote, activity,
level 3 intervention, level 1 domain, start timestamp,
and end timestamp."),
};

```

This prompt again has some changes compared to the last prompt. The changes primarily serve as boundaries for the prompt, ensuring that the model is basing its classification on the provided options, and improving consistency in both the classification of the data and the completeness of the data. The biggest improvement was the addition of instructing the model to check the output again after the initial classification round. Instructing the model to check its results tended to show more consistently correct results. On top of that, a new message emphasizing the inclusion of every data column to ensure no data is missing. The results are illustrated in figure 4.5.

This final output has fewer classification errors compared to the previous result, while no data is missing from any columns. However, even if the event log has improved, it is not perfect. Hallucination does not occur, but there are many instances of wrong classification on the domain level. Furthermore, the inconsistency problem is still present. For example, the same intervention is identified within the same transcript, but receives different domain classifications.

One additional and unexpected anomaly during the event log generation was specifically one transcript in particular not returning any output. The transcripts were processed in sets of three as to prevent reaching the token limit for the GPT-4 model. This way, the token limit was never reached during the event log genera-

## Results

1	Patient	District Nurse	Quote	Activity	Level 3 Intervention	Level 1 Domain
2	P01	DN01	Ja, je merkt meteen dat het stijver w	Bevorderen van	0200 Bevordering van lichamelijke oefe	Elementair Fysiologische Functies
3	P01	DN01	Niet te veel bewegen. Goed water drii	Voorlichting gev	5616 Voorlichting: voorgeschreven medi	Gedrag
4	P01	DN01	En dan een beetje de oefeningen. Ik b	Bevorderen van	0200 Bevordering van lichamelijke oefe	Elementair Fysiologische Functies
5	P01	DN01	Nou u tilt hem nog wel gemakkelijk o	Beoordelen van	0970 Verplaatsing bij beperkte mobilite	Complex Fysiologische Functies
6	P01	DN01	En wat je doet, je wordt benauwd he.	Beoordelen van	3350 Bewaking van de ademhaling	Complex Fysiologische Functies
7	P01	DN01	Daar komt die, het is wel mooi geneze	Beoordelen van	3590 Bewaking van de huidconditie	Veiligheid
8	P01	DN01	Goedzo. Ja, mag u eerst uw voeten. H	Helpen met pers:	1800 Ondersteuning bij persoonlijke zor	Elementair Fysiologische Functies
9	P01	DN01	Ja goed schoonmaken daartussen.	Verzorging van c	1660 Voetverzorging	Elementair Fysiologische Functies
10	P01	DN01	Nog even afspoelen. Is alles er af?	Helpen met was:	1801 Ondersteuning bij persoonlijke zor	Elementair Fysiologische Functies
11	P01	DN01	Gelijk even naar de navel kijken?	Verzorging van c	1875 Verzorging van een aan/afvoerbuis	Elementair Fysiologische Functies
12	P01	DN01	Zo zal ik verder met de achterkant dr	Helpen met drog	1801 Ondersteuning bij persoonlijke zor	Elementair Fysiologische Functies
13	P01	DN01	Wanneer zijn de laatste keer de krulle	Beoordelen van	1670 Haarverzorging	Elementair Fysiologische Functies
14	P01	DN01	Dan gaan we even de bodylotion doen.	Toepassen van h	3583 Huidverzorging: acceptorsite L	Veiligheid
15	P01	DN01	Ja, steeds moeilijker met de onderbro	Helpen met klec	1802 Ondersteuning bij persoonlijke zor	Elementair Fysiologische Functies
16	P01	DN01	Wilt u met een rollator naar de slaapk	Bevorderen van	0970 Verplaatsing bij beperkte mobilite	Complex Fysiologische Functies
17	P01	DN01	Ja, ik zal wel even de pleisters eraf h	Verzorging van e	3660 Wondverzorging	Veiligheid
18	P01	DN01	Nee deze doet het niet goed.	Ondersteuning b	1640 Oorverzorging	Elementair Fysiologische Functies
19	P01	DN01	Nou kom op jongens, zo ja ja.	Ondersteuning b	4974 Bevordering van de communicatie:	Gedrag
20	P01	DN01	Zal ik de beroep meenemen en de san	Helpen met klec	1802 Ondersteuning bij persoonlijke zor	Elementair Fysiologische Functies
21	P01	DN01	Dan had ik mijn oortjes zeker nog niet	Ondersteuning b	1640 Oorverzorging	Elementair Fysiologische Functies
22	P01	DN01	Nou misschien net nog wel voor twaal	Bevorderen van	0200 Bevordering van lichamelijke oefe	Elementair Fysiologische Functies
23	P01	DN01	Leuk om om erbij te zijn dat u gebeld	Emotionele ond	5270 Emotionele ondersteuning	Gedrag
24	P02	DN01	Ja, voor de medicijnen.	Medicijnen klaar	2300 Toediening van geneesmiddelen	Complex Fysiologische Functies
25	P02	DN01	Ja hartstikke goed ga ik die er even b	Medicijnen pakk	2300 Toediening van geneesmiddelen	Complex Fysiologische Functies
26	P02	DN01	Nou, ik zal ze zo meteen op de juiste	Medicijnen opru	2300 Toediening van geneesmiddelen	Complex Fysiologische Functies
27	P02	DN01	Even kijken hoeveel bloedverdunners	Medicatie doser	2300 Toediening van geneesmiddelen	Complex Fysiologische Functies
28	P02	DN01	Zo. Haal ik ze ook voor u uit de verpak	Medicijnen uit v	2300 Toediening van geneesmiddelen	Complex Fysiologische Functies
29	P02	DN01	Er ligt nog. Hoeveel temazepam wilt u	Medicatie doser	2300 Toediening van geneesmiddelen	Complex Fysiologische Functies
30	P02	DN01	Dan heeft het lichaam het toch mooi e	Wondverzorging	3660 Wondverzorging	Complex Fysiologische Functies
31	P02	DN01	Nou, wij laten jullie lekker kapperen.	Sociale interacti	5100 Bevordering van de socialisatie	Gedrag
32	P02	DN01	Nou, goed weekend. Doei dag.	Afscheid nemen	5000 Opbouw van een therapeutische re	Gedrag
33	P03	DN01	Had u ook handschoenen hier liggen o	Zoeken naar har	No match	No match
34	P03	DN01	In de wasmachine of in de?	Vragen waar de	No match	No match
35	P03	DN01	Het water is ook lekker warm, hè?	Informeren naar	No match	No match
36	P03	DN01	En hoe is het verder?	Informeren naar	1160 Bewaking van de voedingstoestand	Complex Fysiologische Functies
37	P03	DN01	En wat heeft u dan precies met de nie	Informeren naar	2020 Bewaking van de elektrolytenhuish	Complex Fysiologische Functies
38	P03	DN01	En merkt u dat ook lichamenlijk?	Informeren naar	1160 Bewaking van de voedingstoestand	Complex Fysiologische Functies

Figure 4.5: Final revision generated event logs.

tion process. However, the 18th transcript returned an empty event log, essentially meaning no activities could be extracted. Furthermore, no message was returned about the point of failure. Initially, this occurred when grouping the transcript in threes as input for the model, but repeated itself when the transcript was the sole transcript. After debugging and isolating variables, the cause appeared to be the relative location of appending the transcript to the prompt in the code. More specifically, the solution involved moving this line `new ChatMessage(ChatRole.User, transcript),` to be the last ChatMessage object of the prompt. Even though the problem has been identified, the root cause of the problem remains unknown. As a result, transcript 18 is excluded from the event log to keep the quality of the output consistent quality, thus a consistent analysis of the results.

Lastly, the final generated event log has been manually enriched by including the level two classification called class. The rationale behind this extension of the log is to add one additional layer of potential granularity, having the full granularity of the NIC taxonomy. Investigating the data on different levels could lead to different knowledge and insights gained, where the level two classification could serve as a middle ground when the domain is too abstract, and the intervention

is too fine-grained. The class granularity level will act as an intermediate level of granularity, potentially revealing patterns that are not visible on the domain level, or too obscure on the intervention level.

#### 4.1.7 Event Log - Emerging Patterns

Before any mathematical analyses, several patterns can be identified in the final event log. On a domain level, one particular domain is predominantly present in the event log. Almost 65% of the total activities belong to the domain of Elementary Physiological Functions, consisting of activities which relate to very basic needs such as washing, exercising, and clothing. In absolute numbers, this means that 217 activities out of the total 314 activities are classified in an intervention belonging to this domain.

The domain that occurs second most frequently is the Complex Physiological Functions, a domain about activities such as drug management and wound care by the nurse. Approximately 16% of the activities are mapped to interventions belonging to this domain, occurring 53 times out of 339 activities.

The third most occurring domain is the Behavioral domain. As the name implies, it is mostly related to emotional or cognitive interventions. Activities falling under this domain occur 50 times in total across the event log, having a relative frequency of about 15%. While the NIC taxonomy defines four more domains, those four have no significant representation in the event log, or even at all. The four domains with the lowest frequency in the event log have a combined relative frequency of about 5%. The frequencies of the domains are summarized in figure 4.1.

Domain	Absolute Frequency	Relative Frequency
Elementary Physiological Functions	207	64.09%
Complex Physiological Functions	51	16.10%
Behavioral	48	14.86%
Health System	8	2.48%
Safety	5	1.55%
Family	3	0.93%

**Table 4.1:** Domain frequencies in the event log.

At the more fine-grained class level, we can identify more specific types of interventions. Here, the core types of classifications are revealed instantly as 41.49% of the events are classified within the class "F Self-Care Facilitation". Interventions



that fall under this class relate to providing or assisting clients with routine activities of daily life. The second and third most occurring classes are "B Elimination Management" with 11.45% and "H Drug Management" occurring in 10.53% of the events respectively. Elimination management is related to maintaining bowel and urinary elimination patterns, while Drug management is to ensure the client is taking the correct medication with the right amount. An overview of these results is shown in figure 4.2.

Class	Absolute Frequency	Relative Frequency
F Self-Care Facilitation	134	41.49%
B Elimination Management	37	11.45%
H Drug Management	34	10.53%

**Table 4.2:** Top 3 class frequencies in the event log. See table A.1 in the appendix for the complete table.

Finally, on the most detailed level of interventions the previously discussed patterns are visible. The dominant presence of the self-care facilitation class, and thus the elementary physiological domain, is revealed. The top three interventions, starting with the most frequently occurring intervention, are: "1802 Self-Care Assistance: Dressing/Grooming", "1801 Self-Care Assistance: Bathing/Hygiene", and "1610 Bathing". These interventions have a very similar frequency, having a combined frequency of 25.08% across the complete event log. These top three results have been compiled in the table shown in figure 4.3.

Intervention	Absolute Frequency	Relative Frequency
1802 SC Assistance: Dressing/Grooming	28	8.67%
1801 SC Assistance: Bathing/Hygiene	27	8.36%
1610 Bathing	26	8.05%

**Table 4.3:** Top 3 class interventions in the event log. (See table A.2 in the appendix for the complete table.)

Specific activities have a very strong presence in the event log, while the remaining activities tend to have a lower frequency as a consequence generally giving the distribution of these activities a very "long" tail. This is especially noticeable on the intervention level where the granularity level is fine-grained enough to express that tail.

#### 4.1.8 Output Validation

To evaluate the quality of the output of the generated event log, a domain expert validated the results. Following the consult, the event log was corrected accord-

Marked			Total	
Initial	Changed	Deleted	Before	After
63	27	16	339	323

**Table 4.4:** Changes to the event log as a result of validation.

ingly which in turn served as the basis for the process models. The scope of the validation was limited to checking if the level 3 intervention matched the quote and activity as extracted from the transcript. The quotes and activities could be validated on a different scale by checking whether the extracted activities were rightfully considered activities for nursing. Both of these were considered when validating the output.

As preparation for the validation discussion, we did a manual check as a first round. The first round served as a preliminary check, seeing whether any obvious mistakes could be identified purely on rational inspection. For example, one of the extracted quotes related to the client offering drinks to the nurse. GPT-4 classified this as an activity related to advice on food intake, which is incorrect considering the context. As a result of this first validation round, we marked 63 out of 339 events with a questionable intervention classification, meaning approximately 18.58% of the event log was marked. However, there are two caveats. First, there were many duplicates, meaning that some questionable classifications were at least consistent, and one single error had a relatively large impact on the accuracy of the classification as a whole. Second, some of the marked events were due to inexperience with the taxonomy and not understanding the nuances of some interventions. One prominent example was the distinction between "1610 Bathing/Showering" and "1801 Self-Care Assistance: Bathing/Hygiene".

The marked event log served as the starting point for the discussion with the domain expert, for which we had scheduled one hour. Seeing as the time was limited, we focused on the 63 marked events instead of inspecting every single event. Even then, it meant we had less than one minute per event on average. At the end of the evaluation, a total of 43 events had to either be deleted due to not being relevant to nursing interventions, or altered due to having a wrong initial classification. More specifically, 15 events had to be deleted entirely, while the remaining 27 had their intervention classification changed. This means that out of the 63 marked events roughly 66.67% had to be either changed or deleted. On the scale of the complete event log, changing or deleting 43 events means approximately 87.65%

Total	Overlap	Identical	Similarity Rate
55	19/55	11/19	42.11%

**Table 4.5:** Compilation of reproducibility sample.

was classified correctly. The changes as a result of the validation are compiled in table 4.4.

Most of the events that were marked were activities that could not be appropriately classified within the constraints of the NIC taxonomy. While this leaves a total of 277 events unchecked by the domain expert, this has to be accepted, respecting the domain expert's effort. These initial 63 marked events were the "low-hanging fruits", the classifications that could be identified without much context or domain knowledge. To evaluate the other classifications, more time should be allocated to the evaluation of the output, which requires multiple sessions with domain experts.

#### 4.1.9 Repeatability

One of the larger challenges with AI models is that they are generally considered black boxes: there is a clear input, and there is a clear output. However, what happens during that transformation is not transparent nor reliably reproducible due to their non-deterministic nature. This means that given the same input, a different output may be generated, as is the case when executing the same prompt with the same transcripts multiple times.

Certainly, there are specific quotes that the model understands as an activity relatively consistently, but others can appear more random. Note that even when the quote is the same, the exact description of the activity may vary, as well as the classification. One such comparison is shown in figure 4.6. This figure compares two newly generated event logs of the first three transcripts contained within the same prompt. The figure does not include the entire output, but rather a fixed comparison between the first 55 events generated. Color coding has been added to highlight the differences between the two. Yellow marked rows mean they refer to the same quote in the transcript, while red font means the classification varies between the respective events.

The results of the comparison have been compiled in table 4.5. Out of the 55 events shown, 19 events are based on the same quote from the transcript, thus these 19 events were included in both event logs. From those 19 events, the model

classified 11 performed activities differently. Similarly, 11 out of 19 domain classifications also varied between the two outputs, leading to a 42.11% similarity rate between the two events shown. Interestingly, even given the same quote and sometimes even activity, the classifications could vary. In particular, note how row 31 on the left side and the respective row 28 on the right side of the figure have the same quote, the same activity extracted, but different classifications. The same goes for row 34 on the left side, and the respective event on row 35 on the right side of figure 4.6. These differences are very strong examples of how the output of the model is not always the same, even given the exact same prompt.

On a similar note, as 19 out of 55 events were based on the same quote, this means 36 of the generated events are based on different quotes. Expressed in a percentage, this means the degree to which the event logs are similar is 34.55%, as this is the percentage of extracted quotes present in both generated event logs. In other words, 65.45% of the events are based on a different quote from the transcript. The degree to which these logs vary is therefore substantial and can have a very large impact on the models that follow.

#### **4.1.10 Comparison of Full Transcript Classification**

As the results from the previous section were very different as summarized in table 4.5, an additional analysis has been conducted on a new, larger set of classifications.

Instead of instructing the model to cherry-pick quotes and activities from the transcript, the model was instructed to explicitly include every line from the same set of transcripts as the previous section. The rationale behind explicitly including all lines from the transcripts is to reduce the randomness as a result of the LLM, where it would sometimes pick up on specific sentences and activities, while at other times it would not. This way we attempt to isolate the accuracy of the intervention classification as a result of the prompt, instead of also measuring the extraction of activities. The prompt used for this experiment is shown below.

Note how the prompt has several differences compared to the last prompt. In particular the removal of the domain classification, and the addition of single-shot prompting by adding an example of a correct intervention classification. The single-shot prompting aids in providing the full classification, instead of merely parts of the classification, an issue that did not arise in the previous prompting. Moreover, to deal with the increased length of the output, the transcripts had to be



cut into smaller segments not to exceed the token limit. The transcripts were divided into smaller transcripts that were about 50 lines, depending on the length of sentences as a guideline for a more equal distribution of tokens between API calls.

```
return new List<ChatMessage>
{
    new ChatMessage(ChatRole.User, "These are
    the options for level 3 interventions: "
    + dutchNicInterventionsString),
    new ChatMessage(ChatRole.User,
    "Iterate over include each line in the
    transcript and classify map the activity
    performed in each line to one of the
    previously defined Dutch NIC taxonomy
    interventions. Do not exclude any lines."
    + "If no fitting match can be found between the
    extracted activity and the predefined
    interventions, set the activity and
    both classifications as \"No match\" while
    keeping the patient,
    district nurse, and referenced transcript
    excerpt in the output."
    + "Respond in Dutch by returning a csv formatted
    event log containing the following columns of
    the respective activity:"
    + "Patient number starting with P, " +
    "district nurse number starting with DN," +
    "The patient or district nurse saying the line" +
    "the exact quote in the transcript excluding
    the timestamp and person saying it" +
    "the performed activity, " +
    "the complete level 3 intervention from
    the Dutch NIC taxonomy as previously
    defined, for example "1801
    Ondersteuning bij persoonlijke zorg:
```

```

    wassen/hygiene", " +
    "Start timestamp, " +
    "End timestamp "),

    new ChatMessage(ChatRole.User, transcript),
    new ChatMessage(ChatRole.User, "Ensure all lines
    in the transcript are included, and the outputted
    level 3 classifications
    are included in the response and are limited to
    the predefined taxonomy."),
    new ChatMessage(ChatRole.User, "Make sure each row
    explicitly includes patient, district nurse,
    quote, activity, level 3 intervention, start
    timestamp, and end timestamp."),

};

```

All the transcripts were run through three times, meaning the output was three completely new event logs based on this revised prompt. The results and the comparison between runs have been compiled in tables 4.6, 4.7, and 4.8, portraying the comparisons for each run in a single table.

Interestingly, the number of events that received a classification from the model (accuracy aside) is a multitude higher than in the original experiment where the model was instructed to extract activities. Whether this is the result of a modified prompt or the smaller transcripts is unclear, although the cause being prompt-related is a starting point.

	Total	Classified	No Match	Run #2		Run #3	
				Overlap	Different	Overlap	Different
<b>Run #1</b>	5497	1187	4310	3994	1461	3911	1505

**Table 4.6:** Comparing results of run 1 with run 2 and 3.

	Total	Classified	No Match	Run #1		Run #3	
				Overlap	Different	Overlap	Different
<b>Run #2</b>	5455	1402	4053	3867	1630	3926	1490

**Table 4.7:** Comparing results of run 2 with run 1 and 3.

Between the runs, there is a noticeable variance across many aspects of the event

	Total	Classified	No Match	Run #1		Run #2	
				Overlap	Different	Overlap	Different
<b>Run #3</b>	5416	1169	4247	3261	2236	3111	2344

**Table 4.8:** Comparing results of run 3 with run 1 and 2.

	Domain							Total Classified		No Match	Total Events
	1	2	3	4	5	6	7	Real	Hallucination		
<b>Run #1</b>	724	211	116	5	2	0	1	1059	128	4310	5497
<b>Run #2</b>	715	314	192	2	3	0	1	1227	175	4053	5455
<b>Run #3</b>	607	211	127	3	0	0	2	950	219	4247	5416

**Table 4.9:** Domain Frequency of every run with each transcript line.

log. For example, the difference in hallucination varies per run. Run 1 has the least amount of hallucination, while run 3 has almost double the number of hallucinated classifications. On a similar note, the overlap between classifying quotes similarly is significantly different in run 3, as shown in the lower numbers in the overlap columns in table 4.8. This is a striking example of how a large language model (GPT-4 specifically in this case) may produce markedly different output even when all else is equal. To provide an overview of all classifications, the results have been compiled in table 4.9.

## 4.2 DECLARE

As we hypothesized the declarative approach to outperform the imperative approach, the initial section will shed light on the performance of the DECLARE model. The DECLARE process model based on the generated event log as described in the previous section is shown in figure 4.7. This model was generated using the ProM plugin Declare Miner in ProM 6.13. It has been recreated in Draw.io to modify the existing elements of the model to restructure the model for readability, and enrich the model using color coding based on the domain level of the identified interventions. Additionally, the activities have been restructured to be read from top to bottom, and from left to right to match the ordering of the domain levels.

Each unique activity performed by the nurse that could be extracted based on the transcript is represented in the model, and the model is constrained based on the following parameters:



1. Support Threshold: 0.95
2. Confidence Level: 0.33
3. Interest Level: 0.25
4. Negative Constraints: True

The model contains limited explicit constraints that were deduced from the event log. Still, this may illustrate the nature of nursing as the process model is very unstructured. The only constraints in the model are related to interventions 1800, 1801, and 1802. Intervention 1801 has the constraint of "Not chain succession" with intervention 1802, while intervention 1802 has the same constraint as intervention 1800. A "Not chain succession" constraint means that in a situation where activity A is constrained by activity B (i.e. activity A points to activity B) then every time activity B is executed, A cannot directly precede B. In other words, when intervention 1801 is performed, activity 1802 must never be performed immediately after, meaning that a dressing should not occur directly after washing. The same goes for the interaction between interventions 1802 and 1800. Similarly, whenever intervention 1800 was performed, activity 1802 should not precede that execution.

By pure logic, these are interesting statements which understandably hold true in several scenarios, but not every scenario. For example, 1801 (washing) not taking place right before 1802 (dressing) is a strong example. For washing to take place, it is likely the client has to undress to some extent. However, if the transcript does not contain such verbalized actions, then the event log will not contain them either. As the nurse takes care of the client, some interventions may be communicated through body language. This means there is no verbal confirmation of the intervention, because the physical presence and context are sufficient to confirm the need. More specifically, if the client knows washing is about to happen, they may prepare themselves by understanding the need to undress and this will not get caught by the audio recording. Using the same parameters for the model, we can use the different granularity levels of the event log for process discovery on each level of the NIC taxonomy.

Figure 4.8 shows the DECLARE model when considering the class level classification as the activities, and figure 4.9 shows the same concept on the domain level. Interestingly, there are very distinct differences between the models. Naturally, the models get smaller when the activities become more abstract, as there are fewer distinct activities. Furthermore, the number of constraints included in the

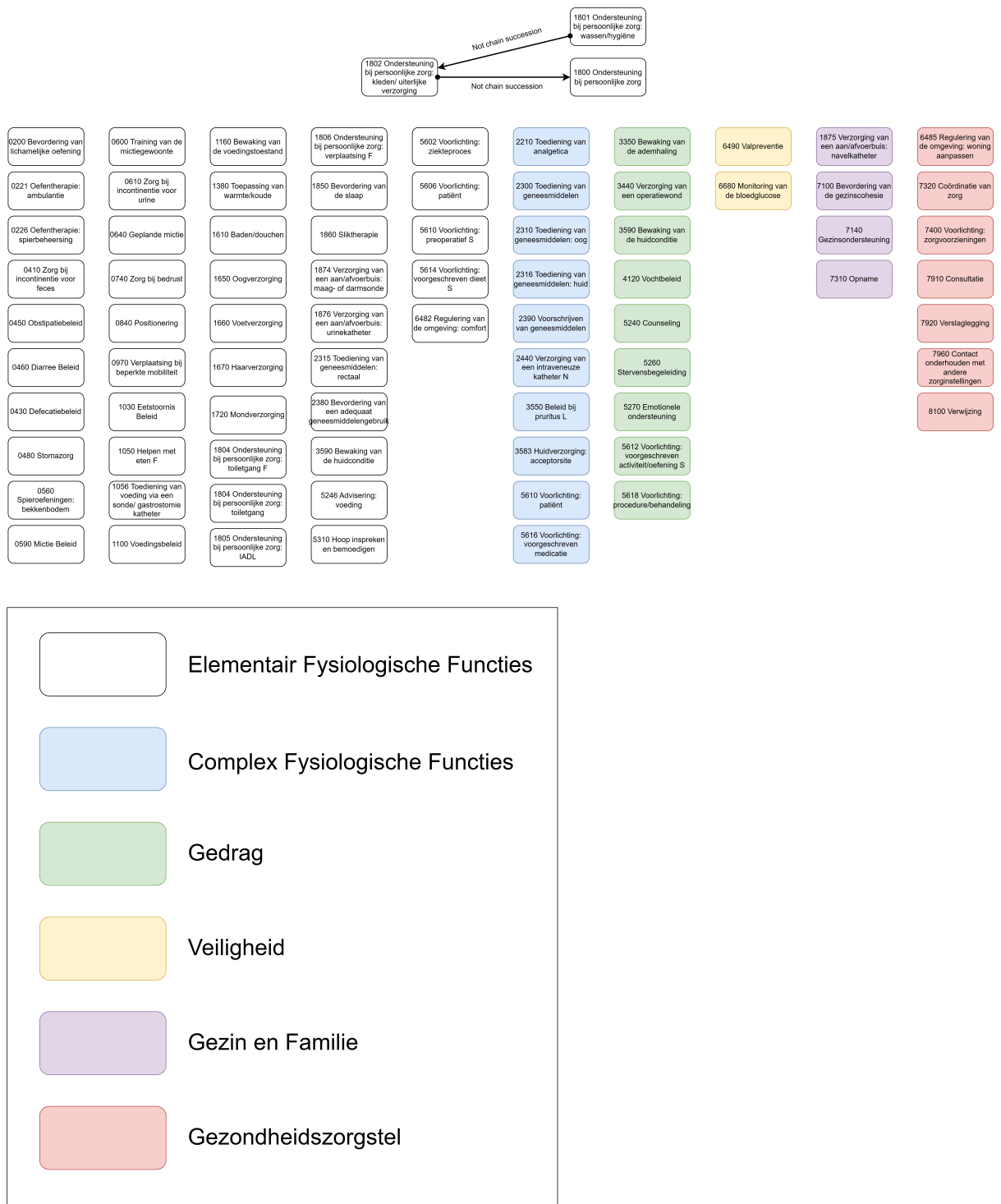


Figure 4.7: Generated DECLARE model on intervention level.

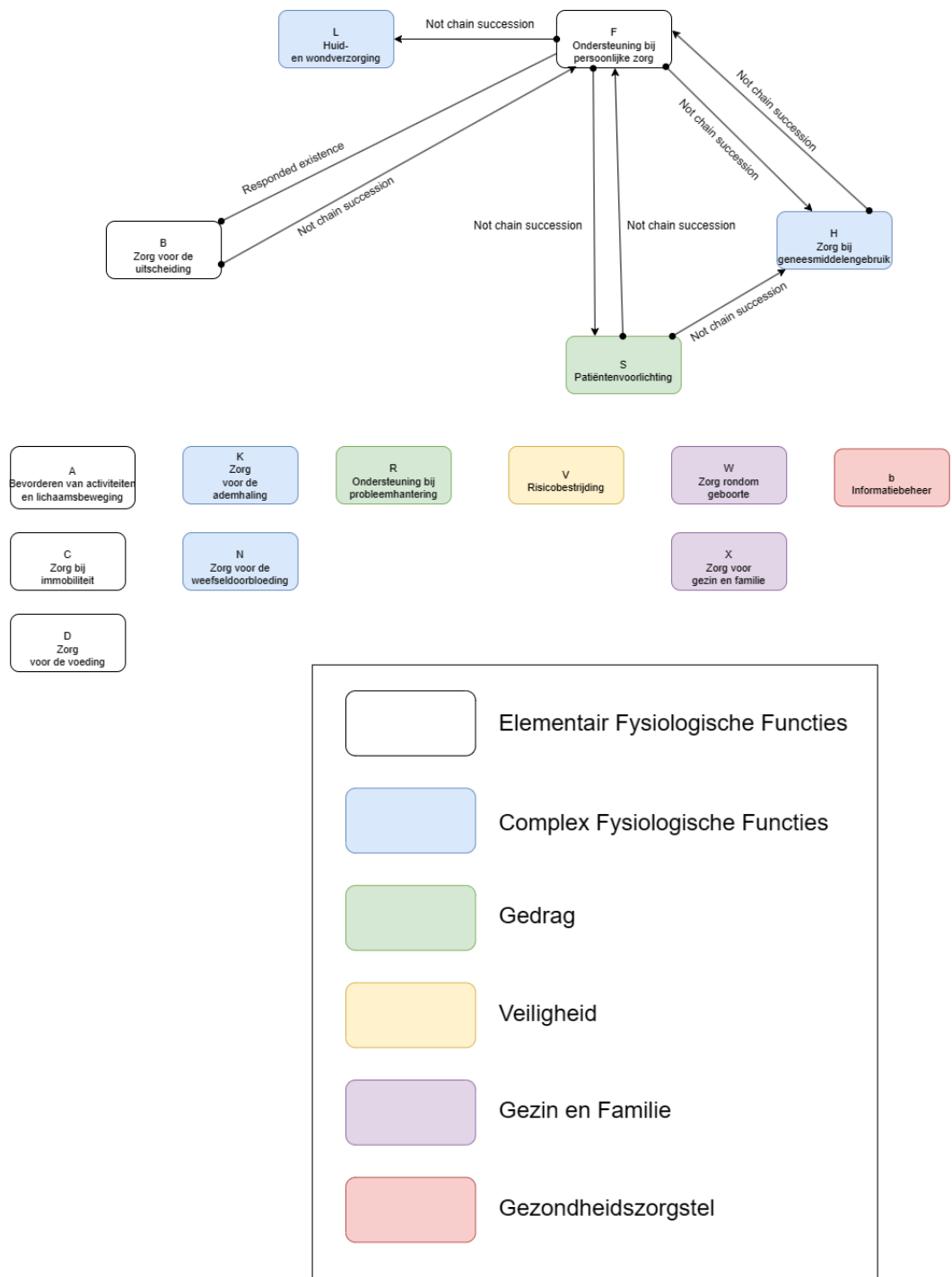


Figure 4.8: Generated DECLARE model on class level.

class-level model is significantly higher than the other models. This is likely due to the number of unique classifications on their respective levels.

For example, on the intervention level there are many unique intervention classifications in the model. Consequently, it is difficult to ascertain a strong relationship between activities, because the absolute frequency of these activities is rather low. On the class level, the absolute frequency of activities and the number of unique activities are more balanced. This means that there are fewer unique sequences, but the absolute frequency balances it out providing enough support for the constraint. However, the balance is skewed again at the domain level, where there are only six unique activities. Due to the absolute frequency of the "Veiligheid", "Gezin en Familie", and "Gezondheidszorgstel" there is not enough support to ascertain any constraints on these activities, while there is support for constraints among the most frequent domains.

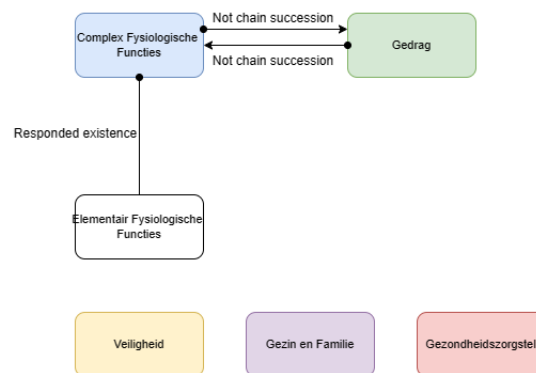
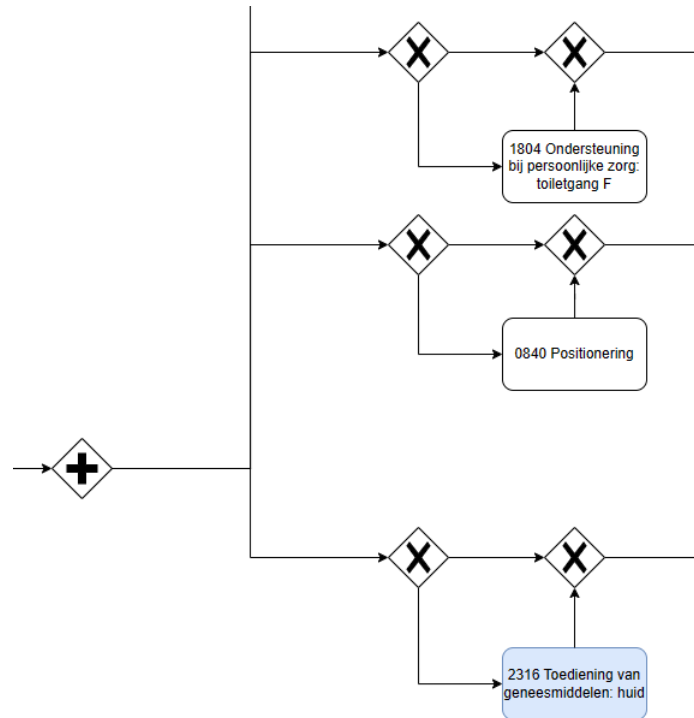


Figure 4.9: Generated DECLARE model on domain level.

### 4.3 BPMN

Using the BPMN Miner plug-in implemented in ProM 6.13 the model shown in figure 4.11 has been generated using the Inductive Miner algorithm. The model in the figure is a model recreated in Draw.io after which it has been restructured for less overlapping elements and additional color coding to enrich the model. This model is divided into three clusters of sequences, which can be distinguished based on how structured they are modelled. Initially, the first several activities have a clear structure to them, there is an OR gateway where a choice is made between multiple activities. However, after this OR gateway starts a giant sequence consisting of one pattern that has a very strong presence. More specifically, this pattern consists of two OR gateways, and one activity. In the context of the model, this means that the

activity either is executed or not without affecting the flow of the process model.



**Figure 4.10:** Frequently recurring BPMN pattern.

The complete model consists of 64 activities, and 109 gateways, summing up to a total of 173 elements in the model. Due to the aforementioned pattern of two OR gateways and one activity, the number of gateways is somewhat skewed, and the current iteration of BPMN has no other syntax to concisely model this behavior.

Evaluating the model's performance by calculating the fitness, precision, and simplicity as defined in the methods section leads to the following scores, which will be summarized along with the other model evaluation metrics in table 4.10. These metrics have been calculated by using a combination of methods. Fitness and Simplicity have been calculated manually, while Precision is calculated using ProM and its plug-ins. ProM does not have a plug-in that can calculate evaluation metrics on a BPMN model, but does have a plug-in to calculate precision for petri nets. Therefore, a petri net was generated by ProM using the plug-in "Mine Petri net with Inductive Miner", using the same inductive miner algorithm to instead generate a petri net. To calculate the precision on this petri net, the plug-in "Check Precision based on Align-ETConformance". The following settings were configured for this plug-in:

1. Representation: Ordered

2. Algorithm: 1-Align Precision
3. Classifier: Event Name
4. Purpose of replay: Measuring fitness
5. Penalize improper completion: Yes
6. Algorithm: Splitting replayer
7. Maximum explored states: 2000
8. Set all costs above to: 1 (For both transitions and events)

Configuring the plug-in with these settings and calculating the metrics accordingly, the resulting metrics are the following:

1. Fitness: 0.69 ( $\frac{18}{26}$ )
2. Precision: 0.06
3. Simplicity: 0.52 ( $\frac{64}{64+69}$ )

These calculated metrics have also been compiled in table (4.10), along with the compiled metrics of the other models.

### 4.3.1 Imperative Class- and Domain Models

To further investigate the performance of BPMN, two additional models have been created based on the same event log. These models distinguish themselves from the previous model by looking at more abstract granularity levels. Instead of looking at interventions as activities, these models consider class and domain respectively.

Figure 4.12 illustrates the model created using the same plug-in as described in the previous section, but instead of considering interventions as activities, it contains classes (i.e. one abstraction level higher). The model is considerably smaller, as following the taxonomy multiple interventions belong to a single class. The evaluation metrics have been calculated accordingly and listed below. Compared to the BPMN model on the intervention level, there is a single improvement in metrics that stands out, namely the precision metric has increased by a notable margin. Precision increased from 0.06 to 0.23, while scoring slightly lower on the other two metrics. This is caused by a combination of factors. First of all, fitness and precision tend to be opponents playing a zero-sum game. This means that if one metric "wins" (i.e. improves) the other metric "loses" (i.e. worsens). The precision of the previous model was very low, so there was much to gain while the model was able

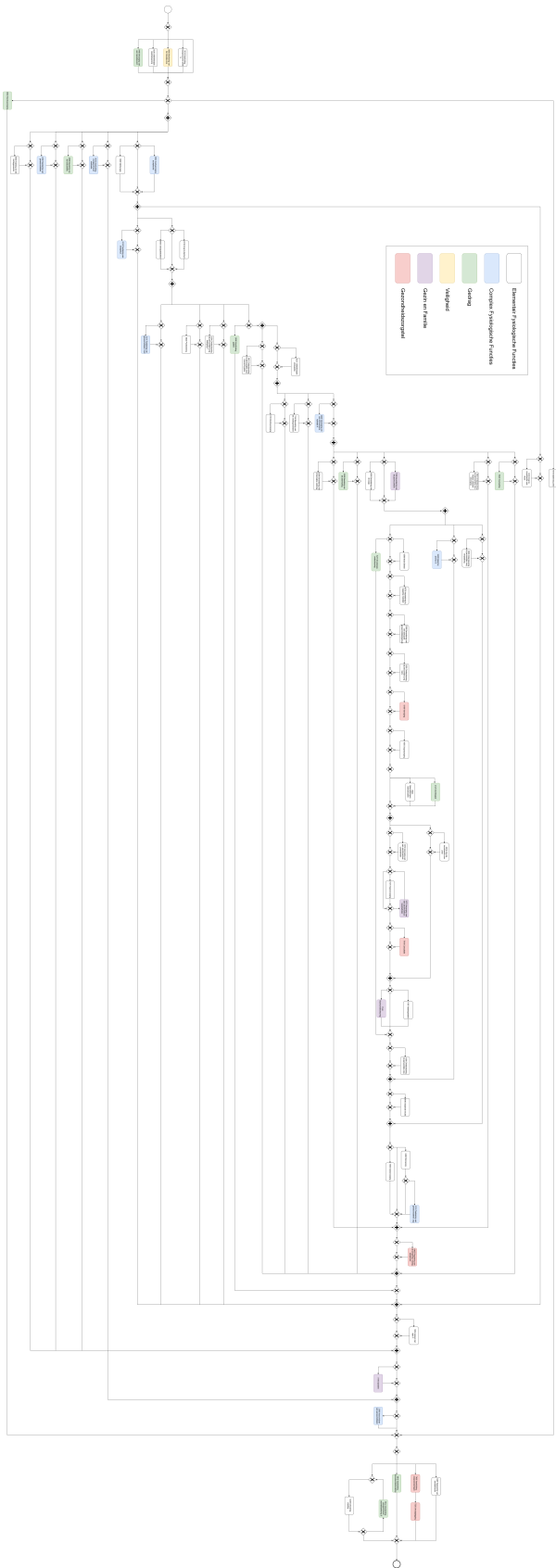


Figure 4.11: BPMN model on intervention level.

to keep fitness relatively unscathed.

While the decrease in simplicity is marginal, it can still be explained by understanding the formula used to calculate it and knowledge of the models. The proposed simplicity formula compares the number of distinct activities in the *model*, and the number of distinct activities in the *log*. The intervention model left out several activities due to the model already being convoluted. Furthermore, seeing as there are many distinct interventions in the NIC taxonomy (approximately 650), the model was still relatively simple by leaving out multiple interventions.

The model on class level does a similar action, where it still leaves out several classes from the model, but leaves out fewer compared to the intervention level model. However, exactly because it leaves out fewer classes, it is relatively more complex. This is why there is a very minor decrease in simplicity.

1. Fitness: 0.65
2. Precision: 0.23
3. Simplicity: 0.53

Figure 4.13 shows the same concept, but on the domain level. The model is even more compact, having only six distinct activities. Similar to the class-level model, the evaluation metrics have been calculated on this model. The trends outlined in the paragraph about the class model are also applicable to this model, as again the fitness and simplicity are slightly lower, while the precision shows considerable increase. The reasons for these changes are the same as described for the class-level model.

1. Fitness: 0.62
2. Precision: 0.45
3. Simplicity: 0.50

## 4.4 BPMN-D

BPMN-D as a notation is a superset of the BPMN notation where it has been extended using DECLARE concepts. As such, the model shares many similarities with the previously illustrated BPMN model. The resulting BPMN-D model is shown in figure 4.14. Due to the BPMN and BPMN-D models having a similar foundation in syntax, the model at large is structured in a similar fashion too. In





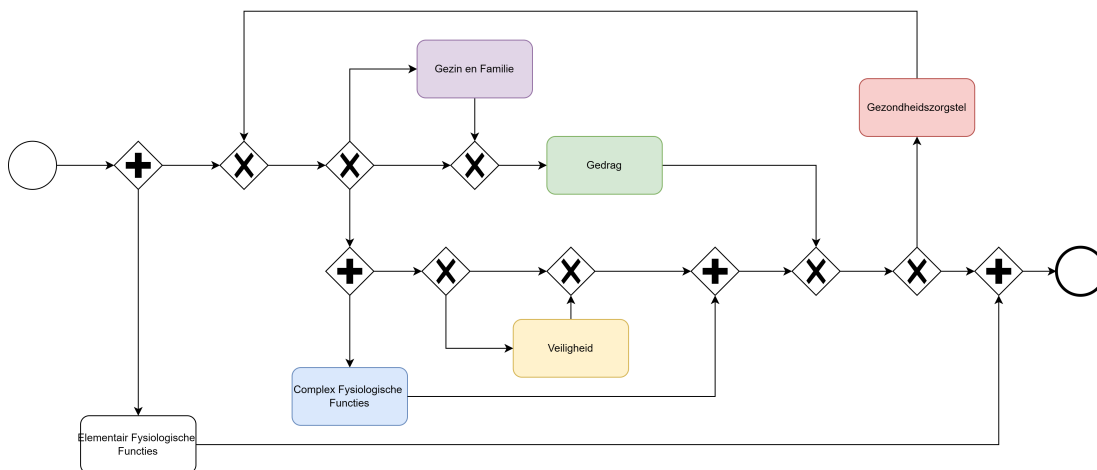


Figure 4.13: BPMN model on domain level.

particular, the same three segments are visible which were also visible in the BPMN model. The very start appears structured, while the larger middle segment appears more convoluted and complex, as this is where the ad hoc nature of nursing makes its presence known. Then finally, at the end, there is a somewhat structured block once again.

However, even though the structure at large appears similar, some of the specific syntaxes have been changed to handle specific model behavior. Specifically, the pattern mentioned in the BPMN section describing the two OR gateways and a single activity has been improved visually while remaining semantically the same as suggested by the original authors [52]. For example, the first activity element in the BPMN-D model contains multiple interventions, which in BPMN is represented by modelling these interventions separately and encapsulating them within two OR gateways. In BPMN-D all of these elements are contained within this single activity element, semantically meaning that exactly one of the interventions listed in the activity element must be executed before continuing in the model.

An additional syntactical difference is the activity elements belonging to a set of horizontal lines on the arrows between elements. They are similar syntactically and semantically, but especially semantically there is an important distinction. The former enforces exactly one of the listed interventions to be executed. On the other hand, the latter extends this behavior by adding the option to perform zero activities, completely skipping the execution of the listed interventions.

Having established the difference in syntax and semantics, the evaluation metrics result in the following numbers:

	<b>Fitness</b>	<b>Precision</b>	<b>Simplicity</b>	<b>Total Score</b>
<b>BPMN</b>	0.69	0.06	0.52	1.27
<b>BPMN-D</b>	0.69	0.06	<u>0.57</u>	<u>1.32</u>

**Table 4.10:** Aggregated metrics for each process model.

1. Fitness: 0.69 ( $\frac{18}{26}$ )
2. Precision: 0.06
3. Simplicity: 0.57 ( $\frac{52}{52+69}$ )

Note that the fitness and precision of the BPMN and BPMN-D models are the same, with the only difference being simplicity. This is due to the way the BPMN-D model is created. The blocks that have been modified as a result of the BPMN extension were ways to represent the same concept more simple. This means that semantically the model is the same, and the only difference is in syntax. In short, this means the resulting fitness and precision will always be the same between the two. Simplicity is the only metric that has changed, with an increase of 0.05 compared to the BPMN model. This is caused by the change in notation where the frequently occurring pattern in BPMN of two XOR gates and one activity was replaced by a single element. Again, the complete metrics for both models are displayed in table 4.10.

#### 4.5 SRQ1: What are the benefits and drawbacks of using Declarative Process Modelling?

Declarative process modelling emphasizes constraints over explicit sequences, fundamentally differing from imperative process modelling approaches. This is no different for DECLARE. With this comes a set of strengths and drawbacks. DECLARE's strength lies in its ability to express relationships between two activities, as this is what DECLARE constraints are. While conceptually this is possible in imperative process modelling languages, DECLARE can express this very clearly and concisely by drawing a visual line between activities. This means that on the scale of individual activities with less clear sequencing, DECLARE can communicate meaningful information that cannot be drawn from an imperative process model. One simple example is the RespondedExistence(A, B) constraint template where an occurrence of A means B has to occur within the same trace, but an occurrence of B does not imply an occurrence of A. This notion is elegantly expressed

#### 4.5 SRQ1: What are the benefits and drawbacks of using Declarative Process Modelling?

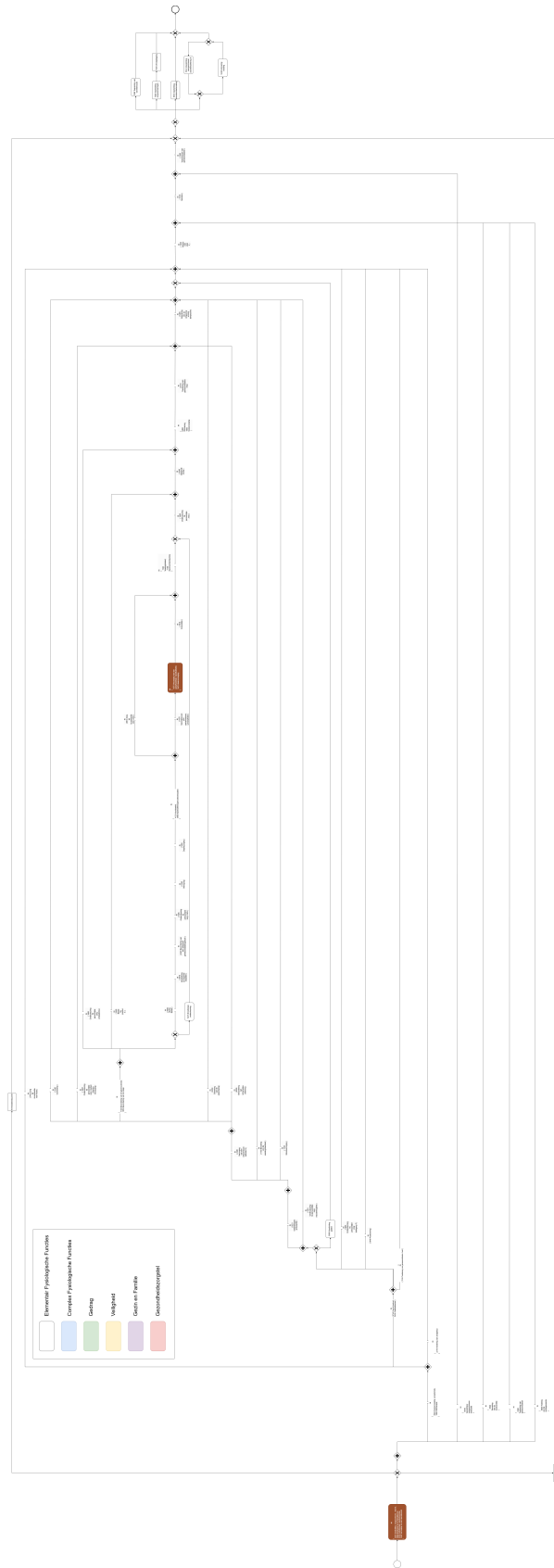


Figure 4.14: Created BPMN-D model on intervention level.

using DECLARE, but is very difficult to express in petri nets or BPMN. In short, isolating activities and investigating their relationship is one of DECLARE's greatest strengths. Furthermore, each element in the model carries meaningful information. No constraint means that any behavior is acceptable so the reader knows that activity does not need any special consideration.

On the other hand, the drawbacks of DECLARE cannot be ignored. A potential strength of DECLARE is that constraints may convey a lot of information, but the reverse is also true. Constraints such as `NotChainSuccession(A, B)` tend to not be as informative, as it simply means A is not directly followed by B. Generally this is intuitive for many activity interactions, and therefore tends to not be a useful addition to the model. As a result, the model may seem informative on a surface level, but could turn out to not add valuable, new information. Moreover, the constraints rely on the parameter thresholds, meaning if a constraint does not exceed the parameter thresholds, it will not be included in the model, or a weaker constraint is shown instead, reducing the value of the model. Finally, the notation tends to have a higher learning curve than the more intuitive imperative notations. When discussing a DECLARE model with stakeholders who are not experienced with DECLARE, this could lead to hesitant stakeholders, as well as increased difficulty when it comes to understanding the process.

### **4.6 SRQ2: What are the benefits and drawbacks of using Imperative Process Modelling?**

Imperative process modelling (specifically BPMN) has a relatively easy-to-understand syntax while remaining semantically simple. There is generally a clear sequence to activities, and it can be enriched in various ways such as color coding activities or adding weights to gateways to symbolize the probabilities of specific choices. Elements are generally modelled explicitly, thus having no "hidden" knowledge. Furthermore, BPMN has the tools to model many situations due to its explicit nature, as the individual elements of the notation are not complex.

However, these benefits do not come without drawbacks. As BPMN has a structured and sequential nature, it may be difficult to portray processes that do not have a fixed structure, but instead rely on guidelines. Making a complex, ad hoc environment such as nursing explicit will quickly make for a very large model as revealed in the previous section. This means that while the model is easily un-

derstandable on a small scale such as small patterns, it can make for a convoluted mode on a large scale and increase the difficulty of identifying patterns in the mode as a whole. This is evident in the models created on the intervention level, where the model got very large and difficult to interpret and analyze manually. Therefore, communicating such a model to stakeholders could prove challenging, as the reader is overloaded with too many details. However, the models on the higher taxonomy levels (i.e. class and domain) show that the procedural nature of BPMN is much less restrictive and convoluted. Due to the limited number of unique activities it makes for a more accurate model, as proven by the calculated metrics.

#### **4.7 SRQ3: What are the benefits and drawbacks of using Hybrid Process Modelling?**

As BPMN-D is a superset of BPMN, they have comparable advantages and disadvantages. This is shown through the evaluation metrics from the previous chapter. Considering the same level of data granularity, the fitness and precision scores were the exact same to due the way the BPMN-D model is a conservative modification of the BPMN model. However, simplicity did show a slight difference in favor of BPMN-D, indicating that by these calculations a BPMN-D model is a more simple representation of the event log. The same strengths and weaknesses apply to BPMN-D as they do to BPMN, considering BPMN-D is an extension of BPMN. Furthermore, due to the grouping of multiple activities, it may be challenging to enrich the model.

#### **4.8 SRQ4: What are important quality aspects for a process modelling technique?**

The four quality dimensions as defined by Van der Aalst exist for a reason, namely as a guideline to aim for a well-balanced process model. The dimension missing in this thesis is generalization, which is to balance a process model such that it does not overfit on the event log, but to some extent can handle unforeseen but likely process executions, for instance, the nurse washing hands an additional time. Ideally, all four dimensions are part of the evaluation as they give guidance to the strengths and weaknesses of the model and how to continue improving it.

#### **4.9 MRQ: How do different modelling techniques compare as a means to capture nursing processes?**

At the most detailed level, the resulting process models were convoluted regardless of notation. The origin of this is the wide variety of activities the nurses perform, ranging from physical activities such as helping the client get dressed, to providing social support by discussing whatever is occupying the client's mind. Furthermore, due to the ad hoc nature of the domain, there is no clear sequence to these activities. If a client is initiating a discussion about medication, then the nurse has to handle this first before continuing with other forms of care. As such, the process appears very unstructured on a detailed level. This is reflected particularly by the precision metric, which is very low for models on this level of granularity.

However, as we abstract away from the intervention level and inspect the process at the class or even domain level the process starts to visibly get more structured, again represented by the increasing precision the further we abstract away from the intervention level of activities. In addition, as the process gets increasingly simple, there is less need for the declarative extensions as suggested by BPMN-D.

Combining these findings, we state two conclusions. First, extending BPMN with declarative elements has a positive effect on the evaluation metrics used in this thesis. With the conservative extension that is BPMN-D, we only saw an increase in simplicity while keeping the other metrics consistent between the models. Second, modelling nursing activities on the most detailed level with an extensive taxonomy appears to make for an unstructured process model, which aligns with the actual domain. The downside is that this is very difficult to interpret for humans, and will likely negatively affect the precision of the process model.

## 5. Discussion

We evaluated the use of three distinct process modelling languages on district nursing processes using the same event log. Additionally, the event log was generated using Azure OpenAI and its state-of-the-art GPT-4 model. Exploring the use cases of LLMs is a trending research topic, but has not yet been thoroughly investigated in the field of nursing. This thesis has set an example of how the GPT-4 model may be used to generate event logs based on transcripts, outlining the performance and the challenges that come with this endeavor.

Furthermore, we can contextualize the results of the models based on the metrics and what the models look like per granularity level. The BPMN model on the intervention level illustrated the enormous complexity and variance of district nurse activities and their execution sequence. The frequent occurrence of the pattern shown in figure 4.10 and the abundance of parallel gateways confirm that BPMN has issues dealing with largely unstructured processes. As a result, the model scores high on fitness, but does so at the cost of precision and simplicity. This means that while the model is syntactically simple, using this model to complete a valid process execution is time-consuming. BPMN-D largely suffers from the same problems, but takes a smaller hit on simplicity due to its more compact notation. As BPMN-D is an extension of BPMN with several declarative properties, it can contain more complex model structures, such as the one shown in figure 4.10, into a single element.

DECLARE had to be evaluated separately, implying DECLARE models could not be directly compared to its BPMN and BPMN-D counterparts. Instead, DECLARE models were compared to DECLARE models on different granularity levels within the NIC taxonomy. DECLARE models were supposed to handle complex, ad hoc processes better than its (more) imperative competition. However, the models were not as informative as hypothesized. This is shown through the established constraint templates, most revolving around weak, negative constraints. For example, one of the most prominent constraints in the models was `NotChainResponse(A, B)`, meaning that A tends to NOT be directly followed by B, but potentially occurs later. While this constraint provides accurate information, it does so



in a very limited manner. Furthermore, as the activities became more abstract (i.e. the classification changed from intervention level to domain level) the constraint increased in quantity and a single case showed a stronger relationship by including a RespondedExistence(A, B) constraint in both the class level model and the domain level model. While this simply means that when A occurs, B is also present within the same trace either before or after A and tends to be more informative than the previously discussed NotChainResponse constraint. One possible explanation for this, is due to the ad hoc nature of nursing a very strict response is difficult to uphold. This works well on domain and class level, as domains and classes could be perceived as a collection of interventions, making establishing relationships and interactions easier.

Taking this as a premise, we expected DECLARE models to generally provide a richer, more informative and accurate representation of the process, but this was not the case. However, as shown through the evaluation metrics, BPMN and BPMN-D also did not perform great. Although the imperative model showed improved scores as the level of abstraction increased. Based on these outcomes, it seems like having a one-sided approach (i.e. fully declarative or fully imperative) is not the best fit for the nursing domain on the intervention level. BPMN-D appeared to outperform BPMN, given the metrics used in this thesis. However, BPMN-D is only a conservative extension of BPMN with a few additions from the declarative paradigm. The effect of adding more declarative concepts (or even removing several) is yet unknown, and discussing this in-depth would be largely speculative, although based on this particular result it appears to have a positive effect.

As illustrated in the results, BPMN benefited from the increased abstraction, where the number of unique activities was reduced and the process as a whole was simplified. As the model went from fine-grained to course-grained, the opposite trend was revealed where declarative constraints were less applicable to the models. Therefore, it may be conducive to understanding the process of selecting the right notation for the right level of granularity. Having multiple levels of granularity in the event log will also aid in understanding the process on different levels, and in communicating the process to different stakeholders. For example, nurses understand the specifics of each intervention well due to their hands-on experience, while a manager by default is less involved in the specifics of the interventions, thus having to resort to a more abstract approach. Having an event log that supports these multiple levels of abstraction facilitates communication with each

of these stakeholders, allowing one to pick the appropriate level of granularity for their audience. This aligns with figure 2.2, where the y-axis represented granularity and the x-axis the declarative-imperative process modelling approach spectrum. As the activities become more fine-grained (i.e. closer to intervention level) we observe through BPMN-D how integrating DECLARE concepts into BPMN results in a better model as shown in figure 4.13.

## 5.1 Limitations

The sky is the limit, but this thesis contains multiple other limitations. To start, this thesis used the GPT-4 model to both extract and classify activities according to the NIC taxonomy. Due to OpenAI introducing the GPT model, AI has seen a rapid surge in traction, and this trend will likely continue as many leading organizations are actively working on various AI models. Examples of these organizations are Microsoft, Google, and Meta. As such, Gen AI as a whole will likely evolve at a high pace, giving an exciting and promising prospect for future developments. The current GPT-4 model was unable to classify all activities correctly, and arguably extracted multiple activities which would not fit the definition of a nursing intervention.

A closely related limitation is the validation of the GPT output. One domain expert was asked for one hour of their time to validate the intervention classifications. Not much time was spent on evaluating whether the extracted activities were valid, nor were all events in the event log carefully validated by the domain expert due to this time constraint. To further improve the event log, more time could be spent by domain experts validating the output to provide a stronger foundation for the process models.

Alternatively, the foundation for these event logs was transcripts of audio recordings. While audio can be a good proxy for understanding what happened, in reality, there is a lot of context still missing. One of the clients made a joking remark about struggling in life, which naturally was part of the transcript. GPT falsely classified this as aiding a dying client. Furthermore, body language is a crucial medium of communication. It could supplement verbal communication, but in some instances completely replace it. Consequently, some interventions are not picked up on because they are performed without explicitly stating them. Dressing and undressing is one such example, where some traces included parts of either activity, while miss-

ing the other. However, this does not make logical sense, because that would mean the client would either start or end the visit without wearing clothes. According to the domain expert who validated the GPT output nurses are generally trained to verbalize these interventions, but situations may arise where this is redundant or the intervention is conducted while holding a casual, unrelated conversation.

One practical limitation was the number of transcripts that served as the input for the event log generation. Summarized by the age-old phrase of "not enough data". The total number of 27 transcripts (of which one was excluded as described in section 5.1.6.) amounted to approximately 340 events. For a complete event log, this is rather small. As a result, interventions that occurred only a single time in the event log had a relatively large presence of anomalies. Moreover, as there was only a single occurrence in the event log, it potentially implies a stronger relationship than in reality holds true. For example, if activity A is directly followed by activity B, and they occur only once in the event log it implies a causal relationship. However, that is much too strong of a relation to conclude based on a single occurrence.

Lastly, this thesis included Dutch clients and the Dutch version of the NIC taxonomy. Conducting a similar study where nurses are followed in different countries could potentially reveal a different distribution of activities, or different activity sequences. Culture may also influence the interactions between nurse and client, thus returning a completely different sequence of activities even if the context is the same. Regarding the Dutch NIC taxonomy, the taxonomy used for this thesis is the third edition originally written in 2000. As of April 2024, the most recent edition of the English book is the seventh edition published in 2018. This is a gap of eighteen years, and four editions. This is a noteworthy difference as a consequence of following the Dutch taxonomy, and following a revised, more up-to-date taxonomy could, again, lead to a vastly different process model.

## 5.2 Future Works

Research following this thesis could implement the limitations as outlined in the previous section. For example, when instructing a new AI model to generate an event log based on transcripts or any other source, a newer or different model could be employed. An improved GPT model could likely exceed the current performance shown in the results. On the other hand, a different model could lead to

different outputs. Perhaps a more specialized model could more accurately extract activities and classify them based on transcripts, or have parameters that could be configured to reach a more desirable output. Closely related to this is the prompt engineering aspect that could benefit from further refinement. A different model handles the same output differently, but a different prompt also affects the output. Techniques such as repair prompts and few-shot prompting could influence the model's behavior. Experimenting with these techniques will help in understanding how these black-box model operate, thus revealing more of their potential.

A different approach of model evaluation could be applied, where domain experts or a demographic that would use that particular model would evaluate the model in a subjective, human evaluation. This could lead to insights or nuances that are missed by objectively evaluating the models using a quantitative analysis.

One recommended avenue related to DECLARE is the use of the ProM plug-in used to generate a process model. This current plug-in uses an outdated engine, compared to the updated stand-alone available on GitHub [55]. The updated engine includes improved measure calculations and pruning mechanism, leading to generally more representative results.

Furthermore, AI models could be employed to generate an object-centric event log and enrich that log by adding classifications of some taxonomy, or a generic classification task. Object-centric event logs are layered due to their focus on how these objects relate to each other. On one hand, it makes for a more complicated and difficult-to-interpret event log, on the other hand, the presence of multiple objects could serve as a method to provide additional context for the model to grasp, thus providing the classification model with an extra tool to perform the task accurately.

Lastly, different types of process models may be applied in future work to evaluate their performance. This could range from declarative, to imperative, to hybrid. The spectrum is wide-ranged, and specifically hybrid has numerous options due to it being simply a combination of declarative and imperative process models, allowing the creator of the model to selectively pick the elements they deem appropriate for their specific context. As the results discussed in this thesis favor a hybrid approach, future research could explore other points on the declarative-imperative process modelling spectrum. More specifically, there could be a more in-depth analysis of the effect of more declarative additions to imperative approaches, or add imperative concepts to a declarative notation and evaluate its effects.

## 6. Conclusion

This thesis investigated the application of imperative, declarative, and hybrid process modelling languages in a nursing context. We compared the accuracy of process models in either language, showing the strengths and weaknesses of each of these notations, and evaluating the process models on their appropriate evaluation metrics. The foundation of these process models was an event log which was generated using the GPT-4 model, using transcripts of real-life nursing interventions. According to these metrics, the hybrid approach performed equally and in a single metric slightly better than the imperative notation, while the declarative process model could not be directly compared due to its constraint-focused approach. The models seemed to show better performance metrics when abstracting away from the most detailed level of the used NIC taxonomy.

## Bibliography

- [1] S. H. Woolf, R. Grol, A. Hutchinson, M. Eccles, and J. Grimshaw, "Potential benefits, limitations, and harms of clinical guidelines," *Bmj*, vol. 318, no. 7182, pp. 527–530, 1999.
- [2] J. Munoz-Gama, N. Martin, C. Fernandez-Llatas, *et al.*, "Process mining for healthcare: Characteristics and challenges," *Journal of Biomedical Informatics*, vol. 127, p. 103 994, 2022.
- [3] M. Rovani, F. M. Maggi, M. De Leoni, and W. M. Van Der Aalst, "Declarative process mining in healthcare," *Expert Systems with Applications*, vol. 42, no. 23, pp. 9236–9251, 2015.
- [4] M. De Regge, P. Gemmel, and B. Meijboom, "How operations matters in healthcare standardization," *International Journal of Operations & Production Management*, vol. 39, no. 9/10, pp. 1144–1165, 2019.
- [5] A. C. Freeman and K. Sweeney, "Why general practitioners do not implement evidence: Qualitative study," *Bmj*, vol. 323, no. 7321, p. 1100, 2001.
- [6] B. Olivares Bøgeskov and S. L. S. Grimshaw-Aagaard, "Essential task or meaningless burden? nurses' perceptions of the value of documentation," *Nordic Journal of Nursing Research*, vol. 39, no. 1, pp. 9–19, 2019.
- [7] P. Lavander, M. Meriläinen, and L. Turkki, "Working time use and division of labour among nurses and health-care workers in hospitals—a systematic review," *Journal of Nursing Management*, vol. 24, no. 8, pp. 1027–1040, 2016.
- [8] J. Jun, M. M. Ojemeni, R. Kalamani, J. Tong, and M. L. Crecelius, "Relationship between nurse burnout, patient and organizational outcomes: Systematic review," *International journal of nursing studies*, vol. 119, p. 103 933, 2021.
- [9] J. Bridel, "Why measure workload?" *Professional nurse (London, England)*, vol. 8, no. 6, pp. 362–365, 1993.
- [10] S. Jenkins-Clarke, "From lamps to laptops," *International Journal of Health Informatics*, vol. 2, no. 1, pp. 11–14, 1992.
- [11] A.-K. Levenstam and I. Bergbom Engberg, "How to translate nursing care into costs and staffing requirements: Part two in the zebra system," *Journal of Nursing Management*, vol. 5, no. 2, pp. 105–114, 1997.
- [12] S. Debois, T. Hildebrandt, T. Slaats, and M. Marquard, "A case for declarative process modelling: Agile development of a grant application system," in *2014 IEEE 18th international enterprise distributed object computing conference workshops and demonstrations*, IEEE, 2014, pp. 126–133.
- [13] L. Maas, M. Geurtsen, F. Nouwt, *et al.*, "The care2report system: Automated medical reporting as an integrated solution to reduce administrative burden in healthcare," in *HICSS*, 2020, pp. 1–10.
- [14] F. M. Maggi, C. Di Francescomarino, M. Dumas, and C. Ghidini, "Predictive monitoring of business processes," in *Advanced Information Systems Engineering: 26th International Conference, CAiSE 2014, Thessaloniki, Greece, June 16-20, 2014. Proceedings 26*, Springer, 2014, pp. 457–472.

- [15] E. M. Midlöv and T. Lindberg, "District nurses' experiences of providing palliative care in the home: An interview study," *Nordic Journal of Nursing Research*, vol. 40, no. 1, pp. 15–24, 2020.
- [16] G. M. Bulechek, H. K. Butcher, J. M. M. Dochterman, and C. Wagner, *Nursing interventions classification (NIC)*. Elsevier Health Sciences, 2012.
- [17] L. S. Lo, "The clear path: A framework for enhancing information literacy through prompt engineering," *The Journal of Academic Librarianship*, vol. 49, no. 4, p. 102 720, 2023.
- [18] B. F. Van Dongen, A. K. A. de Medeiros, H. Verbeek, A. Weijters, and W. M. van Der Aalst, "The prom framework: A new era in process mining tool support," in *Applications and Theory of Petri Nets 2005: 26th International Conference, ICATPN 2005, Miami, USA, June 20-25, 2005. Proceedings 26*, Springer, 2005, pp. 444–454.
- [19] R. Conforti, M. Dumas, L. García-Bañuelos, and M. La Rosa, "Bpmn miner: Automated discovery of bpmn process models with hierarchical structure," *Information Systems*, vol. 56, pp. 284–303, 2016.
- [20] F. M. Maggi, C. Di Ciccio, C. Di Francescomarino, and T. Kala, "Parallel algorithms for the automated discovery of declarative process models," *Information Systems*, vol. 74, pp. 136–152, 2018.
- [21] R. Dijkman, M. Dumas, B. Van Dongen, R. Käärrik, and J. Mendling, "Similarity of business process models: Metrics and evaluation," *Information Systems*, vol. 36, no. 2, pp. 498–516, 2011.
- [22] C. Klinkmüller, I. Weber, J. Mendling, H. Leopold, and A. Ludwig, "Increasing recall of process model matching by improved activity label matching," in *Business Process Management: 11th International Conference, BPM 2013, Beijing, China, August 26-30, 2013. Proceedings*, Springer, 2013, pp. 211–218.
- [23] W. M. van der Aalst and J. Carmona, *Process mining handbook*. Springer Nature, 2022, p. 26.
- [24] J. C. Buijs, B. F. van Dongen, and W. M. van der Aalst, "Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity," *International Journal of Cooperative Information Systems*, vol. 23, no. 01, p. 1 440 001, 2014.
- [25] A. Stefanini, D. Aloini, E. Benevento, R. Dulmin, and V. Mininno, "A process mining methodology for modeling unstructured processes," *Knowledge and Process Management*, vol. 27, no. 4, pp. 294–310, 2020.
- [26] A. Cecconi, C. Di Ciccio, G. De Giacomo, and J. Mendling, "Interestingness of traces in declarative process mining: The janus ltlp approach," in *International Conference on Business Process Management*, Springer, 2018, pp. 121–138.
- [27] W. van der Aalst, *Process Mining: Discovery, Conformance, and Enhancement of Business Processes*. Springer, 2011.
- [28] M. Ghasemi and D. Amyot, "Process mining in healthcare: A systematised literature review," *International Journal of Electronic Healthcare*, vol. 9, no. 1, pp. 60–88, 2016.
- [29] [Online]. Available: <https://www.celonis.com/blog/process-mining-the-missing-link-between-data-science-and-process-science/>.

- [30] R. S. Mans, W. M. Van der Aalst, and R. J. Vanwersch, *Process mining in healthcare: evaluating and exploiting operational healthcare processes*. Springer, 2015.
- [31] R. S. Mans, M. Schonenberg, M. Song, W. M. van der Aalst, and P. J. Bakker, "Application of process mining in healthcare—a case study in a dutch hospital," in *Biomedical Engineering Systems and Technologies: International Joint Conference, BIOSTEC 2008 Funchal, Madeira, Portugal, January 28-31, 2008 Revised Selected Papers 1*, Springer, 2009, pp. 425–438.
- [32] M. R. Dallagassa, C. dos Santos Garcia, E. E. Scalabrin, S. O. Ioshii, and D. R. Carvalho, "Opportunities and challenges for applying process mining in healthcare: A systematic mapping study," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18, 2022.
- [33] D. Fahland, D. Lübke, J. Mendling, *et al.*, "Declarative versus imperative process modeling languages: The issue of understandability," in *Enterprise, Business-Process and Information Systems Modeling: 10th International Workshop, BPMDS 2009, and 14th International Conference, EMMSAD 2009, held at CAiSE 2009, Amsterdam, The Netherlands, June 8-9, 2009. Proceedings*, Springer, 2009, pp. 353–366.
- [34] M. Pesic, "Constraint-based workflow management systems : Shifting control to users," English, Proefschrift., Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences, 2008, ISBN: 978-90-386-1319-2. DOI: 10.6100/IR638413.
- [35] P. Van Roy and S. Haridi, *Concepts, techniques, and models of computer programming*. MIT press, 2004.
- [36] P. Heintz, S. Horn, S. Jablonski, J. Neeb, K. Stein, and M. Teschke, "A comprehensive approach to flexibility in workflow management systems," *ACM SIGSOFT Software Engineering Notes*, vol. 24, no. 2, pp. 79–88, 1999.
- [37] M. Tory and T. Moller, "Human factors in visualization research," *IEEE transactions on visualization and computer graphics*, vol. 10, no. 1, pp. 72–84, 2004.
- [38] R. D. A. Rodrigues, M. D. O. Barros, K. Revoredo, L. G. Azevedo, and H. Leopold, "An experiment on process model understandability using textual work instructions and bpmn models," in *2015 29th Brazilian Symposium on Software Engineering*, IEEE, 2015, pp. 41–50.
- [39] A. Ottensooer, A. Fekete, H. A. Reijers, J. Mendling, and C. Menictas, "Making sense of business process descriptions: An experimental comparison of graphical and textual notations," *Journal of Systems and Software*, vol. 85, no. 3, pp. 596–606, 2012.
- [40] K. Figl and J. Recker, "Exploring cognitive style and task-specific preferences for process representations," *Requirements Engineering*, vol. 21, pp. 63–85, 2016.
- [41] W. M. van Der Aalst, M. Pesic, and H. Schonenberg, "Declarative workflows: Balancing between flexibility and support," *Computer Science-Research and Development*, vol. 23, pp. 99–113, 2009.
- [42] P. Pichler, B. Weber, S. Zugal, J. Pinggera, J. Mendling, and H. A. Reijers, "Imperative versus declarative process modeling languages: An empirical investigation," in *Business Process Management Workshops: BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I 9*, Springer, 2012, pp. 383–394.



- [43] S. Mertens, F. Gailly, and G. Poels, "Supporting and assisting the execution of flexible healthcare processes," in *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, IEEE, 2015, pp. 329–332.
- [44] M. Poulymenopoulou, F. Malamateniou, and G. Vassilacopoulos, "Specifying workflow process requirements for an emergency medical service," *Journal of medical systems*, vol. 27, pp. 325–335, 2003.
- [45] S. Gupta *et al.*, "Workflow and process mining in healthcare," *Master's Thesis, Technische Universiteit Eindhoven*, 2007.
- [46] R. Mans, H. Schonenberg, G. Leonardi, *et al.*, "Process mining techniques: An application to stroke care," in *MIE*, vol. 136, 2008, pp. 573–578.
- [47] R. Nunes-Vaz, P. Arbon, and M. Steenkamp, "Imperatives for health sector decision-support modelling," *International Journal of Disaster Risk Reduction*, vol. 38, p. 101 234, 2019.
- [48] T. T. Hildebrandt and R. R. Mukkamala, "Declarative event-based workflow as distributed dynamic condition response graphs," *arXiv preprint arXiv:1110.4161*, 2011.
- [49] M. Pesic, H. Schonenberg, and W. M. Van der Aalst, "Declare: Full support for loosely-structured processes," in *11th IEEE international enterprise distributed object computing conference (EDOC 2007)*, IEEE, 2007, pp. 287–287.
- [50] N. Martin, J. De Weerd, C. Fernández-Llatas, *et al.*, "Recommendations for enhancing the usability and understandability of process mining in healthcare," *Artificial Intelligence in Medicine*, vol. 109, p. 101 962, 2020, ISSN: 0933-3657. DOI: <https://doi.org/10.1016/j.artmed.2020.101962>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0933365720312276>.
- [51] A. Abbad Andaloussi, A. Burattin, T. Slaats, E. Kindler, and B. Weber, "On the declarative paradigm in hybrid business process representations: A conceptual framework and a systematic literature study," *Information Systems*, vol. 91, p. 101 505, 2020, ISSN: 0306-4379. DOI: <https://doi.org/10.1016/j.is.2020.101505>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0306437920300168>.
- [52] G. De Giacomo, M. Dumas, F. M. Maggi, and M. Montali, "Declarative process modeling in bpmn," in *Advanced Information Systems Engineering: 27th International Conference, CAiSE 2015, Stockholm, Sweden, June 8-12, 2015, Proceedings 27*, Springer, 2015, pp. 84–100.
- [53] I. Teinmaa, M. Dumas, M. L. Rosa, and F. M. Maggi, "Outcome-oriented predictive process monitoring: Review and benchmark," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 13, no. 2, pp. 1–57, 2019.
- [54] C. Di Francescomarino, C. Ghidini, F. M. Maggi, and F. Milani, "Predictive process monitoring methods: Which one suits me best?" In *International conference on business process management*, Springer, 2018, pp. 462–479.
- [55] C. Di Ciccio, *Minerful*, <https://github.com/cdc08x/MINERful>, 2024.

# A. Appendices

## A.1 Appendix A

Table A.1: Complete table of class distribution in event log.

Class	Frequency	Relative frequency
F Ondersteuning bij persoonlijke zorg	134	41.61 %
B Zorg voor de uitscheiding	37	11.49 %
H Zorg bij geneesmiddelengebruik	34	10.56 %
S Patiëntenvoorlichting	28	8.7 %
L Huid- en wondverzorging	24	7.45 %
E Bevorderen van lichamelijke comfort	10	3.11 %
D Zorg voor de voeding	10	3.11 %
K Zorg voor de ademhaling	8	2.48 %
R Ondersteuning bij probleemhantering	8	2.48 %
A Bevorderen van activiteiten en lichaamsbeweging	7	2.17 %
C Zorg bij immobiliteit	5	1.55 %
Y Bemiddeling in zorgvoorzieningen	5	1.55 %
W Zorg rondom geboorte	3	0.93 %
V Risicobestrijding	3	0.93 %
b Informatiebeheer	3	0.93 %
X Zorg voor gezin en familie	2	0.62 %
N Zorg voor de weefseldoorbloeding	1	0.31 %

## A.2 Appendix B

**Table A.2:** Complete table of intervention distribution in event log.

<b>Intervention</b>	<b>Frequency</b>	<b>% Frequency</b>
1802 Ondersteuning bij persoonlijke zorg: kleden/ uiterlijke verzorging	28	8.7 %
1801 Ondersteuning bij persoonlijke zorg: wassen/hygiëne	27	8.39 %
1610 Baden/ douchen	26	8.07 %
2300 Toediening van geneesmiddelen	17	5.28 %
1800 Ondersteuning bij persoonlijke zorg	16	4.97 %
3660 Wondverzorging	12	3.73 %
3590 Bewaking van de huidconditie	11	3.42 %
5616 Voorlichting: voorgeschreven medicatie S	11	3.42 %
1806 Ondersteuning bij persoonlijke zorg: verplaatsing F	10	3.11 %
6482 Regulering van de omgeving: comfort	9	2.8 %
5602 Voorlichting: ziekteproces	8	2.48 %
1160 Bewaking van de voedingstoestand	8	2.48 %
0480 Stomazorg	8	2.48 %
1720 Mondverzorging	7	2.17 %
0430 Defecatiebeleid	6	1.86 %
3550 Beleid bij pruritus E	6	1.86 %
0410 Zorg bij incontinentie voor feces	6	1.86 %
2380 Bevordering van een adequaat geneesmiddelengebruik	6	1.86 %
5606 Voorlichting: patiënt	6	1.86 %
1670 Haarverzorging	5	1.55 %
0200 Bevordering van lichamelijke oefening	4	1.24 %
1804 Ondersteuning bij persoonlijke zorg: toiletgang F	4	1.24 %
1850 Bevordering van de slaap	4	1.24 %
0450 Obstipatiebeleid	4	1.24 %
2210 Toediening van analgetica	4	1.24 %
0590 Mictie Beleid	4	1.24 %
0610 Zorg bij incontinentie voor urine	4	1.24 %
0970 Verplaatsing bij beperkte mobiliteit	3	0.93 %
5270 Emotionele ondersteuning	3	0.93 %
1650 Oogverzorging	3	0.93 %
5240 Counseling	3	0.93 %
2390 Voorschrijven van geneesmiddelen	3	0.93 %
2316 Toediening van geneesmiddelen: huid	2	0.62 %
6485 Regulering van de omgeving: woning aanpassen	2	0.62 %
1876 Verzorging van een aan/afvoerbuis: urinekatheter	2	0.62 %
1660 Voetverzorging	2	0.62 %
6680 Monitoring van de bloedglucose	2	0.62 %
0640 Geplande mictie	2	0.62 %
0221 Oefentherapie: ambulante	2	0.62 %
7310 Opname	2	0.62 %
3350 Bewaking van de ademhaling	1	0.31 %
1875 Verzorging van een aan/afvoerbuis: navelkatheter	1	0.31 %
3583 Huidverzorging: acceptorsite J	1	0.31 %

<b>Intervention</b>	<b>Frequency</b>	<b>% Frequency</b>
0460 Diarree Beleid	1	0.31 %
6490 Valpreventie	1	0.31 %
0560 Spieroefeningen: bekkenbodem	1	0.31 %
0226 Oefentherapie: spierbeheersing	1	0.31 %
1380 Toepassing van warmte/koude	1	0.31 %
4120 Vochtbeleid	1	0.31 %
5618 Voorlichting: procedure/behandeling	1	0.31 %
7400 Voorlichting: zorgvoorzieningen	1	0.31 %
7920 Verslaglegging	1	0.31 %
1860 Sliktherapie	1	0.31 %
5614 Voorlichting: voorgeschreven dieet S	1	0.31 %
2440 Verzorging van een intraveneuze katheter N	1	0.31 %
3440 Verzorging van een operatiewond	1	0.31 %
5246 Advisering: voeding	1	0.31 %
5612 Voorlichting: voorgeschreven activiteit/oefening S	1	0.31 %
1805 Ondersteuning bij persoonlijke zorg: IADL	1	0.31 %
8100 Verwijzing	1	0.31 %
2310 Toediening van geneesmiddelen: oog	1	0.31 %
5310 Hoop inspreken en bemoedigen	1	0.31 %
0600 Training van de mictiegewoonte	1	0.31 %
7100 Bevordering van de gezinscohesie	1	0.31 %
0740 Zorg bij bedrust	1	0.31 %
0840 Positionering	1	0.31 %
7910 Consultatie	1	0.31 %
1100 Voedingsbeleid	1	0.31 %
1050 Helpen met eten F	1	0.31 %
7140 Gezinsondersteuning	1	0.31 %