How Can Interactive Process Discovery Address Data Quality Issues in Real Business Settings? Evidence from a Case Study in Healthcare

Elisabetta Benevento, Davide Aloini, and Wil M. P. van der Aalst[†]

May 1, 2022

Abstract

The focus of this paper is on how data quality can affect business process discovery in real complex environments, which is a major factor determining the success in any data-driven Business Process Management project. Many real-life event logs, especially healthcare ones, can suffer from several data quality issues, some of which cannot be solved by pre-processing or data cleaning techniques, leading to inaccurate results. We take an innovative Process Mining (PM) approach, termed Interactive Process Discovery (IPD), which combines domain knowledge with available data. This approach can overcome the limitations of noisy and incomplete event logs by putting "humans in the loop", leading to improved business process modelling. This is particularly valuable in healthcare, where physicians have a tacit domain knowledge not available in the event log, and, thus, difficult to elicit. We conducted a two-step approach based on a controlled experiment and a case study in an Italian hospital. At each step, we compared IPD with traditional PM techniques to assess the extent to which domain knowledge helps to improve the accuracy of process models. The case study tests the effectiveness of IPD to uncover knowledge-intensive processes extracted from noisy real-life event logs. The evaluation has been carried out by exploiting a real dataset of an Italian hospital, involving the medical staff. IPD can produce an accurate process model that is fully compliant with the clinical guidelines by addressing data quality issues. Accurate and reliable process models can support healthcare organizations in detecting process-related issues and in taking decisions related to capacity planning and process re-design.

Keywords Interactive Process Discovery, Process Mining, Data quality, Business Process Modelling, Healthcare.

1 Introduction

Healthcare organisations are increasingly pushed to improve the quality of care services in an unfavourable and rapidly changing scenario [1]. Advances in treatments, an ageing population,

[*]E. Benevento (e-mail: elisabetta.benevento@ing.unipi.it) and D. Aloini are with the Department of Energy, Systems, Territory and Construction Engineering, University of Pisa, Largo Lucio Lazzarino 1, Pisa, 56122, Italy.

[†]W. M. P. is with Rheinisch-Westfälische Technische Hochschule (RWTH), Ahornstraße 55, 52074, Aachen, Germany and Fraunhofer Institute for Applied Information Technology FIT, Schloss Birlinghoven, 53757, Sankt Augustin, Germany.

increased patient expectations, limited resources are some of the factors that have increased the pressures on hospitals [2]. Therefore, improving process efficiency is of utmost importance and represents a critical success factor for healthcare organisations.

To accommodate these challenges, it is essential to seize the opportunity offered by digitalisation. In fact, digital innovation seems to be the key-driver factor to improve the continuity and the access to care for patients, and to guarantee greater effectiveness, efficiency, and sustainability of healthcare organisations [3].

Advances in software and hardware such as clouds, IoT sensors, and digital platforms have led to an ever-growing volume and variety of process and patient data, thus expanding the spectrum of technologies and techniques applicable to Business Process Management (BPM) [4]. Innovative data-driven techniques, such as Process Mining (PM), Machine Learning (ML), and Data Mining (DM), provide new ways to interpret and implement BPM practices [5–7]. Healthcare organizations can analyse and transform their business processes more effectively and efficiently by leveraging PM and other data-driven approaches, thus improving decision making [8, 6] to attain operational efficiency [9, 10].

These advantages can only be achieved if a high level of data quality is guaranteed. Indeed, data quality can significantly affect the decision-making processes of the organizations and have direct implications for the quality of healthcare provision [6, 5, 11]. Indeed, the accuracy and reliability of any analysis are significantly influenced by the quality of the data [9, 12].

However, real-life event logs (i.e., data input for PM techniques) can suffer from several data quality issues [13], especially in healthcare domain [14]. Indeed, errors in data entry and/or data collection due to excessive workload of professionals, the need for better training in the use of new technologies, and extensive manual recording are causes of low quality in most healthcare datasets [15]. Data quality is thus challenging for BPM [13], particularly in healthcare [14], and is the focus of numerous studies [16, 17]. However, despite this research into overcoming data quality issues during the pre-processing phase [18], some problems remain and can lead to unreliable or misleading results.

We address the data quality issues affecting business process discovery and analysis in the healthcare environment. We exploit an innovative PM approach, namely Interactive Process Discovery [19], which combines domain knowledge with available event logs. This innovative PM approach can overcome the limitations of noisy event logs by putting "humans in the loop" [20], unlike other PM techniques, which are mainly based on raw data. This is particularly valuable in the healthcare context, where physicians have a tacit domain knowledge not available in the event log, and, thus, difficult to elicit. Therefore, Interactive Process Discovery can enhance business process modelling and lead to improved results. However, despite their potential, interactive techniques are still rarely used in real cases and only "high quality" event logs are exploited [19]. Few studies have addressed the problem of data quality in real business settings [58].

Thus, we address the research question of how Interactive Process Discovery can tackle data quality issues in real complex BPM applications, like healthcare, since data quality is recognized as one of the most relevant and critical perspective to analyse the success of BPM data-driven projects [21, 22]. In responding to the research question, we took a two-step approach and conducted a controlled experiment and a case study. In our experiment we reproduced two common data quality issues and estimated their impact on the accuracy of the resulting process models. Our case study came from the healthcare sector, as this is characterised by complex and noisy event logs. We analysed a real dataset from an Italian hospital and a multidisciplinary hospital team was involved in the study. Our aim was to demonstrate the capability of Interactive Process Discovery to handle large and noisy real-life event logs and to produce accurate and comprehensible process

models.

From a scientific perspective, our study responds to the call for more effective business process modelling approaches that address quality problems within the data [23, 13], especially in the healthcare sector [14]. In so doing, we intent to demonstrate with a rigorous approach the effectiveness of Interactive Process Discovery to deal with data quality issues embedded in real-life event log. To the best of our knowledge, this represents one of the first attempts to evaluate the advantages of tacit domain knowledge in improving the quality and comprehensibility of process models generated from noisy event logs. From a practical point of view, accurate and reliable process models can support healthcare organizations in effectively detecting process-related issues (e.g., bottlenecks, process anomalies) and enable them to take decisions related to capacity planning, resource allocation, and process re-design more efficiently. Finally, we demonstrate the suitability of Interactive Process Discovery in real applications and provide a user's guide for the technique, thus assisting practitioners in applying such knowledge and managing data-driven BPM projects more effectively.

The paper is structured as follows. In Section 2 we describe the theoretical background, and we introduce the related research in Section 3. Section 4 describes the Interactive Process Discovery technique and Section 5 illustrates the research design. Section 6 describes the results of the controlled experiment and in Section 7 we present the results of the hospital case study. Finally, Section 8 provides conclusions and future work.

2 Theoretical Background

Business Process Modelling is one of the key ingredients of BPM [24], enabling a common and comprehensive understanding of business processes [25]. Its aim is to produce an effective representation of a business process through a process model, which can be used for analysis and optimization [26]. High-quality process models help organizations improve efficiency [24, 27] and promptly avoid any errors [28, 29]. This is essential as error costs increase exponentially throughout the development lifecycle [30].

Thus, the success of a business process modelling project has been recognised as a critical outcome of BPM, as its consequences can often be significant and lead to the design of new processes and organizational structures, the re-design of existing processes, and the implementation of ISs [31]. Several studies of the quality measures and success factors in BPM have been conducted and approach the problem from multiple perspectives [31, 29, 32–34]. In the following review we discuss the relevant studies and focus on data-driven BPM.

According to the framework proposed in [31, 34], modelling success can be measured in terms of model quality, process impacts, and project efficiency. The first two dimensions represent the effectiveness of the model in terms of the fulfilment of the desired outputs and the effects of the model improvements, while efficiency is measured by resource usage in the project. Both project- and modelling-specific success factors can be identified. The former includes a set of factors related to the project and associated context, i.e., project management capabilities, stakeholder participation, information resources, and modeller expertise. Modelling-specific success factors involve the tool used, the language, and the methodology, as a common understanding and a standardised model development can enhance the quality and reusability of the project outputs [34]. In addition to these factors, the framework of [34] identifies process complexity and project importance as two moderators, as they can influence the success of the modelling. Process complexity generally depends on whether the process being modelled is structured or unstructured, while the importance

of the project to the organization can also determine its perceived success.

This framework fits with model-driven process modelling as it mainly draws on interviews with process participants, direct field observations, and documentation analysis [35, 36]. However, in data-driven process modelling projects, data represents another success factor in addition to project- and modelling-related factors [21, 22]. Although process digitalization often guarantees the availability of data in a BPM project, the data are often "dirty" [13]. Thus, the success of a data-driven project depends on the quality of the dataset and how quality issues are detected and resolved [21].

As shown in the framework below (see Fig. 1), which is adapted from [31, 34], data availability and data quality are factors specific to data originating from ISs. Model accuracy is measured in terms of four metrics commonly used in literature to assess the quality of a process model: fitness, precision, simplicity, and generalization [37].

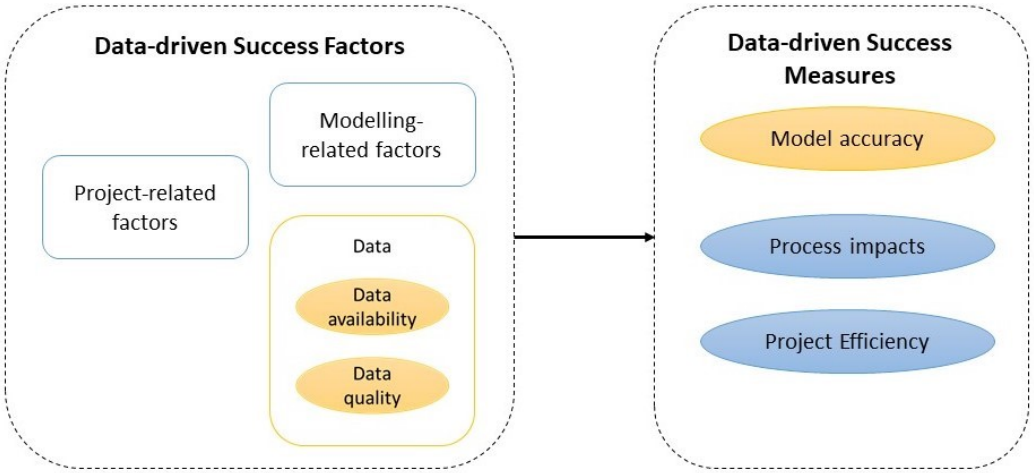

Figure 1: Data-driven process modelling success framework adapted from [31, 34].

The study of [23] suggests that there are two main categories of data quality problems: (i) process-related issues, which involve challenges arising from to the characteristics of the investigated process as recorded in the dataset, such as fine-grained activities and heterogeneity; and (ii) those related to the quality of the event logs in the datasets. Managing these issues is critical, and they can be separated into four sub-categories [23]:

1. **Missing data**: Information may be missing from a dataset even though it should be present.
2. **Incorrect data**: Some data entries in the dataset may be recorded incorrectly.
3. **Imprecise data**: A lack of detail in the data entries can lead to reduced precision.
4. **Irrelevant data**: Some data entries may not be usable in their current format and therefore are irrelevant.

These issues can reduce the applicability of various data-driven techniques and limit the insights gained, which can potentially lead to complex or incomprehensible process models that do not provide a meaningful representation of the reality [15, 23]. This is accentuated in knowledge-intensive processes, like the healthcare ones, that are characterised by high levels of heterogeneity and fewer repeatable activities [38]. Thus, more effective data-driven approaches able to handle data quality issues are required [23].

3 Related Work

In this section, we review previous works relating to process discovery and to data quality issues in the PM literature. We focus on current Automated Process Discovery (APD) techniques, based solely on event logs, that address data quality problems that cannot be solved during the data pre-processing phase and discuss the accuracy and comprehensibility of the models.

Two main strategies for addressing such data quality issues during the discovery phase have been identified in the PM literature. First, filtering out infrequent events and second, managing uncertain events, can improve the quality and reliability of the results.

APD techniques including the Heuristic Miner [39, 40], Inductive Miner [41], Fuzzy Miner [42], and Split Miner [43] are aimed at tackling specific data quality issues such as partial/incomplete traces within the log [23]. They filter out noisy behaviour during the process discovery phase. These techniques can result in a process model that contains only the most frequent traces, thus reducing its complexity. [44] propose an alternative APD technique to manage uncertain events, by discovering a directly- follows graph from event logs, in which the events are recorded with some level of uncertainty [45]. A formal description of uncertainty (as a process model) can then be identified, rather than aiming to eliminate uncertainty so the underlying process can be observed [46].

However, both approaches have drawbacks when handling large event logs with imprecise data, such as imprecise timestamps or activity names [47, 48, 37, 23]. If coarse or mixed granular timestamps occur within the event log [49, 23], event ordering issues can arise, leading to parallelism instead of sequential behaviour in the mined model [13]. APD techniques can then identify paths that do not match any trace in the event log, leading to a lack of precision [47, 50]. This condition can affect the reliability and comprehensibility of the results. In addition, most APD techniques cannot identify duplicate activities [37], which may appear in different parts of the process model. Although they differ, these are often mapped onto one activity in the model, leading to loops [37, 19, 23]. This limitation is particularly significant in contexts such as healthcare, where processes may require that the same activity occurs at various stages (or branches) of the process that have different aims. Such activity cannot be identified and relabelled neither ex-ante nor ex-post during pre-processing. Due to these drawbacks, APD techniques often generate complex, less accurate, or incomprehensible process models [37, 51], that do not provide a significant representation of the reality.

The PM literature has more recently focused on exploiting the tacit knowledge from experts with the event logs when discovering process models [19, 52]. For example, [53] propose a hybrid approach to process discovery, in which a method is developed that can consider a variety of constraints that the analyst can draw on. Here, domain knowledge can be encoded in the form of precedence constraints. In [54], the authors develop a technique that leverages prior knowledge and process execution data to learn a control-flow model. The knowledge is incorporated using ideas from Bayesian statistics. [55] propose a set of techniques to prune declarative process models and remove constraints that are not relevant or implied by other constraints, by exploiting domain knowledge. However, these approaches still present various limitations, such as only enabling users to express their domain knowledge in the form of rules or constraints that the model must respect [55, 53, 54]. The user is then limited by the language used to reproduce the rules. In addition, as the user cannot control the discovery of the process, although the resulting map may be compliant with all rules it may also be extremely complex. [56] propose an approach that incrementally discovers process models using domain knowledge. The developed algorithm enables the user to choose the traces to add to existing process models. The process model constructed is thus gets

incrementally extended. However, when trace variability is high, as in the case of noisy event logs, users will find selecting the correct traces difficult.

IPD [19] provides the user with complete control over the modelling phase, enabling both tacit knowledge and event logs to be applied and the process model to be investigated interactively. The user can then ignore any indications provided by the event log that are deemed unreliable or incorrect. In addition, from a technical point of view IPD can identify particular constructs such as duplicate activities, inclusive choices, and silent activities, which is not possible in many of the other state-of-the-art techniques, either automated or interactive. Thus, IPD can overcome typical representational biases that can negatively affect the accuracy and comprehensibility of the process models [57, 37].

However, to the best of our knowledge, only the study of [58] has tried to apply an interactive approach to a real-life noisy event log for data quality assessment and data cleaning purposes. Consequently, evidence on the effectiveness and suitability of such techniques is still limited and thus valuable to investigate. We contribute to the literature by demonstrating how incorporating domain knowledge into the modelling phase can result in accurate and understandable process models, even with noisy event logs. The success of any business process modelling project depends to a large extent on the ability to manage data quality. We also address the role of domain knowledge, which is typical in knowledge-intensive contexts, like healthcare. Thus, we investigate how using an interactive approach like IPD enables us to capture aspects of tacit domain knowledge that cannot be easily represented in a structured and explicit way.

4 The Interactive Process Discovery Technique: A Brief Overview

IPD is an innovative PM technique developed by [19] for modelling business processes. It combines the information in the event log with domain knowledge, unlike APD techniques that exploit only the dataset during the discovery phase and automatically produce a process model from it. Fig. 2 depicts the general scheme of IPD.

As shown in Fig. 2, IPD enables the interactive editing and construction of process models in an incremental and structured manner. Feedback and suggestions from the event log are provided during model building to help the user decide where to position the activities within the model. The user can decide to take assistance from the data or ignore the suggestions. If the user considers the recommendations inadequate or incorrect, he can deviate from them and place the activity in the correct position based on his knowledge. This is repeated until all activities in the event log have been placed on the map. At the end, the user can export the final map.

IPD uses the following synthesis rules to enable interactive model building: an abstraction rule (AR), place rule (PR), and transition rule (TR) [59, 19]. From these, a minimal workflow net can be expanded by adding one transition and/or one place at a time [19]. AR enables the user to introduce a new place and a new transition between a set of transitions and places. TR enables the inclusion of a new transition into the labelled free-choice workflow net, while PR adds a new place. All applications of these rules are projected onto the net, depending on the feedback from the log and the interaction of the user with IPD [19]. Further mathematical and technical details are available in [19].

To support the user in decision-making, IPD extracts aggregated information about activity and contextual information from the event log. The activity information, i.e., the maximum, minimum, and average occurrences and the percentage of cases, can provide insights about the activity, such

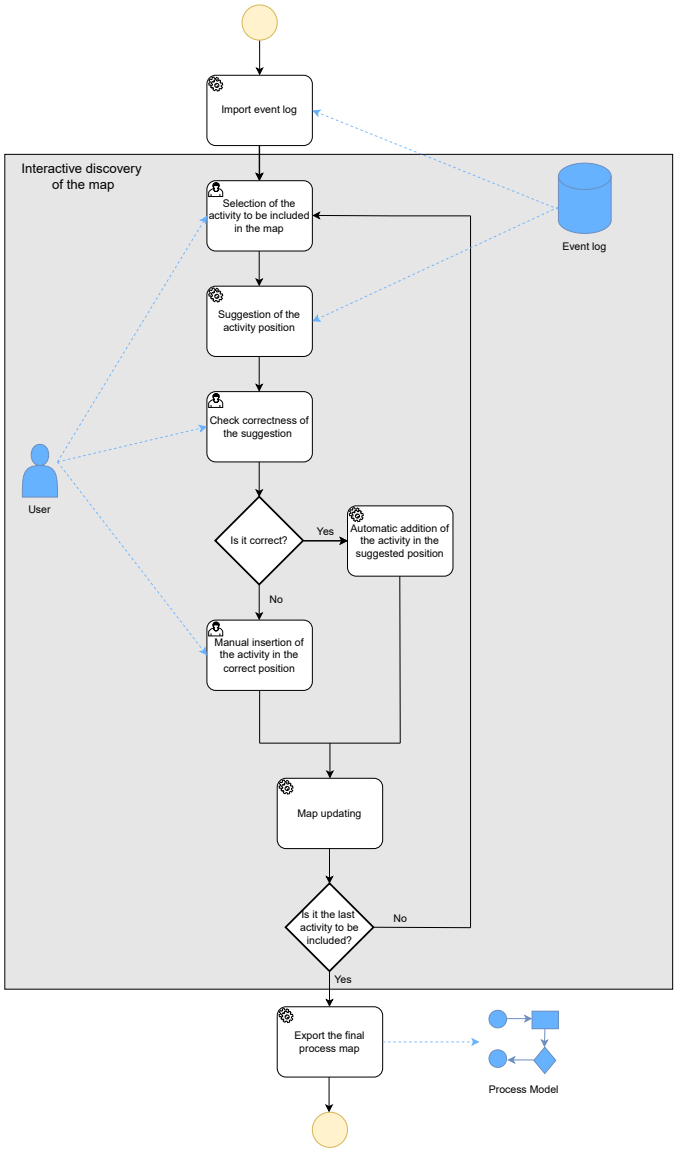

Figure 2: Overview of IPD.

as whether it is prevalent in the event log, if it should be placed in a loop or duplicated, etc. In addition, contextual information about the activity in terms of others that are already present within the net is provided to the user. This supports the user in placing the activity in the net. Such information from the event log is projected directly onto the current net, and depending on the activity, the colouring of the other activities within the map is updated. Colours denote which activities take place before and after the selected one. Purple is used to highlight activities that occur after the selected activity, yellow denotes those that occur before, and white transitions imply that the selected activity and others do not occur at the same time. The projected information can be based either on the eventually follows (precedes) relation or on the directly follows (precedes) relation, as desired by the user. Fig. 3 provides an example and depicts modifications within the net when the user chooses an activity x. Transition $t1$ is completely purple, meaning that x happens after a. Likewise, $t3$ is yellow, which implies that x happens before c. Conversely, $t2$ is half purple and half yellow, and thus x can occur before and after b. Clearly, x must therefore be

placed in parallel to b, before c, and after a. The context information from the event log is also summarized in a table format, thus helping the user take appropriate decisions when the projected information is too vague.

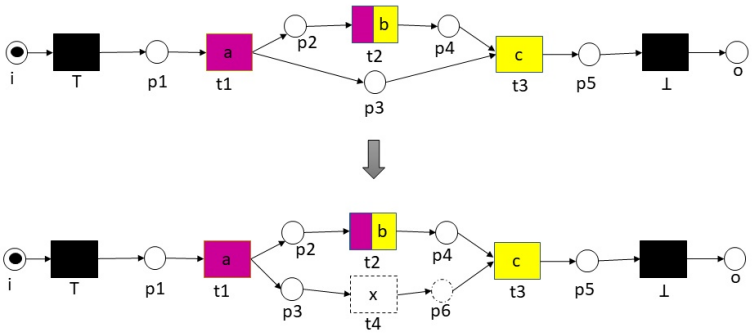

Figure 3: Projected information when an activity x is chosen by the user. The colours suggest suitable positions in the model.

5 Research Design

We address the research question using a two-step approach based on experimental and case study research methods, which are widely used in IS and BPM research [60, 61]. The first step, the controlled experiment, enables us to conduct a more detailed evaluation by examining specific data quality issues and their effects on the IPD results. The second step, the case study, enables a broader and more generalised analysis of how IPD can address data quality issues in complex real-life contexts, like healthcare. This combined approach can be particularly useful when evaluating the suitability of BPM technologies and techniques for specific projects [61, 62].

The controlled experiment is aimed at investigating the behaviour of the process model produced by IPD when two data quality issues, i.e., imprecise timestamps (scenario A) and imprecise events (scenario B), occur and at estimating the impact of each on the accuracy of the model. Therefore, the focus of such experiment is not on the functionality of IPD but on the quality of the resulting process model in the presence of such quality problems. Scenario A involves timestamps of the events that are too coarse, e.g., in the order of a day, thereby making the ordering of events unreliable [23, 63]. Scenario B involves events that have the same activity name but different aims, depending on the context [23, 64]. In the data pre-processing stage, understanding a-priori whether an event that is repeated several times has the same or different purpose in different stages of the process can be difficult. Both of these data quality issues are common in real-life datasets and difficult to manage in the data pre-processing phase [23, 49, 65]. They can affect the reliability of the results [21], and this evaluation thus helps us understand the extent to which IPD can limit the negative effects of such quality problems.

We set our case study in the healthcare context, so we can evaluate the effectiveness of IPD in handling large and noisy event logs in real complex applications. Real datasets suffer from several data quality issues that are difficult to manage simultaneously. The evaluation allows us to examine the benefits of combining knowledge with data about the quality of the results, and also when quality problems occur. The case study also demonstrates how IPD can capture the tacit knowledge within the healthcare context and how it can be used to improve the accuracy and

understandability of the process model. Fig. 4 illustrates the whole research design developed for the evaluation.

In both steps, IPD was compared with current state-of-the-art APD techniques based on event logs, to evaluate the potential advantages of domain knowledge in improving the quality of process models generated from noisy event logs. We used synthetic event logs for the controlled experiment and engaged three process analysts, while we used a real dataset from an Italian hospital for the case study and involved a team of specialists including the health director, IT staff, the head of the thoracic surgery department, an oncologist, a ward doctor, and two ward nurses. We evaluated accuracy (i.e., fitness, precision, and F-score) in both cases, and for the case study only, we also measured the level of compliance with clinical guidelines that healthcare processes should follow when treating patients.

The two steps are detailed in Sections Sections 5.1 and 5.2.

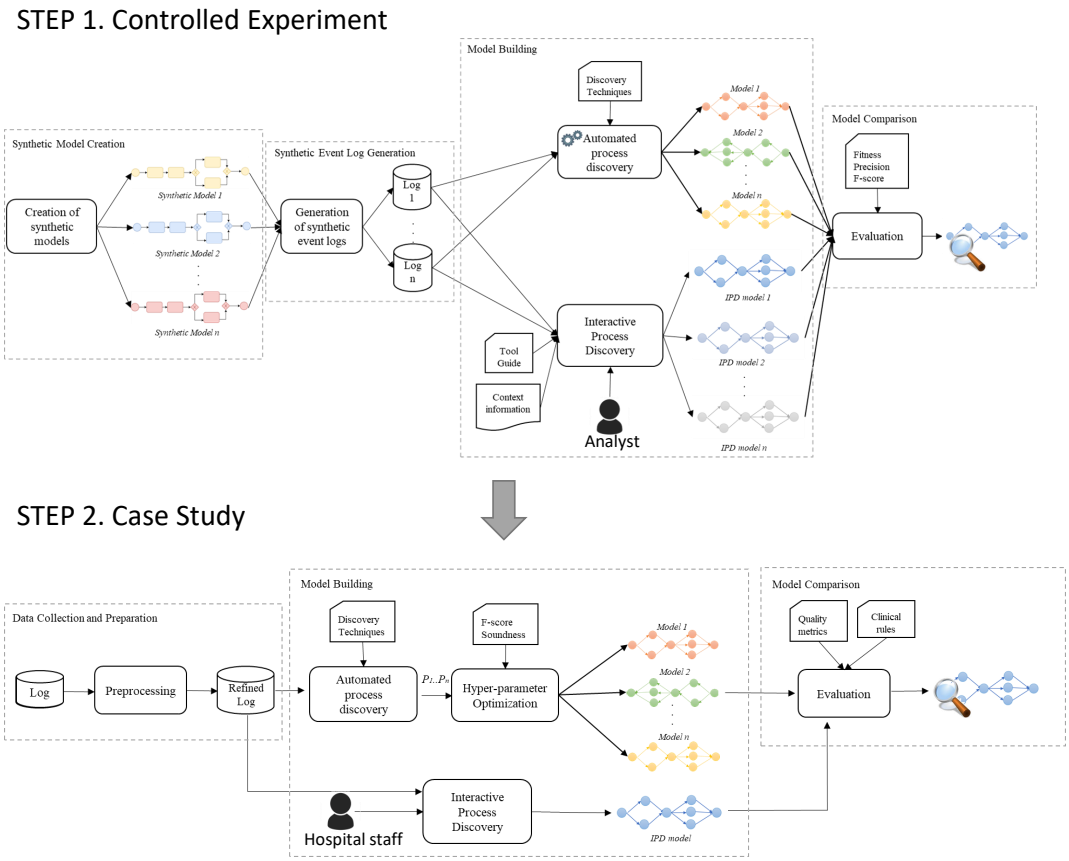

Figure 4: Two-step approach followed to test the effectiveness of IPD in dealing with data quality issues.

5.1 Step 1. Synthetic Experiment

The controlled experiment consisted of four phases, which were repeated for scenarios A (i.e., event log with imprecise timestamps) and B (i.e., event log with imprecise events):

1. **Synthetic Model Creation**. Three synthetic BPMN models were created using the Signavio tool, by setting the number of activities and the type of constructs (the resulting

BPMN models are presented in Supplementary Material). These process models were used to simulate event logs in the next phase.

2. **Synthetic Event Log Generation**. For each BPMN model, a synthetic event log was generated through the BIMP Simulator tool. Noise was introduced in all of the synthetic event logs to produce more realistic datasets. Specifically, starting from each log, three additional modified synthetic logs were generated, by removing time information or by adding/removing occurrences of random activity. We thus obtained 12 synthetic event logs (4 for each BPMN model) at the end of the phase, with each having between 1000 and 9000 events. These synthetic event logs represent inputs to both IPD and APD techniques.

3. **Model Building**. We selected the Inductive Miner (IM) [41], Split Miner (SM) [43], and Directly-Follows Graphs Miner [66] (DFGM) (as a proxy for the commercial tools), as these are common APD techniques. To discover the process models, we conducted the following procedure (for more details see Supplementary Material). Three process analysts, experts in BPM, were engaged for the experiment. Each participant was first provided with: (i) four different synthetic logs generated in Phase 2; (ii) a set of instructions for using IPD; and (iii) a set of rules describing certain properties of the synthetic process model to be discovered, thus mimicking expert knowledge (this is a prerequisite in IPD). Note that each analyst exploited a different set of event logs during the experiment. The participants were then asked to apply IPD and the selected APD techniques to discover the process models using the provided synthetic logs. In the first case (IPD), each participant interactively and actively created a process model for each log, exploiting the rules as domain knowledge. In doing so, each analyst was able to either accept (e.g. when suggesting the starting activity) or ignore (e.g. when suggesting that two activities with a specific relationship be placed in parallel) the suggestions derived from the dataset. In the second case (APD), all analysts used the default parameters of each technique. Thus, it was the APD technique that automatically created the process model.

4. **Model Comparison**. This step aimed at evaluating the accuracy of the discovered process models. More in detail, all maps produced by the IPD and APD techniques were assessed and compared in terms of fitness, precision, and F-score. To calculate fitness and precision, we used the alignment-based algorithms proposed by [67, 68], and due to the trade-off between fitness and precision [69, 70], we also used the F-score metric to ensure the model's balance fitness and precision [69].

5.2 Step 2. Case Study

We conducted the case study in close collaboration with the hospital staff, following three phases:

1. **Data Collection and Preparation**. We gathered and processed data of lung cancer patients, concerning their diagnostic, preoperative, surgical, and postoperative procedures. The treatment of lung cancer is complicated and based on decisions taken by practitioners from various disciplines, who must apply their in-depth knowledge and expertise. Thus, the direct involvement of a specialised multidisciplinary team is essential for accurately modelling this type of treatment processes. Data were extracted from Hospital Information Systems (HISs). As expected, the event log had several data quality issues, and thus we attempted

to refine the event log by (i) removing outliers; (ii) aggregating low-level activities; and (iii) filtering less frequent activities. The resulting event log consisted of about 1000 patient cases, 19 activities, and around 17000 events. Some quality issues still remained after data cleaning.

2. **Model Building**. To ensure against bias and affecting the experts, we first applied IPD to obtain the lung cancer patients process model, supported by the hospital staff. During model construction, we followed some of the suggestions provided by IPD, based on the information recorded in the event log (e.g., placing the activity "nuclear medicine" on the map), as the hospital staff deemed them appropriate. In other cases, we decided to deviate from the suggestions, deeming them incorrect (e.g., positioning the activity "x-ray" on the map), and add the activities into the correct positions according to the knowledge of the hospital staff. In addition, since the lab test may be performed several times and at different points in the process, its placement was critical. The suggestion from the event log was very vague, so it was necessary to intervene with the knowledge of the doctors to put the activity in the right points of the process. As in the synthetic experiment (step 1), we then applied IM, SM, and DFGM. Each APD technique was optimized by testing various parameter values and by selecting the configuration with the highest F-score [69]. The optimization was conducted by using a Rapid Miner extension, called RapidProM. The F-score was computed on Petri nets since the measuring tool work only on them. Conversions of the process model into Petri nets were done using ProM's package.

3. **Model Comparison**. The IPD model and the best configurations of APD techniques were assessed and compared, in terms of accuracy and conformance with clinical guidelines. As in the synthetic experiment, we measured fitness, precision, and F-score as a proxy for accuracy [37, 69]. To evaluate the conformance of the model to the Italian Association of Medical Oncology (AIOM) guidelines, we conducted a qualitative evaluation with the involvement of hospital staff, by defining a set of clinical rules that are necessary for the correct treatment of lung cancer patients and thus must be respected by each model. These rules were extracted from the AIOM guidelines and formalized by using a subset of (Declare) templates. Each template provides a way to specify a dependency between two different classes of activities (e.g., a precedence constraint between the activities involved in the classes "surgery" and "medical examination") [55].

6 Synthetic Experiment Results

This section presents the results of the controlled experiment, divided in two Sections (6.1, 6.2) for scenarios A and B, respectively.

Here, it is convenient to introduce the following notation for the synthetic event logs. In general, they are referenced as $L_a^{b\,x-y}$. The subscript a refers to the synthetic model used to generate the synthetic log. In the superscript $b \in \{o, d, n\}$, o denotes the *original event log* (i.e., with date and time registrations of the events) with 0% noise, d the event log with coarse granular timestamps (i.e., only the date of events is registered) and with 0% of noise, and n the noisy event log with coarse granular timestamps. Superscripts x and y refer to the percentages of noise due to removed and added events, respectively. When both x and y are zero, they are omitted for notation brevity.

6.1 Scenario A - Coarse Granular Timestamps

Table 1 shows the results of the experiment using the 12 synthetic event logs when applying the IPD and APD techniques. The percentage change in accuracy values calculated with noisy and good event logs is given in parentheses in Table 1.

We observe that the fitness of the models produced by IM, SM, and DFGM is almost always higher than that obtained by IPD. Conversely, IPD returns the Petri nets with the highest precision and F-score values in all cases. This is due to the poor quality of the event log (i.e., no time registration of the events). If there is no way to decide whether an activity A always precedes an activity B (or vice versa), APD techniques introduce a parallel construct or a combination of exclusive choice and loop into the resulting Petri nets. They then catch all traces but allow paths that do not match any of the traces in the input log, resulting in more imprecise models. In addition, a significant worsening of the APD technique performance (up to 56%) is observed in terms of precision and F-score when adding noise to event logs (see the values in parentheses in each table). Conversely, IPD appears to limit the negative effects of poor data quality, giving a slight reduction in precision and F-score values (up to 13%). In summary, by incorporating domain knowledge IPD can overcome the problems that arise when coarse granular timestamps are present within the event log and can provide more accurate process models than those obtained using APD techniques.

6.2 Scenario B - Duplicate Events

Table 2 reports the results of a comparative evaluation of the accuracy achieved using IPD and APD techniques. The percentage change in accuracy values calculated with noisy and good event logs is given in parentheses in Table 2.

As shown in these tables, APD techniques achieve higher fitness values than those obtained by IPD. Conversely, IPD yields the highest values for precision and F-scores in all cases. IM, SM, and DFGM do not allow the activities to be duplicated (i.e., two different occurrences of the activity A in the traces match a unique transition in the discovered Petri net), and thus they introduce a cycle within the discovered model. They can then admit paths that are not found in any of the traces, thus reducing the precision of the model. This becomes more evident when adding noisy behaviour. A more significant deterioration in the precision and F-score performance of APD techniques (up to 26%) compared to IPD can be observed with noisy event logs, while the fitness values remain similar for all techniques (see values in parentheses). In summary, incorporating domain knowledge can reduce the negative impact of imprecise events in the event logs on the accuracy of the process models. IPD can provide an accurate model that balances fitness and precision.

7 Case Study Results and Discussion

This section summarizes the results from the case study. Table 3 reports the fitness, precision, and F-score values obtained from the IPD and APD techniques, and the best configuration parameters for the APD techniques, while Table 4 summarizes the results obtained by all techniques in terms of the number of clinical rules met. The process models obtained by IPD and APD techniques are presented in Supplementary Material.

Table 3 indicates that with noisy real-life event logs, all techniques achieve a low level of accuracy in terms of F-score. However, IPD and SM yield acceptable values of 70 and 69, respectively.

Table 1: Quantitative evaluation of IPD, IM, DFGM, SM models discovered for scenario A

Process model M1 - Analyst 1					
Quality measures/Techniques/Logs		L_1^o	L_1^d	L_1^{n15-0}	L_1^{n30-30}
Fitness	IM	1	1 (0)	0.92 (-0.08)	0.86 (-0.14)
	DFGM	1	0.99 (-0.01)	0.89 (-0.11)	0.90 (-0.10)
	SM	1	0.97 (-0.03)	0.87 (-0.13)	0.86 (-0.14)
	IPD	1	0.99 (-0.01)	0.90 (-0.10)	0.75 (-0.25)
Precision	IM	1	0.75 (-0.25)	0.81 (-0.19)	0.45 (-0.55)
	DFGM	1	0.86 (-0.14)	0.95 (-0.05)	0.70 (-0.30)
	SM	1	0.90 (-0.10)	0.84 (-0.16)	0.80 (-0.20)
	IPD	1	1 (0)	1 (0)	1 (0)
F-score	IM	1	0.85 (-0.15)	0.86 (-0.14)	0.6 (-0.40)
	DFGM	1	0.92 (-0.08)	0.91 (-0.09)	0.78 (-0.22)
	SM	1	0.93 (-0.07)	0.85 (-0.15)	0.72 (-0.28)
	IPD	1	0.99 (-0.01)	0.94 (-0.06)	0.86 (-0.14)
Process model M2 - Analyst 2					
		L_2^o	L_2^d	L_2^{n15-0}	L_2^{n30-30}
Fitness	IM	1	1(0)	0.88 (-0.12)	0.82 (-0.18)
	DFGM	1	0.99 (-0.01)	0.93 (-0.07)	0.90 (-0.10)
	SM	1	0.97 (-0.03)	0.82 (-0.18)	0.78 (-0.22)
	IPD	1	1 (0)	0.79 (-0.21)	0.75 (-0.25)
Precision	IM	0.93	0.88 (-0.05)	0.28 (-0.69)	0.37 (-0.60)
	DFGM	0.87	0.68 (-0.21)	0.54 (-0.38)	0.52 (-0.40)
	SM	0.92	0.84 (-0.08)	0.65 (-0.29)	0.76 (-0.17)
	IPD	0.96	0.96 (0)	0.98 (0.02)	0.98 (0.02)
F-score	IM	0.96	0.93 (-0.03)	0.42 (-0.56)	0.51 (-0.46)
	DFGM	0.93	0.80 (-0.13)	0.68 (-0.26)	0.65 (-0.30)
	SM	0.95	0.90 (-0.05)	0.72 (-0.24)	0.76 (-0.20)
	IPD	0.97	0.97 (0)	0.87 (-0.10)	0.85 (-0.12)
Process model M3 - Analyst 3					
		L_3^o	L_3^d	L_3^{n15-0}	L_3^{n30-30}
Fitness	IM	1	1 (0)	0.88 (-0.12)	0.77 (-0.23)
	DFGM	1	1 (0)	0.89 (-0.11)	0.82 (-0.18)
	SM	1	0.90 (-0.10)	0.70 (-0.30)	0.82 (-0.18)
	IPD	1	0.99 (-0.01)	0.90 (-0.1)	0.78 (-0.22)
Precision	IM	0.97	0.78 (-0.19)	0.85 (-0.12)	0.68 (-0.29)
	DFGM	0.91	0.64 (-0.29)	0.80 (-0.12)	0.70 (-0.23)
	SM	0.96	0.86 (-0.10)	0.69 (-0.28)	0.73 (-0.23)
	IPD	0.98	0.94 (-0.04)	0.97 (-0.01)	0.96 (-0.02)
F-score	IM	0.97	0.78 (-0.19)	0.86 (-0.11)	0.72 (-0.25)
	DFGM	0.95	0.78 (-0.17)	0.84 (-0.11)	0.75 (-0.21)
	SM	0.96	0.87 (-0.09)	0.69 (-0.28)	0.77 (-0.19)
	IPD	0.98	0.96 (-0.02)	0.93 (-0.05)	0.85 (-0.13)

Table 2: Quantitative evaluation of IPD, IM, DFGM, SM models discovered for scenario B

Process model M4 - Analyst 1					
Quality measures/Techniques/Logs		L_4^o	L_4^{n15-0}	L_4^{n30-0}	L_4^{n30-30}
Fitness	IM	1	0.86 (-0.14)	0.86 (-0.14)	0.75 (-0.25)
	DFGM	1	0.93 (-0.07)	0.88 (-0.12)	0.86 (-0.14)
	SM	1	0.90 (-0.10)	0.84 (-0.16)	0.83 (-0.17)
	IPD	1	0.90 (-0.10)	0.83 (-0.17)	0.80 (-0.20)
Precision	IM	0.94	0.87 (-0.07)	0.89 (-0.05)	0.69 (-0.26)
	DFGM	0.89	0.82 (-0.07)	0.80 (-0.10)	0.77 (-0.13)
	SM	0.99	0.88 (-0.11)	0.80 (-0.19)	0.87 (-0.12)
	IPD	1	1 (0)	1 (0)	1 (0)
F-score	IM	0.96	0.86 (-0.10)	0.86 (-0.10)	0.72 (-0.25)
	DFGM	0.94	0.87 (-0.07)	0.83 (-0.11)	0.81 (-0.13)
	SM	0.99	0.89 (-0.10)	0.81 (-0.18)	0.84 (-0.15)
	IPD	1	0.95 (-0.05)	0.90 (-0.10)	0.88 (-0.12)
Process model M5 - Analyst 2					
		L_5^o	L_5^{n15-0}	L_5^{n30-0}	L_5^{n30-30}
Fitness	IM	1	0.86 (-0.14)	0.89 (-0.11)	0.76 (-0.24)
	DFGM	1	0.92 (-0.08)	0.86 (-0.14)	0.80 (-0.20)
	SM	1	0.88 (-0.12)	0.77 (-0.23)	0.78 (-0.22)
	IPD	1	0.86 (-0.14)	0.82 (-0.18)	0.77 (-0.23)
Precision	IM	0.96	0.93 (-0.03)	0.85 (-0.11)	0.66 (-0.31)
	DFGM	0.78	0.62 (-0.20)	0.60 (-0.23)	0.59 (-0.24)
	SM	0.94	0.87 (-0.07)	0.89 (-0.05)	0.88 (-0.06)
	IPD	1	1 (0)	1 (0)	1 (0)
F-score	IM	0.97	0.88 (-0.09)	0.87 (-0.10)	0.71 (-0.26)
	DFGM	0.87	0.74 (-0.14)	0.70 (-0.19)	0.67 (-0.22)
	SM	0.96	0.87 (-0.09)	0.82 (-0.14)	0.83 (-0.13)
	IPD	1	0.92 (-0.08)	0.90 (-0.10)	0.87 (-0.13)
Process model M6 - Analyst 3					
		L_6^o	L_6^{n15-0}	L_6^{n30-0}	L_6^{n30-30}
Fitness	IM	1	0.89 (-0.11)	0.95 (-0.05)	0.85 (-0.15)
	DFGM	1	0.94 (-0.06)	0.90 (-0.10)	0.89 (-0.11)
	SM	1	0.80 (-0.20)	0.79 (-0.21)	0.79 (-0.21)
	IPD	1	0.86 (-0.14)	0.78 (-0.22)	0.75 (-0.25)
Precision	IM	0.91	0.80 (-0.12)	0.70 (-0.23)	0.66 (-0.27)
	DFGM	0.91	0.56 (-0.38)	0.57 (-0.37)	0.60 (-0.34)
	SM	0.98	0.93 (-0.05)	0.88 (-0.10)	0.87 (-0.11)
	IPD	1	1 (0)	1 (0)	1 (0)
F-score	IM	0.95	0.84 (-0.11)	0.81 (-0.15)	0.74 (-0.22)
	DFGM	0.95	0.70 (-0.26)	0.69 (-0.27)	0.71 (-0.25)
	SM	0.98	0.85 (-0.13)	0.83 (-0.15)	0.81 (-0.17)
	IPD	1	0.92 (-0.08)	0.87 (-0.13)	0.85 (-0.15)

Table 3: Accuracy achieved by the IPD model and the best configurations of APD techniques

Techniques	Best parameters	Fitness	Precision	F-score
IM	1.0	0.67	0.51	0.58
DFGM	0.2	1	0.21	0.36
SM	0.9 & 0.0	0.81	0.61	0.69
IPD	-	0.70	0.71	0.70

Table 4: Rules satisfied by IM, DFGM, SM and IPD models

	Rules	IM	DFGM	SM	IPD
R1	X-ray should be executed immediately after the Surgical Procedure	0	0	1	1
R2	Invasive tests (ITs) should be preceded by radiological tests (RTs) & ITs should not be followed by RTs	0	0	0	1
R3	Surgical Procedure (SP) should be preceded by invasive tests (ITs) & SP should not be followed by ITs	0	0	0	1
R4	Surgical Procedure (SP) should be preceded by medical tests (MTs) & SP should not be followed by MTs	0	0	0	1
R5	The x-ray should be executed before removing the therapeutic aid	1	0	1	1
R6	The treatment should begin with a general examination	0	0	0	1
R7	Lab test should be executed at least 2 times within the process	1	1	1	1
	Total score	**2**	**1**	**3**	**7**

They can obtain a more precise process model despite inaccurate data, without penalising their fitness. Conversely, DFGM is unable to balance fitness and precision, and thus produces the lowest F-score (0.36), and allows behaviours not present in the event log, thus limiting the reliability of results.

However, the compliance of the model produced by APD techniques with the defined rules is low when compared to that of the IPD model. As shown in Table 4, only some of the rules were respected by SM, IM, and DFGM, as APD techniques consider neither organizational information nor domain knowledge in the care process. Their models fail to meet clinical guidelines if the input information is absent or incorrect, leading to inappropriate behaviours. By putting a human in the loop, IPD can discern the strict relationships that are not visible in the data due to their poor quality and can produce a compliant model.

As an example, we can analyse in detail the behaviour of each discovered model in terms of rule R6. The guidelines indicate that a general physical examination should be the starting point for the lung cancer process. However, the models produced by IM, SM, and DFGM do not respect this rule. The process in those generated by IM and DFGM may start with an activity other than the general physical examination (See Figs. 5 and 7). Similarly, the only starting activity in the SM model is the lab test, as shown in Fig. 6. This inconsistency is mainly due to timestamp granularity problems in the event log, which are not handled by APD techniques. As the events in the log are recorded in the order of days, the automated techniques find it difficult to understand the correct sequence of events. Conversely, the IPD technique can identify the appropriate care pathway due to the inclusion of expert knowledge, as shown in Fig. 8.

The main difference between APD techniques and IPD relies in the interaction with the experts,

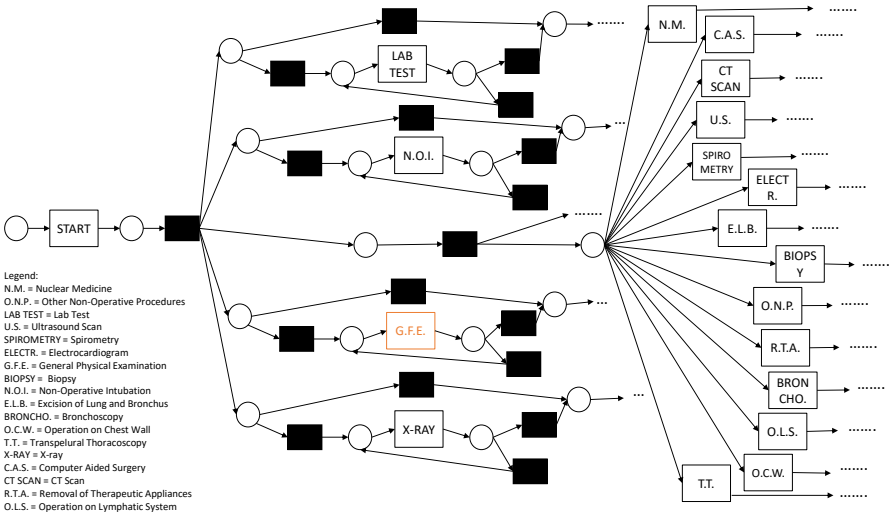

Figure 5: Extract of the Petri net produced by IM for lung cancer patients.

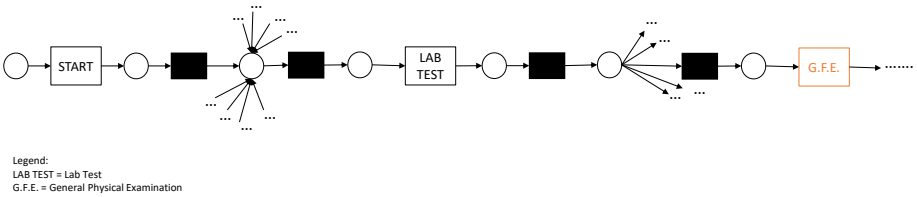

Figure 6: Extract of the Petri net transformed from SM for lung cancer patients.

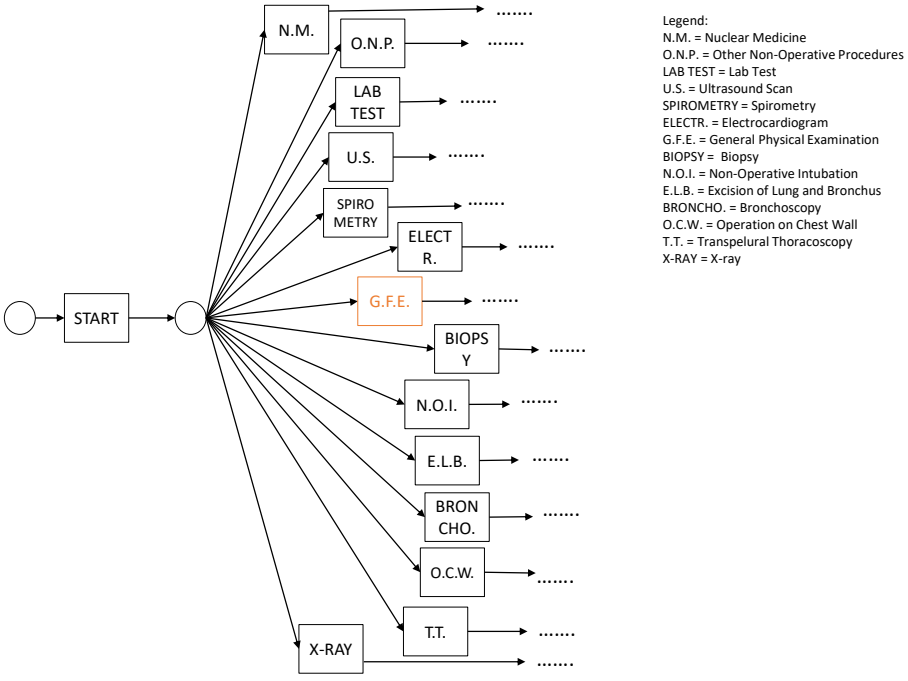

Figure 7: Extract of the Petri net converted from DFGM for lung cancer patients.

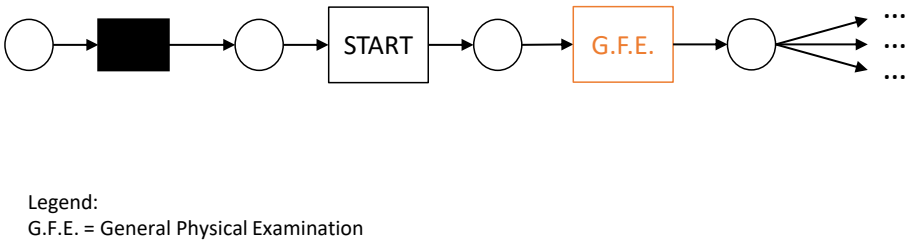

Figure 8: Extract of the Petri net produced by IPD for lung cancer patients.

ie., the hospital staff, during the discovery phase, which led to different results in terms of model quality. APD techniques provide automatic approaches, where users directly discover the process model from the event log only. Although the hospital staff acknowledged that the APD approach can be less time-consuming, they did not appreciate the lower quality of the process model and its poor adherence to clinical guidelines. On the contrary, the hospital staff have appreciated both the interactivity of IPD, since they can directly exploit their knowledge to conduct the discovery process incrementally, and the quality of the results in terms of accuracy and compliance with the guidelines.

To summarize, the results of the case study show that in the presence of noisy and large real-life event logs, all techniques perform worse than in the synthetic experiment. However, by exploiting domain knowledge, IPD is able to overcome the limitations caused by low data quality, and produces a process model that is both accurate and fully compliant with the guidelines. The other techniques produce inconsistent models that cannot be exploited by practitioners. These results suggest that IPD can be effectively used by healthcare managers conducting business process modelling projects when data quality issues occur, by considering their tacit knowledge.

8 Conclusions

In this study, we offer a more effective approach to data quality issues [23]. Data quality is critical to the success of any data-driven BPM project. This is particularly true in the healthcare context as data quality can impact on hospital managers and physicians' decision-making processes and have direct implications on the quality of healthcare delivery.

However, this issue has been scarcely investigated in the BPM literature [32, 33, 31, 34, 29]. Accordingly, we focused on data quality, by exploiting PM techniques to obtain a more effective business process modelling in the healthcare environment. In so doing, we responded to the need to develop new approaches to manage the problem of data quality in healthcare datasets [14].

More in detail, in our study, we demonstrated the effectiveness and suitability of IPD to deal with data quality issues in real complex contexts, like healthcare. IPD provides the analyst with a flexible way to interact with model construction, directly exploiting the domain knowledge along with the event log. Prior knowledge from domain experts represents a precious resource in the discovery of process models, providing critical advances with respects to automated discovery techniques [19].

Our study made two main contributions. First, we demonstrated the advantages of exploiting domain knowledge to improve the quality and comprehensibility of process models generated from noisy event logs. These include identifying the correct paths, discarding unacceptable sequences

or redundant activities, and positioning activities according to their purpose. This is valuable in healthcare where processes may require that certain activities have to follow strict relationships and the same activity occurs at various stages of the process with different aims, as shown in the case study. On the contrary, APD techniques do not have such advantages, as they solely use the information from the event logs, leading to unreliable and inaccurate results. This confirms suggestions in the literature that the capacity of APD techniques to produce comprehensible process models is limited in cases of noisy and large event logs [37, 51].

Second, we propose IPD as an alternative approach for data-driven business modelling projects in a real setting where quality problems occur, which cannot be solved in the pre-processing phase. This is the case for particular problems such as imprecise events and timestamps, which are typical for healthcare datasets. The presence of these problems can be due to the fact that data entry often requires manual action by clinicians or administrative staff. This may lead to errors, omissions, or delays in recording the activities or timestamps. No extant studies have addressed the problem of data quality in real applications during the process discovery phase. In our study, we confirm that an accurate and reliable process model can be obtained from IPD, even if generated by noisy real-life event logs. This represents the optimal output of a successful business modelling project and supports the evidence in the PM literature concerning the importance of integrating knowledge in the model discovery phase [47, 54]. The results from the controlled experiment and the healthcare case study provide the following insights regarding the impact of data quality issues on the final quality of PM results:

1. The presence of imprecise timestamps and/or events in the event log significantly worsen the performance of APD techniques, particularly in terms of precision and the related F-score. Conversely, IPD appears to limit the effects of such issues.

2. Noisy real-life event logs negatively affect the accuracy of the process models produced by APD and IPD techniques. However, IPD and SM can achieve acceptable results.

3. Noisy real-life event logs negatively impact compliance with guidelines or procedures in the models discovered using APD techniques, as they lead to incorrect or ambiguous behaviours. Conversely, IPD can reduce data quality problems through the inclusion of domain knowledge, as this compensates for missing or incorrect information and results in a fully compliant process model.

4. IPD can capture and incorporate tacit knowledge from domain experts such as medical staff within the model (e.g., how to manage x-ray activity during the treatment of lung cancer patients).

From a managerial perspective, this work provides clear guidance on how to effectively use IPD and how to leverage it, to ensure a data-driven business process modelling project is successful when data quality issues occur. Obtaining an accurate and comprehensible process model is valuable in the process analysis, monitoring, and improvement phases, especially in complex environments, like healthcare. As a starting point, it can be effective in detecting bottlenecks and deviations, efficient resource planning, and business process reengineering [26]. IPD can also effectively support health managers in quality assessment procedures, such as compliance with clinical guidelines, by providing reliable results. Finally, the model developed using IPD can be an input for simulation models or for assisting in the implementation of process-oriented Information Systems [71].

Although IPD is effective, it can be more expensive than APD techniques in terms of time. Only one activity can be added at a time in the model construction, and the user is only provided

with information concerning the control-flow aspect from the event log. Other data attributes can be useful when making decisions during the development of the process model. Finally, our case study is exploratory and may be affected by the bias of the context, i.e., the specific process (e.g., complexity, flexibility, and multi-disciplinarity) and the actors involved (highly specialized professionals).

We are currently planning to conduct a more in-depth assessment and replicate the study in other healthcare contexts, to further test the suitability of IPD in terms of addressing data quality issues and its generalizability. Furthermore, as future work, it may be interesting to test the interactivity of IPD or other promising functionalities to have a more complete evaluation of the technique. In addition, decision-making during the process discovery phase can be further supported through design improvements, such as by providing intermediate conformance results or additional information about event-level and case-level attributes. Finally, another future direction could be to allow the user to choose the type of model to be obtained in output, including the BPMN model. This may be valuable in healthcare as the Petri net is often difficult for doctors to interpret without the help of an expert.

References

[1] F. Lega, A. Prenestini, and P. Spurgeon, "Is management essential to improving the performance and sustainability of health care systems and organizations? a systematic review and a roadmap for future studies," *Value in Health*, vol. 16, pp. S46–S51, jan 2013.

[2] G. Gopal, C. Suter-Crazzolara, L. Toldo, and W. Eberhardt, "Digital transformation in healthcare – architectures of present and future information technologies," *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 57, pp. 328–335, dec 2018.

[3] S. Kraus, F. Schiavone, A. Pluzhnikova, and A. C. Invernizzi, "Digital transformation in healthcare: Analyzing the current state-of-research," *Journal of Business Research*, vol. 123, pp. 557–567, 2021.

[4] J. Mendling, B. T. Pentland, and J. Recker, "Building a complementary agenda for business process management and digital innovation," *European Journal of Information Systems*, vol. 29, pp. 208–219, may 2020.

[5] R. Vidgen, S. Shaw, and D. B. Grant, "Management challenges in creating value from business analytics," *European Journal of Operational Research*, vol. 261, pp. 626–639, sep 2017.

[6] N. Côrte-Real, P. Ruivo, and T. Oliveira, "Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value?," *Information & Management*, vol. 57, p. 103141, jan 2020.

[7] G. Hindle, M. Kunc, M. Mortensen, A. Oztekin, and R. Vidgen, "Business analytics: Defining the field and identifying a research agenda," *European Journal of Operational Research*, vol. 281, pp. 483–490, mar 2020.

[8] F. E. Horita, J. P. de Albuquerque, V. Marchezini, and E. M. Mendiondo, "Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in brazil," *Decision Support Systems*, vol. 97, pp. 12–22, may 2017.

[9] N. Gorla, T. M. Somers, and B. Wong, "Organizational impact of system quality, information quality, and service quality," *The Journal of Strategic Information Systems*, vol. 19, pp. 207–228, sep 2010.

[10] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," *International Journal of Production Economics*, vol. 165, pp. 234–246, jul 2015.

[11] A. P. Kurniati, E. Rojas, D. Hogg, G. Hall, and O. A. Johnson, "The assessment of data quality issues for process mining in healthcare using medical information mart for intensive care iii, a freely available e-health record database," *Health informatics journal*, vol. 25, no. 4, pp. 1878–1893, 2019.

[12] F. Fox, V. R. Aggarwal, H. Whelton, and O. Johnson, "A data quality framework for process mining of electronic health record data," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, jun 2018.

[13] R. Andrews, C. van Dun, M. Wynn, W. Kratsch, M. Röglinger, and A. ter Hofstede, "Quality-informed semi-automated event log generation for process mining," *Decision Support Systems*, vol. 132, p. 113265, may 2020.

[14] J. Munoz-Gama, N. Martin, C. Fernandez-Llatas, O. A. Johnson, M. Sepúlveda, E. Helm, V. Galvez-Yanjari, E. Rojas, A. Martinez-Millana, D. Aloini, I. A. Amantea, R. Andrews, M. Arias, I. Beerepoot, E. Benevento, A. Burattin, D. Capurro, J. Carmona, M. Comuzzi, B. Dalmas, R. de la Fuente, C. D. Francescomarino, C. D. Ciccio, R. Gatta, C. Ghidini, F. Gonzalez-Lopez, G. Ibanez-Sanchez, H. B. Klasky, A. P. Kurniati, X. Lu, F. Mannhardt, R. Mans, M. Marcos, R. M. de Carvalho, M. Pegoraro, S. K. Poon, L. Pufahl, H. A. Reijers, S. Remy, S. Rinderle-Ma, L. Sacchi, F. Seoane, M. Song, A. Stefanini, E. Sulis, A. H. ter Hofstede, P. J. Toussaint, V. Traver, Z. Valero-Ramon, I. van de Weerd, W. M. van der Aalst, R. Vanwersch, M. Weske, M. T. Wynn, and F. Zerbato, "Process mining for healthcare: Characteristics and challenges," *Journal of Biomedical Informatics*, vol. 127, p. 103994, mar 2022.

[15] C. Fernandez-Llatas, *Interactive Process Mining in Healthcare*. Springer, 2021.

[16] L. Vanbrabant, N. Martin, K. Ramaekers, and K. Braekers, "Quality of input data in emergency department simulations: framework and assessment techniques," *Simulation Modelling Practice and Theory*, vol. 91, pp. 83–101, 2019.

[17] N. Martin, "Using indoor location system data to enhance the quality of healthcare event logs: opportunities and challenges," in *International conference on business process management*, pp. 226–238, Springer, 2018.

[18] I. Davidson and G. Tayi, "Data preparation using data quality matrices for classification mining," *European Journal of Operational Research*, vol. 197, pp. 764–772, sep 2009.

[19] P. M. Dixit, H. M. W. Verbeek, J. C. A. M. Buijs, and W. M. P. van der Aalst, "Interactive data-driven process model construction," in *Conceptual Modeling*, pp. 251–265, Springer International Publishing, 2018.

[20] P. M. Dixit, J. C. A. M. Buijs, and W. M. P. van der Aalst, "ProDiGy : Human-in-the-loop process discovery," in *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, IEEE, may 2018.

[21] R. Andrews, F. Emamjome, A. H. ter Hofstede, and H. A. Reijers, "An expert lens on data quality in process mining," in *2020 2nd International Conference on Process Mining (ICPM)*, IEEE, oct 2020.

[22] R. S. Mans, R. Hajo, B. Hans, B. Wasana, and P. Rogier, "Business process mining success," in *21st European Conference on Information Systems, ECIS 2013*, 2013.

[23] R. P. J. C. Bose, R. S. Mans, and W. M. P. van der Aalst, "Wanna improve process mining results?," in *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pp. 127–134, IEEE, apr 2013.

[24] S. Fan, Z. Hua, V. C. Storey, and J. L. Zhao, "A process ontology based approach to easing semantic ambiguity in business process modeling," *Data & Knowledge Engineering*, vol. 102, pp. 57–77, mar 2016.

[25] R. S. Aguilar-Savén, "Business process modelling: Review and framework," *International Journal of Production Economics*, vol. 90, pp. 129–149, jul 2004.

[26] M. Dumas, M. La Rosa, J. Mendling, H. A. Reijers, *et al.*, *Fundamentals of business process management*, vol. 1. Springer, 2013.

[27] S. Windle, "H. Smith and P. Fingar, Business Process Management (BPM): the Third Wave," *Journal of Information Systems*, vol. 18, no. 1, pp. 128–131, 2004.

[28] L. Sánchez-González, F. García, F. Ruiz, and J. Mendling, "Quality indicators for business process models from a gateway complexity perspective," *Information and Software Technology*, vol. 54, pp. 1159–1174, nov 2012.

[29] I. M.-M. de Oca, M. Snoeck, H. A. Reijers, and A. Rodríguez-Morffi, "A systematic literature review of studies on business process modeling quality," *Information and Software Technology*, vol. 58, pp. 187–205, feb 2015.

[30] D. L. Moody, "Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions," *Data & Knowledge Engineering*, vol. 55, pp. 243–276, dec 2005.

[31] W. Bandara, G. G. Gable, and M. Rosemann, "Factors and measures of business process modelling: model building through a multiple case study," *European Journal of Information Systems*, vol. 14, pp. 347–360, dec 2005.

[32] O. Lindland, G. Sindre, and A. Solvberg, "Understanding quality in conceptual modeling," *IEEE Software*, vol. 11, pp. 42–49, mar 1994.

[33] J. Krogstie, G. Sindre, and H. Jørgensen, "Process models representing knowledge for action: a revised quality framework," *European Journal of Information Systems*, vol. 15, pp. 91–102, feb 2006.

[34] W. Bandara, G. G. Gable, M. Tate, and M. Rosemann, "A validated business process modelling success factors model," *Business Process Management Journal*, vol. ahead-of-print, may 2021.

[35] C. D. S. Garcia, A. Meincheim, E. R. F. Junior, M. R. Dallagassa, D. M. V. Sato, D. R. Carvalho, E. A. P. Santos, and E. E. Scalabrin, "Process mining techniques and applications – a systematic mapping study," *Expert Systems with Applications*, vol. 133, pp. 260–295, nov 2019.

[36] Á. Rebuge and D. R. Ferreira, "Business process analysis in healthcare environments: A methodology based on process mining," *Information Systems*, vol. 37, pp. 99–116, apr 2012.

[37] W. M. P. Van der Aalst, *Process Mining*. Springer Berlin Heidelberg, 2016.

[38] C. Di Ciccio, A. Marrella, and A. Russo, "Knowledge-intensive processes: Characteristics, requirements and analysis of contemporary approaches," *Journal on Data Semantics*, vol. 4, pp. 29–57, apr 2014.

[39] A. Weijters and A. A. van der Aalst, W. M. P. and De Medeiros, "Process mining with the heuristics miner-algorithm," *Technische Universiteit Eindhoven, Tech. Rep. WP*, vol. 166, pp. 1–34, 2006.

[40] A. Weijters and J. Ribeiro, "Flexible heuristics miner (FHM)," in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, IEEE, apr 2011.

[41] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, "Discovering block-structured process models from event logs containing infrequent behaviour," in *Business Process Management Workshops*, pp. 66–78, Springer International Publishing, 2014.

[42] C. W. Günther and W. M. P. van der Aalst, "Fuzzy mining – adaptive process simplification based on multi-perspective metrics," in *Lecture Notes in Computer Science*, pp. 328–343, Springer Berlin Heidelberg, 2007.

[43] A. Augusto, R. Conforti, M. Dumas, M. L. Rosa, and A. Polyvyanyy, "Split miner: automated discovery of accurate and simple business process models from event logs," *Knowledge and Information Systems*, vol. 59, pp. 251–284, may 2018.

[44] M. Pegoraro, M. S. Uysal, and W. M. P. van der Aalst, "Discovering process models from uncertain event data," in *Business Process Management Workshops*, pp. 238–249, Springer International Publishing, 2019.

[45] M. Pegoraro and W. M. P. van der Aalst, "Mining uncertain event data in process mining," in *2019 International Conference on Process Mining (ICPM)*, IEEE, jun 2019.

[46] M. Pegoraro, M. S. Uysal, and W. M. van der Aalst, "Conformance checking over uncertain event data," *Information Systems*, vol. 102, p. 101810, dec 2021.

[47] A. Bottrighi, L. Canensi, G. Leonardi, S. Montani, and P. Terenziani, "Interactive mining and retrieval from process traces," *Expert Systems with Applications*, vol. 110, pp. 62–79, nov 2018.

[48] B. N. Yahya, M. Song, H. Bae, S.-o. Sul, and J.-Z. Wu, "Domain-driven actionable process model discovery," *Computers & Industrial Engineering*, vol. 99, pp. 382–400, 2016.

[49] X. Lu, D. Fahland, and W. M. P. van der Aalst, "Conformance checking based on partially ordered event data," in *Business Process Management Workshops*, pp. 75–88, Springer International Publishing, 2015.

[50] J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "On the role of fitness, precision, generalization and simplicity in process discovery," in *On the Move to Meaningful Internet Systems: OTM 2012*, pp. 305–322, Springer Berlin Heidelberg, 2012.

[51] C. Diamantini, L. Genga, and D. Potena, "Behavioral process mining for unstructured processes," *Journal of Intelligent Information Systems*, vol. 47, pp. 5–32, feb 2016.

[52] D. Schuster, S. J. van Zelst, and W. M. van der Aalst, "Utilizing domain knowledge in data-driven process discovery: A literature review," *Computers in Industry*, vol. 137, p. 103612, may 2022.

[53] G. Greco, A. Guzzo, F. Lupia, and L. Pontieri, "Process discovery under precedence constraints," *ACM Transactions on Knowledge Discovery from Data*, vol. 9, pp. 1–39, jun 2015.

[54] A. J. Rembert, A. Omokpo, P. Mazzoleni, and R. T. Goodwin, "Process discovery using prior knowledge," in *Service-Oriented Computing*, pp. 328–342, Springer Berlin Heidelberg, 2013.

[55] F. M. Maggi, R. P. J. C. Bose, and W. M. P. van der Aalst, "A knowledge-based integrated approach for discovering and repairing declare maps," in *Advanced Information Systems Engineering*, pp. 433–448, Springer Berlin Heidelberg, 2013.

[56] D. Schuster, S. J. van Zelst, and W. M. P. van der Aalst, "Cortado—an interactive tool for data-driven process discovery and modeling," in *Application and Theory of Petri Nets and Concurrency*, pp. 465–475, Springer International Publishing, 2021.

[57] W. van der Aalst, "On the representational bias in process mining," in *2011 IEEE 20th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*, IEEE, jun 2011.

[58] N. Martin, A. Martinez-Millana, B. Valdivieso, and C. Fernández-Llatas, "Interactive data cleaning for process mining: A case study of an outpatient clinic's appointment system," in *Business Process Management Workshops*, pp. 532–544, Springer International Publishing, 2019.

[59] J. Desel and J. Esparza, *Free choice Petri nets*, vol. 40. Cambridge university press, 2005.

[60] Hevner, March, Park, and Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, p. 75, 2004.

[61] J. Recker, B. Mutschler, and R. Wieringa, "Empirical research in business process management: introduction to the special issue," *Information Systems and e-Business Management*, vol. 9, pp. 303–306, aug 2010.

[62] J. Recker and J. Mendling, "The state of the art of business process management research as published in the BPM conference," *Business & Information Systems Engineering*, vol. 58, pp. 55–72, nov 2015.

[63] M. T. Wynn and S. Sadiq, "Responsible process mining - a data quality perspective," in *Lecture Notes in Computer Science*, pp. 10–15, Springer International Publishing, 2019.

[64] L. Canensi, G. Leonardi, S. Montani, and P. Terenziani, "A context-aware miner for medical processes," *Journal of e-Learning and Knowledge Society*, vol. 14, no. 1, 2018.

[65] R. S. Mans, W. M. P. van der Aalst, and R. J. B. Vanwersch, *Process Mining in Healthcare*. Springer International Publishing, 2015.

[66] W. M. P. Van der Aalst, R. De Masellis, C. Di Francescomarino, and C. Ghidini, "Learning hybrid process models from events," in *Lecture Notes in Computer Science*, pp. 59–76, Springer International Publishing, 2017.

[67] A. Adriansyah, B. F. van Dongen, and W. M. P. van der Aalst, "Conformance checking using cost-based fitness analysis," in *2011 IEEE 15th International Enterprise Distributed Object Computing Conference*, IEEE, aug 2011.

[68] A. Adriansyah, J. Munoz-Gama, J. Carmona, B. F. van Dongen, and W. M. P. van der Aalst, "Measuring precision of modeled behavior," *Information Systems and e-Business Management*, vol. 13, pp. 37–67, jan 2014.

[69] J. De Weerdt, M. De Backer, J. Vanthienen, and B. Baesens, "A robust f-measure for evaluating discovered process models," in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, IEEE, apr 2011.

[70] M. F. Sani, S. J. van Zelst, and W. M. P. van der Aalst, "Improving process discovery results by filtering outliers using conditional behavioural probabilities," in *Business Process Management Workshops*, pp. 216–229, Springer International Publishing, 2018.

[71] L. Măruşter and N. R. T. P. van Beest, "Redesigning business processes: a methodology based on simulation and process mining techniques," *Knowledge and Information Systems*, vol. 21, pp. 267–297, jun 2009.