How to Evaluate Bias from Hospital Data: Automatic Preprocessing of Patient Pathways for Data Analysis

Laura Uhl, Vincent Augusto, Benjamin Dalmas, Youenn Alexandre, Paolo Bercelli, Fanny Jardinaud, Saber Aloui

Submitted to: JMIR Medical Informatics
on: March 29, 2024

Table of Contents

Original Manuscript **5**
Supplementary Files **53**
Figures **54**
Figure 1 ..... **55**
Figure 2 ..... **56**
Figure 3 ..... **57**
Figure 4 ..... **58**
Figure 5 ..... **59**
Multimedia Appendixes **60**
Multimedia Appendix 1 ..... **61**
Multimedia Appendix 2 ..... **61**
Multimedia Appendix 3 ..... **61**
Multimedia Appendix 4 ..... **61**
Multimedia Appendix 5 ..... **61**

How to Evaluate Bias from Hospital Data: Automatic Preprocessing of Patient Pathways for Data Analysis

Laura Uhl[1] MSc, ING; Vincent Augusto[1] Prof Dr; Benjamin Dalmas[1] PhD; Youenn Alexandre[2] PhD; Paolo Bercelli[2] MD; Fanny Jardinaud[3] PhD; Saber Aloui[4] PhD

[1]Mines Saint-Etienne Centre CIS UMR 6158 LIMOS CNRS Saint-Etienne FR
[2]Groupe Hospitalier Bretagne Sud Université de Bretagne Occidentale Lorient FR
[3]Direction Anticipation & Usages Enovacom Marseille FR
[4]Inserm, UMR 1085 Ester CHU Angers Angers FR

Corresponding Author:
Laura Uhl MSc, ING
Mines Saint-Etienne Centre CIS
UMR 6158 LIMOS
CNRS
158 cours Fauriel
Saint-Etienne
FR

Abstract

Background: Optimisation of patient care pathways is crucial for hospital managers in a context of a scarcity of medical resources. Assuming unlimited capacities, the pathway of a patient would only be governed by pure medical logic, to meet at best patient's needs. However, logistical limitations (e.g., resources such as inpatient beds) are often associated with delayed treatments and may ultimately affect patient pathways. This is especially true for unscheduled patients: when a patient at the emergency department needs to be admitted to another medical unit without disturbing the flow of planned hospitalisations.

Objective: In this study, we propose a new framework to automatically detect activities in patient pathways which may be unrelated to patients' needs and rather induced by logistical limitations.

Methods: The scientific contribution lies in a method that turns a database of history pathways with bias into two databases: (i) a labelled pathways database where each activity is labelled as relevant (related to patient's need) or irrelevant (induced by logistical limitations) and (ii) a corrected pathways database where each activity corresponds to the activity that would occur assuming unlimited resources. The labelling algorithm is assessed through medical expertise. Two case studies quantify the impact of our preprocessing method of healthcare data by using respectively process mining and discrete event simulation.

Results: Focusing on unscheduled patient pathways, we collected data covering 12 months of activity at the Groupe Hospitalier Bretagne Sud in France. Our algorithm has an error of 13% and has demonstrated its usefulness to preprocess traces and obtain a clean database. The two case studies show the importance of our preprocessing step before any analysis.

Conclusions: Patient pathways data reflect the actual activity of hospitals, governed by medical requirement and logistical limitation. Before any use of these data, these limitations should be identified and corrected. We anticipate the generalisation of our approach to obtain unbiased analyses of patient pathways for other hospitals. Clinical Trial: The study was approved by the French Data Protection Authority (CNIL) under the number 922243.

(JMIR Preprints 29/03/2024:58978)
DOI: https://doi.org/10.2196/preprints.58978

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

How to Evaluate Bias from Hospital Data: Automatic Preprocessing of Patient Pathways for Data Analysis

Laura Uhl. CNRS UMR 6158 LIMOS, Mines Saint-Etienne, Centre CIS. Saint-Etienne, France.
Vincent Augusto. CNRS UMR 6158 LIMOS, Mines Saint-Etienne, Centre CIS. Saint-Etienne, France.
Benjamin Dalmas. CNRS UMR 6158 LIMOS, Mines Saint-Etienne, Centre CIS. Saint-Etienne, France.
Youenn Alexandre. Groupe Hospitalier Bretagne Sud. Lorient, France.
Fanny Jardinaud. Direction Anticipation & Usages, Enovacom. Marseille, France.
Paolo Bercelli. Groupe Hospitalier Bretagne Sud. Lorient, France.
Saber Aloui. Centre Hospitalier Universitaire d'Angers, Inserm UMR 1085 Ester. Angers, France.

Corresponding Author: Laura Uhl, 158 cours Fauriel, 42000, Saint-Etienne, France. +33477420123, l.uhl@emse.fr

Abstract

Background: The optimisation of patient care pathways is crucial for hospital managers in the context of a scarcity of medical resources. Assuming unlimited capacities, the pathway of a patient would only be governed by pure medical logic, to meet at best the patient's needs. However, logistical limitations (e.g., resources such as inpatient beds) are often associated with delayed treatments and may ultimately affect patient pathways. This is especially true for unscheduled patients: when a patient in the emergency department needs to be admitted to another medical unit without disturbing the flow of planned hospitalisations.
Objective: In this study, we propose a new framework to automatically detect activities in patient pathways that may be unrelated to patients' needs but rather induced by logistical limitations.
Methods: The scientific contribution lies in a method that transforms a database of history pathways with bias into two databases: (i) a labelled pathways database where each activity is labelled as relevant (related to patient's need) or irrelevant (induced by logistical limitations) and (ii) a corrected pathways database where each activity corresponds to the activity that would occur assuming unlimited resources. The labelling algorithm is assessed through medical expertise. Two case studies quantify the impact of our method of preprocessing healthcare data using process mining and discrete event simulation.
Results: Focusing on unscheduled patient pathways, we collected data covering 12 months of activity at the Groupe Hospitalier Bretagne Sud in France. Our algorithm has 87% accuracy and has demonstrated its usefulness for preprocessing traces and obtaining a clean database. The two case studies show the importance of our preprocessing step before any analysis. The process graphs of the processed data have on average 40% fewer variants than the raw data. The simulation revealed that 30% of the medical units had greater than one bed difference in capacity between the processed and raw data.
Conclusions: Patient pathway data reflect the actual activity of hospitals that is governed by medical requirements and logistical limitations. Before using these data, these limitations

should be identified and corrected. We anticipate that our approach can be generalised to obtain unbiased analyses of patient pathways for other hospitals.
Trial Registration: The study was approved by the French Data Protection Authority (CNIL) under the number 922243.

Keywords: preprocessing, framework, healthcare data, patient pathway, bed management

Introduction

Context

Bed management is a critical task for hospitals to provide coherent care pathways. Daily bed management consists of finding beds for patients coming from the emergency department (ED) in appropriate medical units without cancelling planned hospitalisations. Therefore, bed management involves two distinct flows: unscheduled flow (life-threatening emergencies and patients coming to the emergency department) and scheduled flow (planned hospitalisations). Despite the complexity of the task, bed management is most often organised without the help of any decision support tools and involves multiple phone calls to find a bed in a medical unit matching the patient's needs [3]. When medical units are facing high occupation rates, it is not always possible to find a bed to match patient needs.
In these situations, patients are either kept in the short stay hospitalisation unit of the ED, or transferred to an overflow medical unit, to wait for a bed. Consequently, the medical units visited by a patient do not always correspond to her or his medical needs. For example, a patient from the ED can be transferred into a surgery unit and then into a cardiology unit. This is the pathway observed in the data. The patient did not receive any surgical treatment. He was admitted to the surgery unit waiting for a cardiology unit bed. Therefore, the location of the patient does not always match the cause of hospitalisation. The succession of medical units is called a patient pathway. Unscheduled pathways describe the pathways of patients coming from the ED. In this work, we only considered patients who visited the ED and were subsequently hospitalised.
The study of patient pathways reveals several challenges due to the variety of pathways, the lack of complete guidelines and references, and the heterogeneity of patient management between hospitals (due to equipment and organisational differences). Unscheduled pathways are difficult to explain because management rules or clear indicators are not available to identify them. In addition, the high number of pathway variants makes individual studies of each pathway impossible (e.g. more than a thousand variants for French hospitals of average size) [26]. Process mining is an interesting tool for studying a set of pathways with several variants because a pathway can be seen as a patient care process [26, 32]. Nevertheless, a large variance in pathways leads to uninterpretable process graphs. Strategies exist to make a process graph easier to read, such as trace clustering or graph size reduction using filters or aggregation [26], but these methods cannot identify which activities are relevant, and which activities are induced by logistical limitations.
In this paper, we sought to develop a method to assess observed pathways extracted from a hospital information system. We want to identify which medical units match the cause of hospitalisation (relevant) or not (irrelevant) in a patient pathway. The medical relevance or the relevance of treatments is not evaluated, nor is the choice of the bed manager. Only the relevance of the patient's location was evaluated. An "irrelevant" medical unit means that the patient would have been hospitalised in another unit if there were an infinite number of beds. The identification of such "irrelevance" is important to avoid any misinterpretation of further analysis results. In this paper, we often use word *bias* to denote a wrong, inaccurate, or incomplete interpretation of a real situation because the data do not represent reality. We use expression bias in pathways or data bias, to refer to data that represent pathways that do not always correspond to patients' medical needs.

Related work

We did not find proper literature on the task of assessing pathways, but rather heterogeneous papers dealing with bias or phases of a pathway. In 1989, Selker et al. [32] designed the “Delay Tool”, which detects medically unnecessary hospital days. It is based on a taxonomy of delays. Each stay was manually evaluated using patient records with the Delay Tool method. In an article on the prediction of the disposition of emergency department patients, El Bouri et al. [13] considered the fact that ED patients can be admitted to an inpatient unit inappropriate for their diagnosis. Patients were filtered according to whether their primary diagnosis code for the ED visit clearly corresponded to the admission inpatient unit. Their aim was to avoid learning from biased data. These methods require a thesaurus of all possible diagnoses linked to appropriate wards. To study patient pathways, Franck et al. [17], designed a generic framework to model pathways and distinguished three different phases: (1) a waiting phase – the patient waits in the ED (unscheduled) or at home (scheduled) to be admitted to the relevant medical unit; (2) an acute phase – the patient receives care in the medical unit; and (3) a rehabilitative phase – rehabilitative care of the patient. They also differentiated scheduled patients from unscheduled patients. To analyse the clinical pathways, they defined relevant pathways for each type of patient by considering only the acute phase and substitution options. To identify the relevant pathways and substitutions, they used process mining on administrative data. This method is very accurate but time-consuming given that a relevant pathway and substitutions must be defined for each pathology. They applied this method exclusively to stroke patients.

Data quality in health research is a shared problem and solutions have been proposed to improve several dimensions of quality [4]. However, methods are often not suggested to correct specific bias in healthcare data due to missing details about a piece of information.

Patient pathways can be seen as processes, with the succession of medical units being the succession of events. Therefore, pathways can be studied with process mining techniques. “The goal of process mining is to use event data to extract process-related information” [36]. The first rough representation of patient pathways using process discovery algorithms provides a spaghetti-like process model. Indeed, process discovery algorithms are not successful with event logs that involve numerous variants and many events [26]. A typical solution to untangle a spaghetti-like model is to cluster the whole set of traces (trace clustering) and represent each cluster using a process model that should be smaller and more comprehensive. The main challenges of the clustering of patient pathways are the integration of medical knowledge (medical logic) and the evaluation of the resulting clusters. In the literature, several methods for trace clustering have been proposed. Some of these methods are distance-based clustering algorithms. The core of these methods is to compute distances between traces to apply classic clustering algorithms (trace clustering [34], trace clustering based on conserved patterns [5], context aware clustering [6], and Delias’ method [11]). Others are model-based; these methods gradually build a process model that represents a cluster, and a trace is assigned to the cluster with the nearest process model (sequence clustering [41], active trace clustering [10], disjunctive workflow schema [18], graph-based approach and Markov models [14], and behavioural topic analysis [20]). In terms of cluster evaluation, different metrics are used. Some metrics analyse cluster intrahomogeneity and others analyse the complexity of the process model of each cluster. There is no consensus on these metrics, and they do not guarantee that the clusters computed using the algorithm have an expert logic. Hence, trace clustering does not give us complete satisfaction in characterising pathways. Another approach for simplifying process models was proposed by Fahland and van der Aals [16] based on unfolding.

Some data preparation techniques can also reduce the “spaghettiness” of process models. Data preparation is an unavoidable step in a process mining project and impacts on the resulting process

graph, as highlighted by De Roock and Martin in their most recent state-of-the-art study [9]. Several methods have been suggested in the literature to simplify process models. Semantic log purging was proposed by Ly et al. [23] in 2012 to clean log data. This method is based on the identification of "fundamental constraints that a process has to obey" thanks to experts. Only a qualitative evaluation and one experiment using one dataset were performed. Van Zelst et al. [38] reviewed the literature on event abstraction in process mining. However, this technique is not related to the problem addressed in this paper because our dataset does not provide information on the granularity of events. Several papers address the issue of timestamp inaccuracy.
Martin et al. [24] proposed interactive data cleaning. Dixit et al. [12] created a method to detect and repair event ordering mistakes. Rogge-Solti et al. [31] presented a similar approach to repairing missing events based on alignment. In addition, these researchers created a method for time repairing.

To rigorously prepare the data and event log, different frameworks have been developped. Andrew et al. [2] applied the Cross Industry Standard Process for Data Mining (CRISP-DM) method to identify data quality issues. The data quality dimensions used in data mining are also useful for assessing data quality in process mining. Nevertheless, the researchers do not consider dimensions specific to processes such as trace coherence. Therefore, the fourth step, namely pre-study process mining analysis, is important to assess this dimension. Bose et al. [7] noted 27 event log quality issues based on four categories (missing data, incorrect data, imprecise data, and irrelevant data) and nine components of an event log (case, event, belongs to, case attributes, event attributes, position, activity name, timestamp, and resource). The researchers also distinguished four process characteristics: (1) voluminous data, which refers to a large number of cases or events, (2) case heterogeneity, which refers to a large number of distinct traces, (3) event granularity, which refers to a large number of distinct activities, and (4) process flexibility and concept drifts. The issues caused by case heterogeneity; are a part of the problem we attempt to address in this paper. Van Eck et al. [37] suggested PM^2, a process mining project methodology. Data processing is the third step and consists of creating views (creating the event log), aggregating events, enriching logs (addition of attributes), and filtering logs. Vanbrabant et al. [39] presented a data quality framework based on three previous frameworks and applied it to a case study: pretreatment of emergency department data before simulation. These researchers divided quality problems into hierarchical classes. Verhulst [42] defined very precisely the different data quality dimensions for process mining and their scoring methods. All these papers on data preparation are general to process mining dataset and do not answer the question of pathway bias.

Process mining is not the unique method used to analyse patient pathways and the method can be combined with other methods such as discrete event simulation (DES). Prodel et al. [28, 29] developed a framework to automatically convert a process model discovered with process mining into a simulation model of clinical pathways. Abohamad et al. [1] used process mining to discover emergency department processes and then used DES to study bottlenecks. Wood and Murch [43] modelled patient pathways with Markov chains to study transfer delays between medical units and discharge delays. Karakra et al. [21] also used a DES to model an emergency department and added a real-time connection to real-time patient data to create a digital twin. The digital twin of the patient enables the monitoring of his or her pathway and activities as well as near future predictions. Some models reproduce an entire hospital. Holm et al. used a DES to model an entire hospital and patient flows through the wards and to determine bed utilisation. Demir et al. used a similar model to anticipate an increase in the number of patients and to adapt resources. Ordu et al. [27] achieved an even more complete model of a hospital and patient flows.

To conclude, this review of the literature reveals that process mining and simulation are the principal

methods used to study patient pathways. Process mining is a standard tool for discovering patient pathways (or other healthcare processes), but important limitations are noted in the literature, especially the complexity of the model graphs. Data preparation techniques and clustering methods are suggested to compensate for this issue. Some methods are based on expert interviews/expert knowledge integration. This is similar to our construction of rules. Several papers focus on timestamp correction or missing events or labels but none of the studies focus on our problem of biased events. Simulations do not consider input data quality. In this paper, we focus on enriching the log by adding an attribute that can define an event as relevant or irrelevant.

Objectives

The objective of this paper is twofold: (1) building a framework to model and analyse patient pathways, and (2) proposing a method to automatically identify bias in patient pathway data. In other words, this method turns a database of observed pathways with bias into two databases, one database includes labelled pathways (identified bias) and one database includes of corrected pathways (without bias). Such a method is intended to ease the preprocessing of real data for data analysts or hospital managers who seek a clean database with unbiased medical pathways.
The scientific contributions of this paper are listed below:

1. This study provides new framework to model patient pathways taking into account hospital management constraints (e.g. bed occupancy, resource availability). This framework is used to assess patient pathways.
2. This study develops a new method to automatically label and correct pathways based on hospital data. Pathway labelling aims to identify the steps in a patient pathway due to a difficult bed management, and path correction aims to correct irrelevant activities. The labelling algorithm is assessed by comparing its outputs with the experts' answers.
3. Two case studies are reported.
   i A quantitative comparison of the observed pathways database and the corrected pathways database is performed based on process mining using process model comparison and classical process mining indicators.
   ii A discrete-event simulation model based on the observed pathways database and the corrected pathways database was used to evaluate the impacts of data correction on the occupation of medical units.

Methods

Unscheduled Hospital Pathway Modelling Framework

Formal Definition of the Framework for the Study Patient Pathway

In this section we propose a set of definitions that will be used to formalise the unscheduled hospital pathway modelling framework.
In this article, we are interested in medicine, surgery, and obstetric *medical units* (MCO). In French the initials MCO stand for Médecine, Chirurgie, Obstétrique et Odontologie. The medical units belong to a *hospital* that itself can belong to an *hospital group*. Figure 1 represents the dependencies among the hospital group, *site*, and medical unit. Here we are interested in the pathways inside the same hospital group, which we call the *MCO-stay*. See Multimedia Appendix 5 for detailed definitions of the abovementioned concepts.

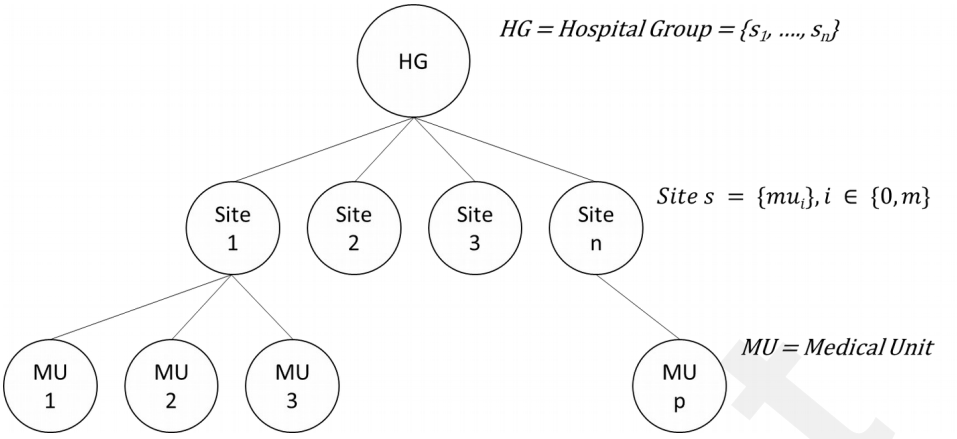


Figure 1. Hospital group structure.

The hospital pathways are defined using a process mining formalism [36].

Definition 1 (*Event*)**.** Let *E* be the event universe, i.e. the set of all possible event identifiers, E* is the set of all sequences over E and *T* is the time domain. We assume that events are defined by several attributes; however, the case id, timestamp, and activity name are mandatory for case identification, trace ordering, and event labelling respectively.

Let *AN* be a set of attribute names. For any event $e \in E$ and name z $\in AN$, $\#_n(e)$ is the value of the attribute *z* for event *e*. We consider $\#_{activity} \in E \rightarrow A$ and $\#_{time} \in E \rightarrow T$ functions that assign an activity name from a finite set of process activities A and a timestamp respectively to each event. For convenience, we assume the following standard attributes:

- $\#_{activity}$(e) is the activity associated with event e.
- $\#_{time}$(e) is the timestamp of event e.
- $\#_{trans}$(e) is the transaction type associated with event e, examples are schedule, start, complete, and suspend.

The transaction type attribute $\#_{trans}$(e) refers to the life cycle of activities. In most situations, activities take time. Therefore, events may point out, for example, the start or completion of activities.

Definition 2 (*Trace*)**.** A trace is a finite sequence of events denoted as $\sigma = < e_1, e_2, ..., e_n > \forall e_i \in E^*$ such that each event appears only once: $e_i \neq e_j$ for $1 \leq i < j \leq |\sigma|$.

Specifically, $|\sigma|$ denotes the length of the trace. Here $|\sigma| = n$.

Definition 3 (*Stage*)**.** In a trace, several events can have the same activity name. In the following, the subset of events with the same activity name that contains a single start event and a single completion event subsequent to the start event is referred to as a stage:

Let e_i and e_m such that:

- $\#_{activity}(e_i) = \#_{activity}(e_m) = a$,
- $\#_{trans}(e_i)$ = start, $\#_{trans}(e_m)$ = complete,
- $\#_{time}(e_i) = t_i \leq \#_{time}(e_m) = t_m$,
- $\nexists$ e_j such that $\#_{time}(e_j) = t_j \geq t_i$ and $t_j \leq t_m$, $\#_{activity}(e_j) = a$ and $\#_{trans}(e_j)$ = completion.

$$\text{Stages} = \{e_j | \#_{activity}(e_j) = a \text{ and } t_i \leq t_j \leq t_m\}$$

The duration of a stage is defined as the time between the start of the stage and its completion:

$$\#_{duration}(s)) = \begin{cases} \#_{time}(e_m) - \#_{time}(e_i) \\ 0, \quad \text{otherwise} \end{cases}$$

In the following, we will note $e_{x1, ex2, ..., } e_{xi}$ as all the events that compose stage x.
Furthermore, the duration of the trace σ of length n is determined as follows:

$$duration(\sigma) = \#_{time}(e_n) - \#_{time}(e_1)$$

In our study framework, a trace always begins at the ED stage and ends at the last unit of the MCO stay (the unit before the MCO discharge).

Definition 4 (*Event log*)**.** An event log is a set of traces, representing the execution of the underlying process. An event can only occur in one trace, however events from different traces can share the same activity.

Definition 5 (*Patient athway*). A patient pathway describes the succession of medical events inside a healthcare facility. In this work, each pathway is linked to an MCO stay.

A patient pathway is a set *(p, s, σ, d)* where $p \in \mathbb{N}$ is the identifier of the patient, $s \in \mathbb{N}$ is the identifier of the MCO stay, *σ* is the trace of the MCO stay and *d* is the MCO discharge disposition.
We consider two types of pathways:

- *Scheduled pathways*: These pathways are planned before patient admission.
- *Unscheduled pathways*: Neither the admissions nor the pathways are planned. The patients are hospitalised from the emergency department or admitted to a specific unit for life-threatening emergencies.

Definition 6 (*Relevance of stage*)**.** A stage of a patient pathway is relevant if it is adequate that at this moment the patient is still hospitalised (first condition) and if the patient is in the medical unit intended to care for his pathology (second condition).

We consider three levels of relevance: (level 2) both conditions are met; (level 1) the second condition is not met, i.e., the patient is not hospitalised in the ideal medical unit for his pathology; and (level 0) no condition is met, and there is no more medical reason that justifies the patient being still hospitalised in this discipline.

Definition 7 (*Bias*)**.** In our context, a bias in the data is noted when some details about a piece of information are missing, which leads to a misinterpretation of a situation.

For example, a pathway {ED, Surgery, Geriatrics} without additional information, suggests that the patient needs surgery after the ED followed by geriatric care. The bias is that the patient just stays in surgery while waiting for a bed in geriatrics.

Definition 8 (*Activity labels*)**.** In this study, we consider two levels of activity names:

- Level 1: *U* is the set of labels corresponding to all the names or IDs of the medical units constituting the hospital group. Consequently, an event activity is a medical unit that a patient has visited.

$$GH = \{mu_1, ..., mu_n\}$$
$$U = \{id(mu_i)\} \text{ with } i \in [1, n]$$
$$\#_{activity}(e_x) = id(mu_m)$$

- Level 2: Let *L* be the set of labels corresponding to the relevance levels. *A* is the set of labels corresponding to the product of *U* and *L*:

$$L = \{level0, level1, level2\}$$
$$A = U \times L$$

$$\#_{activity}(e_x) = (id(mu_m), level\ l)$$

Hence, level 1 characterises the activity of an event based on the id of the medical unit and level 2 adds a level of relevance. For more convenient reading, in the following the activity of an event will be noted by the name of the medical unit.

Motivation

The pathway of a patient is governed not only by pure medical logic (healthcare needs), but also by logistical limitations. In other words, the pathway of a patient depends not only on his or her medical needs but also on the availability of inpatient beds and the possibility of discharge. Therefore, a patient can go to an unsuitable medical unit (unit b) because of a lack of beds in the suitable unit (unit a). The patient can later be transferred to unit a. Discharge also has an impact on patient pathways. Indeed, patients do not always immediately leave the hospital when they are medically fit for discharge because they are waiting for a discharge disposition. The challenge is to automatically identify these irrelevant steps in any MCO pathway. Indeed, these pathways are not clearly identified in the electronic health records (EHR) and there is no generally applicable thesaurus of ideal pathways and no clear indicator of the adequacy of a unit in the EHR. The same medical unit can have different functions, (see Table 1 for an example), and identical patients in terms of pathology can have different pathways according to hospital occupancy [17].

Our objective is to find a function (an algorithm) that evaluates the relevance of each stage. It is important to understand the word ***relevance*** as defined in the previous paragraph (Definition 6). Medical practices or medications are not judged here. Only the relevance of the patient's location was evaluated. In our framework, an input trace with events labels composed only of the activity name is converted into an output trace with events labels composed of the activity name and the level of relevance.

Table 1. Example of the different roles a medical unit can play in a patient pathway.

ED → Neurology	Acute care in neurology
ED → Neurology → Neurovascular Intensive Care	Waiting in neurology for a bed in neurovascular intensive care
ED → Neurovascular Intensive Care → Neurology	Waiting for a discharge solution in neurology

Definition of the Function Evaluating the Relevance of Stages

Let $\sigma = < e_{11}, e_{12}, e_{nx} > \forall e_{xi}, \#_{activity}(e_{xi}) \in U$ and $\sigma' = < e'_{11}, e'_{12}, ..., e'_{nx} > \forall e'_{xi}, \#_{activity}(e'_{xi}) \in A$.
σ and σ' are two traces of the same case, σ is the **history trace** and σ' is the **labelled trace**.

$$f: \sigma \rightarrow \sigma'$$

$$\text{with } |\{e'_i | \#_{act} e'_i = a\}|_{\forall e'_j \in \sigma'} \geq |\{e_i | \#_{act} e_i = a\}|_{\forall e_i \in \sigma}$$

The function f identifies the relevance levels of each stage in a trace. A stage can be divided into several phases with different levels of relevance. Therefore, the number of events that correspond to activity a in the trace σ is smaller or equal to the number of events that correspond to the activity a in trace σ'.

Example

This paragraph illustrates the definitions and the transformation of a pathway using the function *f*. In the following fictive example, the hospital group is named GHBS and is composed of two sites, named Scorff and Villeneuve. One patient arrived on the 4th of January at 5h36 at the Emergency Department of the Scorff Hospital. At 10h13 he was admitted to the observation unit (OU), but the patient was actually waiting for a bed in the geriatric unit. On the 5th of January at 9h45, the patient was transferred to the geriatric unit of the Villeneuve Hospital, another site of the hospital group. He arrived at 10h15. On the 10th of January at 14h00, the patient was medically fit for discharge. On the 12th of January at 13h30, the patient was discharged, and he returned home with additional community nursing services. Figure 2 illustrates the pathway of the patient according to the framework defined above.

The pathway of patient *00000056098* can be formalised as follows:

$\sigma =< e_{11}, e_{12}, e_{21}, e_{22}, e_{31}, e_{32} >$ with

$\#_{activity}(e_{11})$ = ED, $\#_{trans}(e_{11})$ = start

$\#_{activity}(e_{12})$ = ED, $\#_{trans}(e_{12})$ = end

$\#_{activity}(e_{21})$ = OU, $\#_{trans}(e_{21})$ = start

$\#_{activity}(e_{22})$ = OU, $\#_{trans}(e_{22})$ = end

$\#_{activity}(e_{31})$ = GERIATRICS, $\#_{trans}(e_{31})$ = start

$\#_{activity}(e_{32})$ = GERIATRICS, $\#_{trans}(e_{32})$ = end

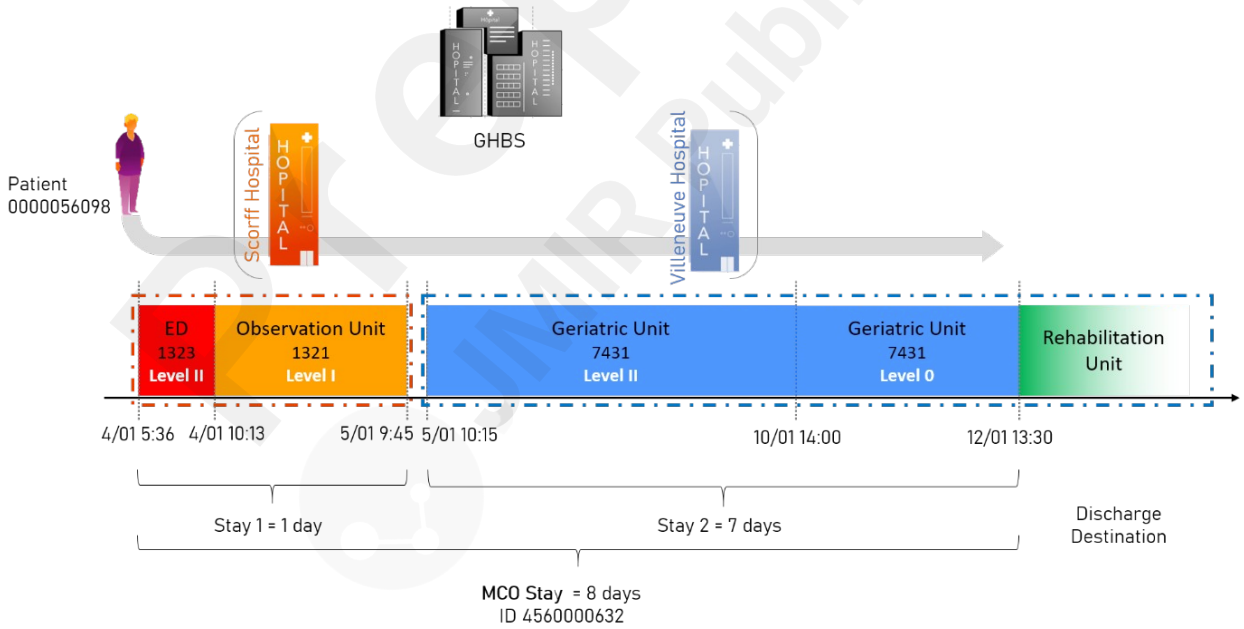


Figure 2. Illustration of the framework using a fictive pathway and patient.

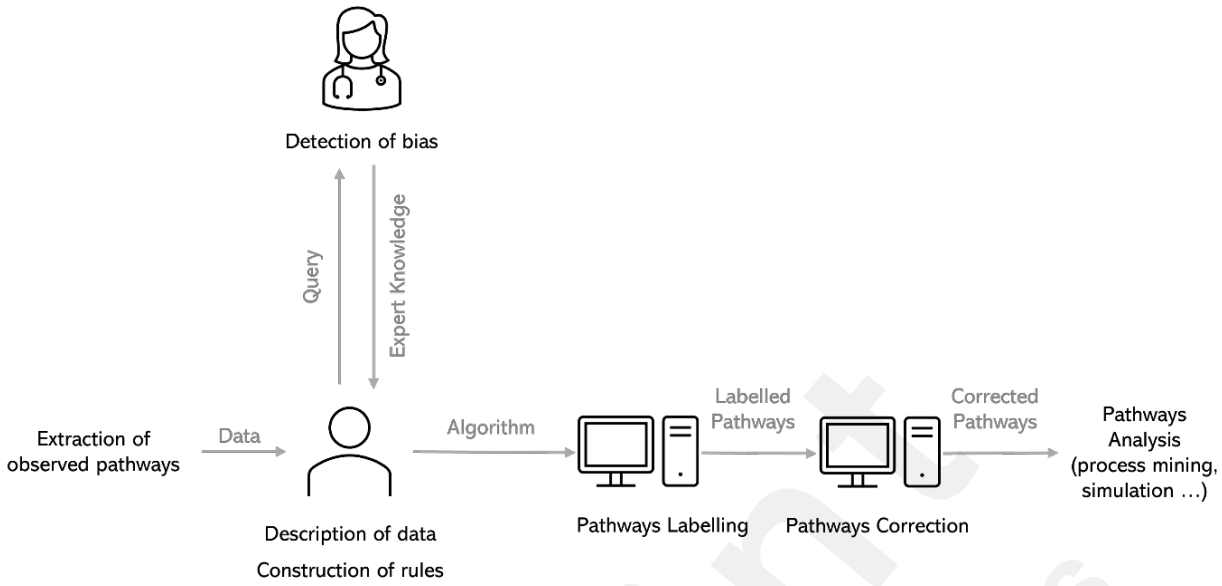
The function *f* takes the trace σ as input and returns the trace $\sigma' =< e'_{11}, e'_{12}, e'_{21}, e'_{22}, e'_{31}, e'_{32} >$ with the following features:

$\#_{activity}(e'_{11})$ = (ED, level2), $\#_{trans}(e'_{11})$ = start

$\#_{activity}(e'_{12})$ = (ED, level3), $\#_{trans}(e'_{12})$ = end

$\#_{activity}(e'_{21})$ = (OU, level1), $\#_{trans}(e'_{21})$ = start

$\#_{activity}(e'_{22})$ = (OU, level1), $\#_{trans}(e'_{22})$ = end



$\#_{activity}(e'_{31})$ = (GERIATRICS, level2), $\#_{trans}(e'_{31})$ = start
$\#_{activity}(e'_{32})$ = (GERIATRICS, level2), $\#_{trans}(e'_{32})$ = end
$\#_{activity}(e'_{41})$ = (GERIATRICS, level0), $\#_{trans}(e'_{41})$ = start
$\#_{activity}(e'_{42})$ = (GERIATRICS, level0), $\#_{trans}(e'_{42})$ = end

The succession of stages is described by level 2 activity labels. The initial stage *Geriatrics* has been divided into a relevant phase (level 2) and an irrelevant phase (level 0).

Automatic pathway labelling and correction

In this section, the method for identifying bias in patient pathway data and the method for building an algorithm to label the stages of the pathways as relevant or irrelevant are described. Then we present an algorithm for automatically transforming a history trace into a labelled trace and an algorithm for automatically transforming a labelled trace into a corrected trace. These algorithms are based on a symbolic approach. In other words, the rules are if/then propositions. Our approach can be visualised in Figure 3.

Identification of Bias and Rules Definition

The methodology for identifying the bias and then defining the rules consists of three steps, which are described below.

1. Description of the pathway dataset. The objective was to distinguish the different pathways and identify frequent and rare patterns.
2. Identification of bias with expert interpretation. The objective is to analyse the patterns with experts to determine the bias.
3. Definition of rules. The rules are defined based on the experts' analysis.

Figure 3. Pathway Labelling Method.

The first step can be achieved by computing the frequency and representativeness of each pathway variant, the number of events per variant, grouping some events, and computing the new variants to identify repetitive patterns. The second step enables us to identify bias through discussions with experts. Preferably, the discussions are conducted with several physicians to have different opinions. The third step requires the translation the analyses of physicians into rules. A rule deduces from the EHR data whether an event is relevant or not. According to our definition of relevance (cf. Definition

15) we consider three levels of relevance: (a) level zero, in which the stage is completely irrelevant (none of the conditions is met); (b) level one, the stage is not totally relevant (only the first condition are met); (c) level two, in which the stage is relevant (both conditions are met). The aim of the rules is to identify in a pathway the phases with different levels of relevance. We remind the reader that only the hospitalisation ward (patient location) is evaluated.

Example. A rule states that if the first stage of the trace lasts less than 10 units then the stage is completely irrelevant (level 0).
Let $\sigma =< e_{11}, e_{12}, e_{21}, e_{22} >$ be a trace with two activities a_1 and a_2 such that the following conditions are met:

- $\#_{activity}(e_{11}) = \#_{activity}(e_{12}) = a_1$,
- $\#_{activity}(e_{21}) = \#_{activity}(e_{22}) = a_2$,
- $\#_{time}(e_{11}) = 0$,
- $\#_{time}(e_{12}) = 7$,
- $\#_{time}(e_{21}) = 7$, $\#_{time}(e_{22}) = 20$.

The first stage is a_1 and lasts 7 units therefore e_{11} and e_{12} are labelled level 0.

Algorithm 1: Pathway Labelling

Once the rules have been defined, Algorithm 1 labels a pathway according to these rules (lines 6-9). A level of one or zero is assigned if a stage is identified as irrelevant, and a level of two is assigned, if the stage is relevant (lines 7-9). A stage can be labelled by several rules. In this case, the worst label is applied, i.e. level 0 has a priority over level 1 and level 1 over level 2 (lines 10-12).
We want to emphasise that this algorithm is purely based on logic and administrative rules. It does not include medical reasoning and is therefore inaccurate. However, the aim of the next paragraph is to evaluate this inaccuracy, i.e. the number of errors between an algorithm with simple rules and the complex reasoning of an expert (expert knowledge).

Algorithm 2: Pathway Correction

We also implemented an algorithm to correct the pathways once labelled using Algorithm 1. The idea is to transform the observed pathway into a theoretical pathway by correcting irrelevant stages. The different corrections applied to a labelled pathway are deduced from the rules. The irrelevant activities are replaced by the relevant activities (lines 4-6). Only the label of an event is changed, and the timestamp remains the same. At the end of the correction, subsequent identical activities are merged (lines 7-13). For example, let us note a stage (activity name, relevance level, start, or end). The pathway *<(ED, level2, t1, t2), (OU, level1, t2, t3), (Geriatrics, level2, t3, t4)>* is corrected and becomes *<(ED, t1, t2), (Geriatrics, t2, t3), (Geriatrics, t3, t4)>.* In addition, the pathway can be merged to become *<(ED, t1, t2), (Geriatrics, t2, t4)>*.

Algorithm 1 Pathway Labelling

1: Let e_x be an event of the history pathway
2: Let e'_x be an event of the labelled pathway
3: Let t_1 be the start date of the stay.
4: Let t_n be the end date of the stay.
5: Let $\sigma =< e_{11}, e_{12}, ..., e_{n1}, e_{n2} >$ be the trace representing the history pathway.
6: Let L be a list that stores the result of each rule.
7: **for** each rule r_k **do**
8: Add $r_k(\sigma)$ to L
9: **end for**

10: **for** each $e_x \in \sigma$ **do**
11: Apply the modification of each rule that has changed e_x. The lowest relevance level has priority.
12: **end for**
13: Return $\sigma' = < e'_{11}, e'_{12}, ..., e'_{m1}, e'_{m2} >$, the labelled trace.

Algorithm 2 Pathway Correction
1: Let $\sigma' = < e'_{11}, e'_{12}, ..., e'_{m1}, e'_{m2} >$, be the labelled trace.
2: Let r be the rule applied at e'_x.
3: Let $\sigma'' = < e''_{11}, e''_{12}, ..., e''_{m1}, e''_{m2} >$, be a copy of σ'.
4: **for** each $e''_x \in \sigma''$ **do**
5: $\#_{activity}(e''_x)$ = the corrected activity according to the rules of correction
6: **end for**

7: **for** each $e''_x \in \sigma''$ with $t_x < t_m$ **do**
8: **if** $\#_{activity}(e''_{x1}) = \#_{activity}(e''_{x+1,1})$ **then**
9: $\#_{time}(e''_{x2}) = \#_{time}(e''_{x+1,2})$ with $\#_{trans}(e''_{x+1,2})$ = complete and all the events of stage *x+1* are deleted from σ''
10: **end if**
11: **end for**
12: Return $\sigma'' = < e''_{11}, e''_{12}, ..., e''_{p1}, e''_{p2} >$, the corrected trace.

Evaluation of the Performance of the Labelling Algorithm

This subsection describes the method used to assess the labelling algorithm. Because there is no reference to compare the results of the algorithm with a ground truth, the evaluation of the algorithm has to be made by comparing its results with the analyses of experts. The methodology used for this study is inspired by the framework developed by the French think-tank Ethik IA for its humane warranty college [15]. A representative sample of patient pathways was analysed by two experts. They had access to information from the electronic patient records. Each expert performed the analysis separately. The results of the first expert are compared with those of the second expert. When the results did not match, medical experts discussed them to find a common answer. Then their answers are compared with the algorithms' answers. For each difference, a discussion with the experts allows us to determine whether the algorithm is wrong and if so, to qualify the errors.
The method presented here is general and can be applied at any hospital. In the next section, we apply these methods to a real case study to create rules to label and correct a real dataset extracted from a hospital database, and we evaluate the accuracy of these rules.

Results

Data

This work was performed with the Groupe Hospitalier Bretagne Sud (GHBS), a French hospital group located in the Lorient area. It has two general hospital sites with an emergency department and six other sites. In total, there were 89,791 emergency department visits, and 108,875 hospitalisations and sessions (values for 2021). The study was based on data collected at the GHBS. Data were retrospectively collected for the time period from July 2020 to July 2021. The data cover 12 months of activity, 62,839 different MCO stays (including simple ED visits), 18,796 different MCO stays, and 47,888 unique patients. The multiple MCO-stays of the same patient were treated as separate instances. Three sources of data were used: (a) electronic patient records, (b) administrative

healthcare databases, and (c) data from the software used for rehabilitation and home hospitalisation. Only structured data were used to save time in the data analyses; no plug-and-play natural language processing tool was available for our data. The paediatric and obstetric pathways were excluded from the study dataset, as were the pathways with only a visit to the emergency department. The study was approved by the French Data Protection Authority (CNIL: Commission Nationale de l'Informatique et des Libertés) under the number 922243.

Identification of Bias and Rules Definition

In this section, we detail the results obtained with our method to identify bias from our data.

Results of the Step 1: Description of the Pathway Dataset

In our dataset, there are 19,905 pathways and 1013 trace variants. Some variants are very frequent such as (ED, OU) representing 23% of the pathways (4528 out of 19,905) and others are very rare such as (ED, neurology intensive care, cardiology) occurring just once.
We observed that most pathways had only one stage after the ED visit (6% of the variants – 60 out of 1013 – and 79% of the pathways – 15,699 out of 19,905). Pathways with more than three stages after the ED are rare. They appear between one and three times in the dataset and represent 1% of the pathways (195 out of 19,905) but 19% of the variants (191 out of 1013). Therefore, the diversity of pathways is mainly due to pathways with many activities. We identified five types of pathways at the GHBS: a) mono-disease pathways, which include one necessary medical unit; b) seriously ill patient pathways, which include transfer to an intensive care unit, c) older person pathways, which include geriatric units; d) frequent and processed pathways, which include strokes, and e) multi-diseases and complex pathways, which include several medical units.
The most frequent medical units are observation units, polyvalent medicine units, geriatric medicine units, post-emergency units, surgery units, and specialised medical units. By categorising the medical units into four groups (ED, medicine, surgery, and intensive care unit (ICU)), we obtained 165 patterns, and the 10 most frequent structures of the pathways are listed in Table 3.

Within the pathways (MEDICINE, MEDICINE), several patterns were frequently observed. Pathways such as heart failure or stroke pathways are normally composed of two steps after the ED: admission to a cardiology (respectively neurology) intensive care unit followed by cardiology (respectively neurology). The second type of pattern is the pathway with an admission to a polyvalent unit before an admission to a specialised unit or another polyvalent unit. Here, a polyvalent unit, is a medical unit such as an observational unit, polyvalent medicine, or post-emergency unit where patients with multiple diseases or who do not require specialised treatment are treated. We also observe a few transfers between specialised units. Finally, patients can be transferred between the weekly hospitalisation unit and the full hospitalisation unit.
For most of the pathways that follow the pattern (SURGERY, MEDICINE), the activity of surgery is noted as an overflow bed. For the few others, a surgical act is performed before a transfer for medical reasons to a medicine unit.
The pattern (MEDICINE, SURGERY) mainly concerns an admission to an observation unit (while waiting for surgery) followed by a surgery unit. In the other pathways, patients are first admitted to a specialised unit (e.g. hepatogastroenterology) before surgery.

Hence, we obtained four patterns for the pathways in two steps:

- Pattern 1: A polyvalent unit is followed by a specialised unit
- Pattern 2: A surgery unit is followed by a medical unit
- Pattern 3: A daily or weekly hospitalisation unit followed by a full hospitalisation unit of the same speciality

- Pattern 4: A specialised unit followed by another specialised unit

Results of Step 2: Interpretation of patterns by experts

We discussed the patterns identified in the first step with the experts. According to experts, pattern 1 (a polyvalent unit followed by another unit) usually indicates that the polyvalent unit is used as a buffer. In fact, when beds are lacking, patients can be admitted to a polyvalent unit to begin their treatment while waiting for a bed in the ideal unit. The second pattern has the same explanation, a patient requiring medical treatment is placed in a surgery unit while waiting for a bed in the ideal unit. Pattern 3 (transfers between daily or weekly hospitalisation units and full hospitalisation units) is explained by a lack of beds in the full hospitalisation unit. The last pattern (transfers between specialised units) has no general explanation. Occasionally, a specialised unit can also be used as a buffer, but the transfer can also be medically explained.
The length of stay was also an aspect of the pathway discussed with the experts. Some pathways are too long because patients cannot be discharged as soon as they are medically fit for MCO-discharge. The delay is mainly due to an insecure back home or a lack of beds in rehabilitation centres. Length of stay can be compared with a national reference. In France, the national reference is the average length of stay used for hospital stays invoicing (DMS: Durée Moyenne de Séjour in French). It is defined for each diagnosis related group. However, each pathway is unique and a length of stay above the national reference does not guarantee that the stay was too long because of discharge difficulties. For example, longer stays for patients receiving palliative care are frequent and normal. Occasionally, the diagnosis related group does not correctly report the seriousness of the patient because the patient was transferred to another hospital.
The length of stay in the emergency department was also discussed. According to experts, the length of stay in the ED is occasionally too long because patients are waiting to be hospitalised. The ED LoS should not exceed 5 to 10 hours.

Results of Step 3: Rules Construction

From these observations and discussions, we deduced several dimensions to investigate in a pathway:

- Time spent in the ED: ED LoS can be prolonged because of a lack of inpatient beds in acute care units. Rule 1 evaluates whether the time spent in the ED is too long.
- Length of stay: The length of stay can be prolonged because of a delayed discharge. Rules two and three evaluate this condition.
- Overflow bed: Occasionally, patients are admitted to a medical unit but are treated by the physicians of another unit. The location of the patient is entered (i.e. the activity name) in the hospital data and the unit medically responsible for the patient is also indicated. For example, when a patient is in surgery and awaits a bed in cardiology, the activity is surgery and the medically responsible unit is cardiology.
- Sequence of activities: A typical pattern is the transfer between a polyvalent unit and a specialised unit. According to physician experience, when the transfer occurs within one week, in general, the polyvalent unit stage was irrelevant. In rule five, the threshold is 7 and a half days to consider the transfer time. Rule six is dedicated to the observation unit because in our dataset a transfer in the observation unit is labelled as “back home within 24 h”, “observation” or “awaiting bed”. The label “awaiting bed” indicates an irrelevant stage because the patient should have been transferred immediately to the appropriate unit.

The challenge was to identify which structured data could be used to investigate these different

dimensions and therefore, to create the rules. We detail the rules obtained from our dataset in Multimedia Appendix 1 as well as the algorithms of the seven rules.
From these rules, we deduced how to correct the labelled pathways. The different corrections are listed in Table S1 in Multimedia Appendix 1.

Example

Figure 4 illustrates the labelling and correction of a pathway. Two phases are considered irrelevant by the algorithm: the observation unit (rule six) and the end of the stay in the geriatric unit (rule three). To correct the pathway, the time spent in the observation unit is replaced by the time spent in the geriatric unit which is the relevant stage following the observation unit and the patient is discharged earlier, at the presumed discharge date.

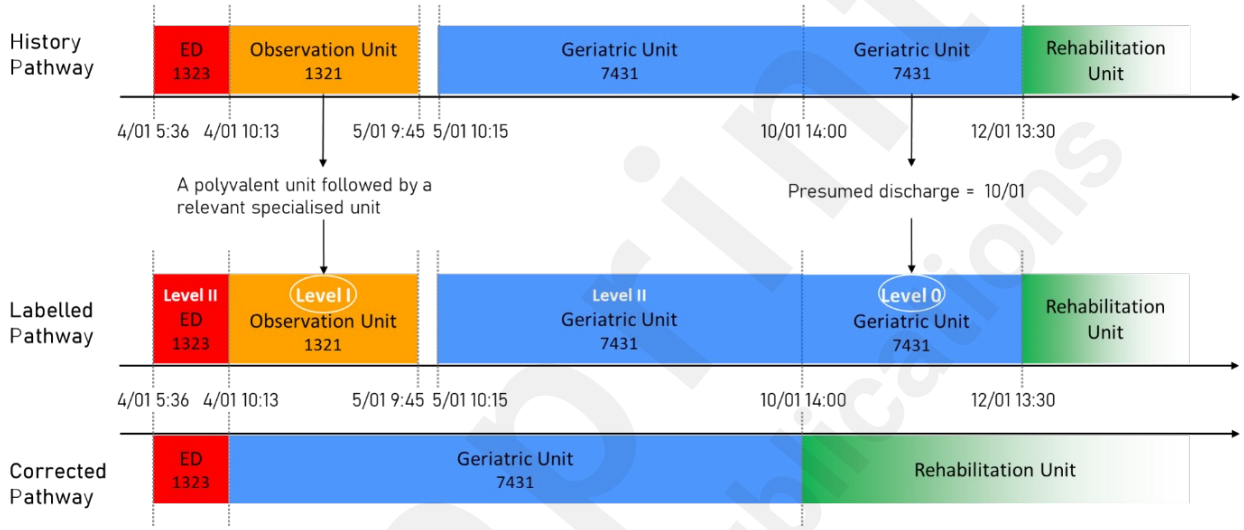


Figure 4: Example of the correction of a pathway.

Accuracy of the Labelling Algorithm

One hundred and eighteen different pathways were analysed by six different duos of physicians from the GHBS (only one physician of the twelve had previously participated in the discussion to define the rules) and compared with the algorithm output. Only rules 3, 4, 5, 6 and 7 could be evaluated because the physicians could not obtain a consensus on either the ideal length of stay or the ideal length of ED visit within the allotted time. The physicians had access to the patient's pathway, his age, the chief complaint, the diagnoses, the national length of stay reference for the diagnosis related groups, the discharge destination, and requests for rehabilitation and home hospitalisation. They also had access to patient records if more information was needed.
We counted the number of errors per stage (except for the first ED visit) and per pathway. For example, the pathway (ED, OU, CARDIOLOGY) has two stages (we do not count ED), and experts assess that OU is not relevant and that CARDIOLOGY is relevant, but the algorithm assesses that the two stages are relevant. Therefore, we identified 1 pathway error / 1 pathway and 1 stage error / 2 stages. Table 2 presents the results of the evaluation on the 118 pathways. The algorithm failed to evaluate 19% (23 out of 118) of the pathways and 13% of the stages (30 out of 232). Among these errors, false-positive errors are five (resp. three) times more important than false-negative errors. This means that the main error of the algorithm considers a path or a stage as relevant, whereas it is actually irrelevant.
The main errors are related to medical knowledge. When patients are hospitalised in a specialised medical unit (e.g. oncology) but pertain to another specialised medical unit (e.g. cardiology) the algorithm does not detect this irrelevance. Similarly, hospitalisation in the general intensive care unit

before transfer to the cardiology intensive care unit is occasionally irrelevant but not detected by the algorithm. In addition, the algorithm considers that a polyvalent unit following a specialised unit is irrelevant; however in some cases it is wrong.

Table 2. Performance of the algorithm based on rules 3, 4, 5, 6 and 7

			Actual values		Precision	Recall
			Positive	Negative		
Pathways	Predicted values	Positive	56	19	0.75	0.93
		Negative	4	39		
Stages	Predicted values	Positive	162	22	0.88	0.95
		Negative	8	40		

Positive = Relevant pathway or stage (level 2)
Negative = Irrelevant pathway or stage (level 0 and level 1)

Preprocessing of Pathway Data

In the following, we used only rules 3 to 7 to preprocess our data given that the rules 1 and 2 could not be assessed. We evaluated 19,832 pathways, including 24,989 stages (excluding the first ED stage). Among these pathways, 2669 pathways (14%) were evaluated as irrelevant, and 2802 (11%) stages were also evaluated as irrelevant. Considering the error margin, between 2162 and 3176 pathways were irrelevant and between 2438 and 3166 stages were irrelevant. The main irrelevant movements detected by the algorithm are overflow bed in surgery, overflow bed in polyvalent units, and waiting time in the observation unit. The 19,832 history pathways included 986 variants. Once corrected, 792 variants were noted. Table 3 details the distribution of the structures of the variants. We observed that the “ED, MEDICINE, MEDICINE” and “ED, SURGERY, MEDICINE” traces are less frequent than before correction, which is due to the correction of the overflow beds. The trace ED appeared because the OU stage was assessed as irrelevant for several “ED, OU” (included in “ED, MEDICINE”) traces.

Statistical Analysis of Relevant and Irrelevant Pathways

Once the pathways were labelled, we compared the relevant pathways with the irrelevant pathways to understand the causes of irrelevance in the pathways. Several causes are already known among medical and administrative staff: bed occupation rates, ED crowding, age, and discharge destination. We tested these hypotheses with four bivariate analyses. The four variables to explain were an ED LoS greater than 5 hours, an ED LoS greater than 10 hours, the presence of overflow beds and delayed discharge. For categorical variables, the proportions were compared using a chi-square test. For quantitative variables, the distributions were compared with a Student’s t-test. We studied different explanatory variables: weekday corresponds to the start of the ED visit or admission to the inpatient unit; the arrival period is divided into four periods [morning 7 h-12 h, afternoon 12 h-17 h, night 17 h-23 h, and deep night 23 h-7 h]; the next history stage is the inpatient unit where the patient was admitted and the next corrected stage is where the patient should have been admitted (based on the evaluation of the pathways); the last stage is the medical unit from which the patient was discharged; and the ED crowds are the number of patients present in the ED when the patient arrives. Table 4 reports the bivariate analyses. See Table S2 and Table S3 in Multimedia Appendix 2 for the detailed results. Several observations can be made from the statistical analysis. (1) The seasons and the irrelevance of the next stage are not significant for a delay in the ED greater than 5 hours, but they are significant for a delay greater than 10 hours. (2) Counterintuitively, fewer irrelevant post-ED admissions occur for long ED delays. This finding is probably because patients who stay in the ED for a long time are ultimately admitted to a relevant unit. (3) The weekday of patient arrival

influences the ED delay. On Mondays more patients wait more than 5 hours, and the proportion decreases throughout the week and increases again on Sundays. This phenomenon is caused not only by the greater number of patients admitted on Monday but also by the difficulty in hospitalising patients during the weekend. Therefore on Sunday many patients are waiting for hospitalisation. (4) This is the same observation and explanation for the overflow beds; more patients are admitted to an irrelevant unit on Sunday. (5) Crowding in the emergency department is less important for the longest ED delays. Indeed patients arriving at night or late at night are less likely to be transferred to an inpatient unit and this is also the period which fewer patients arrive at the ED. (6) A greater number of patients are admitted to irrelevant units when the ED is more crowded. (7) Increased age is a factor of long delays in accessing the ED. (8) Similar features are noted for the occupation rate of the next stage. (9) The next corrected stages had an occupation rate (95%) higher than that of the next history stages (92%). (10) Age does not impact the risk of being in an overflow bed. (11) The season has an impact on discharge delays. Specifically, in summer more discharges are delayed, perhaps because the healthcare supply is lower during the summer holidays. (12) The proportion of delayed discharges varies according to the destination of the discharge. Discharges at a psychiatric centre have the highest rate of delay (47%: 92 delayed stays out of 203 discharges in psychiatry) followed by discharges at rehabilitation centres (39%: 1285 delayed stays out of 3294 discharges in rehab). Delayed discharges for death correspond to requests for palliative care at home or at another centre that were not accepted in time. (13) Age does not affect the risk of delayed discharge.

Table 3: Main structures of history and corrected pathways.

Variant	Percentage of occurrence	
	History	Corrected
ED, MEDICINE	68.68	68.57
ED, MEDICINE, MEDICINE	10.67	7.24
ED, SURGERY	9.66	10.90
ED, SURGERY, MEDICINE	2.88	0.33
ED, MEDICINE, SURGERY	1.7	0.63
ED, MEDICINE, MEDICINE, MEDICINE	1.2	0.80
ED, ICU, MEDICINE	0.88	1.05
ED, ICU	0.53	0.56
ED, SURGERY, SURGERY	0.47	0.16
ED, ED, MED	0.39	
ED		7.18

Table 4: Bivariate Analysis

Variable to explain	Features	*P* value
ED visit > 5 h	Age	<.001
	ED crowds	<.001
	Occupation Rate History Next Stage	<.001
	Occupation Rate Corrected Next Stage	<.001

	Weekday	<.001
	Season	0.513
	Arrival Period	<.001
	Next Stage is Irrelevant	0.053
	Next History Stage	<.001
	Next Corrected Stage	<.001
ED visit > 10 h	Age	<.001
	ED crowds	<.001
	Occupation Rate History Next Stage	<.001
	Occupation Rate Corrected Next Stage	<.001
	Weekday	<.001
	Season	<.001
	Arrival Period	<.001
	Next Stage Irrelevant	<.001
	Next History Stage	<.001
	Next Corrected Stage	<.001
Overflow Beds	Age [b]	0.033
	ED crowds [c]	<.001
	Occupation Rate Corrected Unit [d]	<.001
	Arrival Hour [d]	0.365
	Weekday	<.001
	Season	0.782
	Arrival Period	<.001
Delayed Discharge	Age	0.722
	Discharge Destination	<.001
	Last Stage	<.001
	Season	<.001

a. The analysis was performed exclusively with the data from the principal site (Scorff) because ED crowds and age differ between the principal site and the smaller site (Villeneuve).
b. We compared the set of pathways without the overflow stage and the set of pathways with at least one overflow stage.
c. In each pathway, the ED crowds were only computed for the first medical unit subsequent to the ED stage.
d. We compared the sets of relevant stages and irrelevant stages.

The period of study was impacted by the COVID-19 pandemic. These results could be more robust with access to a longer period of study (three years instead of one year) and the seasons variably could be assessed several times during a longer timeframe. Furthermore, the quality of the data was imperfect, especially for the computation of the occupation rate. Therefore, the results should not be extrapolated to other periods or hospitals. However, the analysis allows us to compare the relevant and irrelevant pathways because they are derived from the same dataset. Hence, we can conclude that significant differences are observed between relevant and irrelevant pathways. Logistic factors such as the day of the week, the hour of arrival, medical unit occupation and the discharge

destination influence the risk of overflow.

Synthesis of Patient Pathway Labels

To summarise this section, from the analysis of the structure of the patient pathways and expert knowledge, we built seven rules that detect irrelevant stages in a patient pathway (description of the dataset and construction of rules). Based on these rules, we used a pathway labelling algorithm that labels the stages of a pathway according to three levels of relevance (pathway labelling). Then, we used a pathway correction algorithm that transforms a labelled pathway into an ideal pathway (pathway correction). The evaluation of our algorithm shows that it exhibits 87% accuracy. In our dataset, 14% (2669 out of 19,832) of the pathways were labelled irrelevant. Finally, a statistical comparison between relevant and irrelevant pathways demonstrated that logistic constraints influence the quality of patient pathways.
The next two sections show the importance of this preprocessing step before analysing patient pathways with process mining and using these data for hospital management.

Case Study 1: Analysis of Patient Pathways Using Process Mining

Motivation

The first case study investigates the impact of our preprocessing technique on process discovery. We evaluated the ability of our preprocessing method to simplify process graphs. We compared the process graph of an event log composed of history traces with the process graph of an event log composed of corrected traces. We used the ProM framework 6.12 [30] to discover the graphs and estimate different metrics. The process graphs were computed using the Fodina algorithm [40], which outputs a causal graph that was transformed into a Petri-net with the plugin “Convert Causal net (C-Net) to Petri net (F. Mannhardt)”.

Metrics

The metrics were computed using the plugin “Show Petri-net Metrics” (H.M.W. Verbek). This plugin computes (1) the extended Cardoso metric (ECaM), (2) the extended Cyclomatic metric (ECyM), (3) the structuredness [22] and (4) the density [25].
The ECaM counts the splits (XOR, OR, AND) in the net and penalises each of them. The ECyM is the difference between the number of edges and vertices plus the number of strongly connected components. According to Lassen et al. [22], a high ECaM score can be caused by a “high degree of fan-out from places” and numerous parallelisms can increase the ECyM. Structuredness recognises different types of structures and scores each structure by giving it a penalty value. Finally, the density relates the number of arcs to the number of all possible arcs for a given number of nodes. Therefore, these four metrics quantify the different structural characteristics of a graph.

Quantitative Analysis

The process graph discovered from the whole dataset of pathways is spaghetti like because the number of variants is large with or without correction (986 vs. 792). To avoid spaghetti-like effects, we reduced the analysis to one medical unit, i.e. the event log included only the traces that contained this activity. Table 5 shows the metrics calculated for five medical units. As expected, the corrected event log contains fewer variants and activities than the historical event log. Consequently, the number of arcs, places and transitions on the graph also decreases. The history graph density is greater than the corrected graph density. This finding indicates that the corrected graph is more

compact than the history graph. The ECaM and ECyM of the corrected graph are lower than those of the history graph. Indeed, we can observe more places with many output transitions and more parallelisms in the history Petri-net. Finally, the structuredness of the history graph is also greater than the structuredness of the corrected graph, except for the neurology unit. This means that more complex structures and/or more unstructured components are observed in the history graphs than in the corrected graph. The neurology exception can be explained by the fact that a state-machine is identified in the history graph, but only an unstructured component is identified in the corrected graph. In conclusion, the corrected Petri-nets can be considered simpler than the history Petri-nets.

Qualitative Analysis

A qualitative analysis can also be performed (see Multimedia Appendix 3 for the pictures of the graphs). The cardiology history graph (Figure S1 in Multimedia Appendix 3) shows that several activities can occur before admission to cardiology, but it is not easy to distinguish between groups of patients. The corrected graph is easier to read and four groups of patients can be identified: (1) serious patients who need intensive care or continuous care before being admitted to cardiology, (2) patients who need to be permanently monitored for cardiac examination, (3) patients who need pulmonology care before cardiology care, and (4) patients who do not need other care before transfer and are directly admitted to cardiology (with eventually a step in the observation unit before).
The history graph of the polyvalent medicine unit (Figure S2 in Multimedia Appendix 3) is complex to read, and several specialised units are present but not related to the polyvalent unit in the graph. The corrected graph is much simpler to read. Three groups of stays are identified: (1) stays with intensive care or continuous care before admission to polyvalent medicine, (2) stays with direct admission, and (3) stays with a step in the observation unit or seasonal unit before admission to polyvalent medicine.
Equivalent analyses can be performed on other medical units. The neurology graph (Figure S3 in Multimedia Appendix 3) is less simple than the other graphs, possibly because patients going to neurology have complex pathways, or because the correction of neurology pathways requires particular rules. However, the corrected graph is again more interpretable than the history one.

Table 5. Comparison of the history and corrected process graph

Process Graph		**No. Variants**	**No. Activities**	**No. Arcs, Places, Transitions**	**Density**	**ECaM**	**ECyM**	**Structure-dness**
Cardiology	History	13	10	32, 7, 16	0.14	11	16	64
	Corrected	9	7	22, 6, 11	0.17	8	11	22
Visceral Surgery	History	19	15	42, 8, 21	0.13	12	20	42
	Corrected	8	6	20, 6, 10	0.17	8	10	20
Polyvalent Medicine	History	12	13	76, 16, 38	0.06	33	27	76
	Corrected	7	6	28, 9, 14	0.11	13	11	56
Geriatric Medicine	History	5	6	14, 4, 7	0.25	3	7	9.5
	Corrected	3	4	14, 4, 5	0.25	3	5	6.5
Neurology	History	24	13	52, 11, 26	0.09	21	26	208
	Corrected	16	8	37, 10, 18	0.10	17	24	1602

Case Study 2: Estimation of Ward Capacity by Simulation

Motivation

Computer simulations can be used to estimate the number of beds necessary in each medical unit to admit unscheduled patients and help solve capacity planning problems. Indeed, the actual number of patients admitted to each medical unit does not include all the patients not admitted because of a lack of beds and the number of the patients who should not be admitted to each unit is considered. Hence, it does not reflect the real need for beds. To solve this problem, pathway correction be used. This method can be useful for estimating the capacities when building or renovating a hospital or for organising medical teams.

DES Model

In this case study, we simulate patient flow through the medical units of one hospital to compute the level of occupation of the medical units. We compared a simulation with the history pathways (scenario 1) and a simulation with the corrected pathways (scenario 2). To do so, we modelled the medical units and patient flow using AnyLogic software. Figure 5 shows the DES model.
The model represents one general hospital and 20 full hospitalisation units. Each patient has a succession of units to follow. The model simulates the admissions of patients to medical units, their stays, and their discharges. A patient exits the simulation when (s)he has completed his pathway. To simulate the source of patients, we used the dataset described in Section 5.1 (only patients from the main general hospital site (named Scorff) were included). We filtered the dataset based on the following criterion: only the pathways with at most 3 activities (ED visits plus one or two units) were included. We excluded pathways with weekly or daily units, the rare variants (coverage percentage below 0.001) and patients from sites other than Scorff. The corrected pathway dataset was filtered to keep only the stays included.

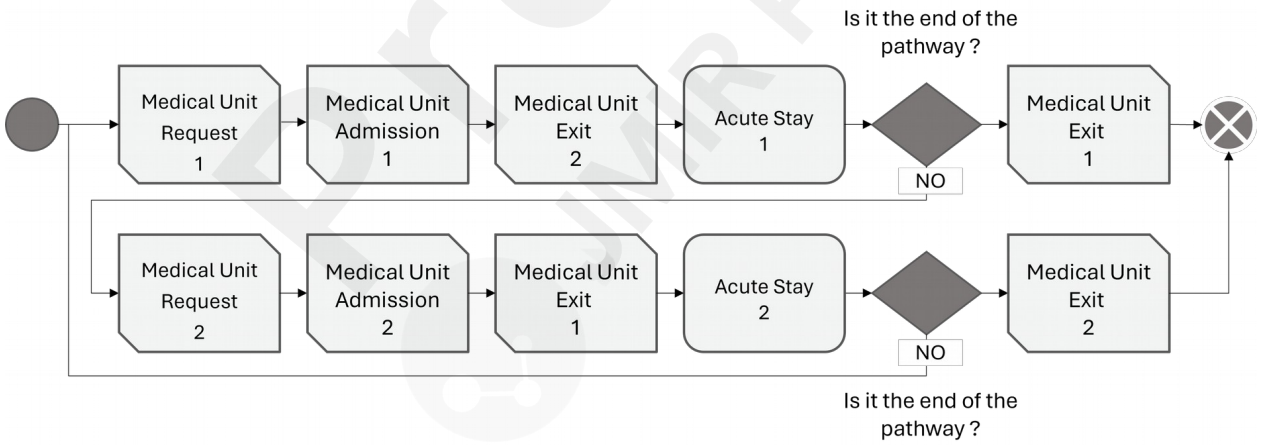


Figure 5. Modelling of Patient Flows through Medical Units.

Experimental Settings

The capacities of the medical units were set to infinity to calculate how many unscheduled patients needed to be admitted to each unit each day. The length of stay in each unit was randomly generated according to a probability law. To choose this probability law we compared several distributions (normal, beta, gamma, log-normal, Weibull and exponentiated Weibull) and fitted them on the stays of our dataset (between 150 and 5000 stays per medical unit). The log-normal distribution best fits

the lengths of stay in each unit. The parameters of the log-normal distributions were adjusted for each unit by fitting the distribution to the real values. The log-normal distribution tends to generate more extreme values than those observed in reality; thus the length of stay was limited to 28 days in a unit and 24 hours in the ED. The simulated patients' arrival dates are fixed and equal to the real patients' arrival dates. The simulated patients either followed the history pathways (scenario 1) or the corrected pathways (scenario 2). The simulation duration is one year and the simulation run included 13,366 pathways.

To choose the warm-up time, we monitored the mean number of present patients in each medical unit over the simulation time. After 50 days, a stable situation is reached (see Figure S6 in Multimedia Appendix 4). The warm-up time was set to two months. The number of replications was chosen to have an error less than 10%. Fifteen replications allow this target to be reached and a reasonable simulation time to be reached: approximately 1 minute is required for one run and the fifteen replications take 10 minutes. The results are the mean values of fifteen replications.
The simulation model was validated by comparing the mean length of stay of each medical unit from the simulation results to the real dataset. This is the unique source of randomness in the model since the simulated patients arrive according to the real dataset and follow a deterministic pathway. The mean absolute error is 0,6 days, which is less than 10% of the mean length of a hospital stay (see Table S4 in Multimedia Appendix 4).

Table 6. Mean number (95% confidence interval) of acute patients in medical units over one year.

Medical Unit	Scenario 1	Scenario 2	Difference
3O Surgery [a]	7.6 ± 0.27	1.1 ± 0.14	-6.5
Orthopaedic Surgery	10.1 ± 0.19	13.3 ± 0.24	3.1
Visceral Surgery	6.5 ± 0.28	5.9 ± 0.19	-0.7
Pulmonology	14.2 ± 0.30	14.3 ± 0.39	0.1
Cardiology ICU	2.0 ± 0.00	2.0 ± 0.00	0.0
Cardiology	10.1 ± 0.14	10.1 ± 0.34	0.1
Post emergency	17.0 ± 0.35	14.5 ± 0.34	-2.5
Polyvalent Medicine	35.5 ± 0.57	33.9 ± 0.45	-1.7
Neurology ICU	3.0 ± 0.00	3.0 ± 0	0.0
Neurovascular	3.9 ± 0.24	5.1 ± 0.19	1.2
Neurology	3.7 ± 0.32	3.1 ± 0.14	-0.7
Hepatogastroenterology	13.3 ± 0.32	13.7 ± 0.44	0.4
Rheumatology	11.4 ± 0.34	11.3 ± 0.26	-0.1
Observation Unit	8.7 ± 0.24	6.5 ± 0.28	-2.2
Geriatric Medicine	39.3 ± 0.55	38.7 ± 0.38	-0.5
ICU [b]	1.8 ± 0.22	2.0 ± 0.20	0.2
Seasonal Unit [c]	2.9 ± 0.19	2.0 ± 0	-0.9

Oncology Haematology	5.0 ± 0.20	5.2 ± 0.22	0.2
Nephrology Endocrinology	3.5 ± 0.28	3.4 ± 0.27	-0.1
CCU [d]	0.6 ± 0.27	0.6 ± 0.27	0.0
Total	200	190	-10

[a] ENT, ophthalmologic, orthopaedic surgery
[b] Intensive Care Unit
[c] The seasonal unit is only open during the winter months. Therefore the occupation figures computed over a year do not reflect reality.
[d] Continuing Care Unit

Results

Table 6 shows the mean number of acute patients present in each unit in both situations. For several units the number of patients is differed between scenario 1 and scenario 2. For example, the polyvalent medicine unit has on average two patients less with corrected pathways and the neurovascular unit has one more patient. We also observed that, globally fewer patients present at the same time with the corrected pathways compared with the historic pathways. Indeed, there are fewer stages in the corrected pathways because some are judged irrelevant; therefore, in the simulation the patients stay less in the hospital.
In conclusion, preprocessing pathway data is important for addressing capacity planning problems in hospitals. In this example, we observed that using historic pathways can lead to biased numeric interpretations for the capacity planning of medical units.

Discussion

Principal Findings

A framework and a methodology to study patient pathways were presented in this paper. They were used to develop a pathway labelling algorithm that automatically detects whether a patient pathway is irrelevant i.e. contains stages due to resource limitations (as defined in Definition 6). Two main methods are available to achieve such a task: (1) building a thesaurus with medical experts (or using supervised learning) that links the main diagnosis (or the chief complaint) with an ideal pathway, and (2) building a symbolic algorithm. The first method is the most accurate but is very time-consuming. This method would require hours of work with experts to build a thesaurus or to annotate data for training a machine. None of these methods provide general results because the thesaurus and rules need to be adapted to each hospital. We chose the second option, an algorithm based on logic and administrative data, because it can be built quickly and is easily adaptable to organisational changes. We provided a general method to build this algorithm. We applied our algorithm to our dataset, and we were able to estimate the gap between our algorithm and an expert assessment. Our results demonstrate that a nonnegligible gap exists (13% to 19% of errors); however, we believe that the rate error was small enough for globally evaluated pathways. The estimation of this error also enabled us to identify the source of errors of the algorithm. Based on this labelling, a correction of the pathways is then performed to represent pathways that would be considered “ideal”.

We also demonstrated that resource limitations impact the choice of pathway using a statistical analysis that compares relevant and irrelevant pathways. The factors identified as increasing difficulties in managing patient flows could be included in hospitals’ strategies to improve patient pathways.

The two case studies illustrate the importance of preprocessing patient pathway data before any analysis. Studying and representing patient pathways using process mining is complicated (cf. Section 2). By focusing on pathways with a common medical unit, we demonstrated that a corrected graph is more interpretable than a history graph. Hence, our algorithm is an efficient preprocessing tool for the analysis of patient pathways using process mining. The simulation of patient pathways is useful for testing bed management changes, but numeric results can be false if the input data include bias. In our example, the determination of the mean number of beds required for acute patients differed for the historic pathways and corrected pathways. Some medial units need fewer beds and others need more beds.

Pathway labelling should be applied before any analysis such as process mining (case study 1), simulation (case study 2) or training of machine learning models to predict hospital pathways. In another work, we studied the prediction of the medical unit where a patient will be admitted after an emergency department visit. If raw data are used to train a machine learning model, the training will be biased. Indeed, the model will learn, for example, that some patients who do not need surgical treatment should be transferred to surgery. In contrast, if the model is trained from relevant pathways, it will learn the ideal medical unit for the patients [35].

Limitations

We proposed a general method to study patient pathways and identified bias in the data. However, our approach could only be tested one dataset because of legal constraints. Therefore, additional studies with other hospitals should be performed to validate the generalisability of our approach. Our labelling algorithm is not 100% accurate. To avoid errors, more rules could be created by exploiting textual data using natural language processing. Indeed, to build our algorithm, we only used structured data because of the unavailability of an adequate tool to treat textual data in our hospital.

Conclusion

This work suggests a new approach to preprocess data on pathways of unscheduled patients. To our knowledge, there are no other studies that have evaluated nonspecific disease pathways. Our approach has the advantages of being explicable, simple to implement, and adaptable to each hospital.
Future research could develop process discovery techniques that consider the relevance labels of the activities.

Acknowledgement

The authors thank the Groupe Hospitalier Bretagne Sud (GHBS), especially the GHBS Information System Department teams and the SIB Company Data Department teams, for providing access to electronic health records for this study. The authors would like to thank the physicians who helped us construct and validate our algorithms, including Dr. Bry, Dr. Chevallier, Dr. Dollon, Dr. La Combe, Dr. Le Corf, Dr. Girard, Dr. Henry, Dr. Laurichesse, Dr. Le Merlay, Dr. Lenoir, Dr. Luquet, Dr. Maigre, Dr. Vaillant. The authors also thank Dr. Quiguer for her advice and Pr. Saulnier for his help with the statistical analyses.

Conflicts of Interest

This work was funded by the company Enovacom. The authors have no other competing interests to declare that are relevant to the content of this article.

Appendix 1: Algorithms of the Rules

Rule 1: If the ED stage is longer than the reference duration (e.g., 5 hours), then the stage is split into two stages: a stage (ED, level 2) from admission to + 5 h and a stage (ED, level 1) from + 5 h to discharge from the ED.

The only necessary data were the date of ED admission and the date of ED discharge.

Algorithm 3 Rule 1: Evaluation of the ED length of stay.

Let $e_{11} = (ED, start, t_1)$ and $e_{12} = (ED, end, t_2)$ be the two events of stage s_1.
if $\#_{duration}(s1) > 5h40$ **then**
$e'_{11} = ((ED, level2), start, t_1)$
$e'_{12} = ((ED, level2), end, t_1 + 5h)$
$e'_{21} = ((ED, level1), start, t_1 + 5h)$
$e'_{22} = ((ED, level2), end, t_2)$
end if

Rule 2: If the stay is longer than the time reference, then the stage that begins before the time reference and ends after the time reference is split into two stages: one stage (activity, level 2) from the start of the initial stage to the time reference and one stage (activity, level 0) from the time reference to the end of the initial stage. All the following stages are labelled with a level 0. Exceptions: pathways with a transfer to another hospital, stays with palliative care, death of the patient during the stay, and requests for home hospitalisation or rehabilitation. The following data were used: date of stay, diagnoses related group, time reference for the diagnoses related group, discharge destination to determine whether the patient died or was transferred, stay coding to determine whether palliative care occurred.

Algorithm 4 Rule 2: Evaluation the length of stay.

Let *reference* be the national reference of length of stay that corresponds to the diagnoses related group.
if the patient is transferred to another hospital for a serious reason or presumed discharge date is filled or received palliative care or died during the stay **then**
the rule 2 is not applied.
end if

Let t_1 be the start date of the stay.
Let t_n be the end date of the stay.
Let $\sigma =< e_{11}, e_{12}, ..., e_{n1}, e_{n2} >$ be the trace representing the history pathway.
if $\#_{duration}(\sigma) > 1.5 * reference$ **then**
$t_{end} = t_1 + reference$
if t_{end} is a Saturday or a Sunday and $t_n >$ next Monday 20h **then**
t_{end} = next Monday at noon
end if
for e_{x1} such as $e_{x1} = (mu, start, tx_1)$ with $t_{x1} < t_{end}$ and $e_{x2} = (mu, end, t_{x2})$ with $t_{x2} > t_{end}$ **do**
$e'_{x1} = ((mu, level2), start, t_{x1})$
$e'_{x2} = ((mu, level2), end, t_{end})$
$e'_{y1} = ((mu, level0), start, t_{end})$
$e'_{y2} = ((mu, level0), end, t_{x2})$
end for

for each e_{x1} such as $e_{x1} = (mu, start, tx_1)$ with $t_{x1} > t_{end}$ **do**
$e'_{x1} = ((mu, level0), start, t_{x1})$
$e'_{x2} = ((mu, level0), end, t_{x2})$
end for
end if

The last exception is addressed by another rule. Indeed, when a request for home hospitalisation or rehabilitation is made, a presumed discharge date is entered.

Rule 3: If the discharge date is later than the presumed discharge date, then the stage that begins before the presumed discharge date and ends after discharge is split into two stages: one stage (activity, level 2) from the start of the initial stage to the presumed discharge date and one stage (activity, level 0) from the presumed discharge date to the end of the initial stage. All the following stages are labelled with a level 0.

The data needed are the requests for home hospitalisation and rehabilitation. Notably, if a presumed discharge date was entered for every stay, rule 2 would not have been necessary.

Algorithm 5 Rule 3: Detection of delayed discharges caused by rehabilitation or home hospitalisation requests.

Let t_{presum} be the presumed discharge date.
if $t_n > t_{presum} + 1day$ **then**
for e_{x1} such as $e_{x1} = (mu, start, t_{x1})$ with $t_{x1} < t_{presum}$ *and* $e_{x2} = (mu, end, t_{x2})$ with $t_{x2} > t_{presum}$ **do**
$e'_{x1} = ((mu, level2), start, t_{x1})$ $e'_{x2} = ((mu, level2), end, t_{presum})$ $e'_{y1} = ((mu, level0), start, t_{presum})$
$e'_{y2} = ((mu, level0), end, t_{x2})$
end for
for each e_{x1} such as $e_{x1} = (mu, start, t_{x1})$ with $t_{x1} > t_{presum}$ **do**
$e'_{x1} = ((mu, level0), start, t_{x1})$
$e'_{x2} = ((mu, level0), end, t_{x2})$
end for
end if

Rule 4: If the activity is different from that of medically responsible unit, then the stage is labelled with a level 1.

Algorithm 6 Rule 4: Detection of overflow bed.

Let mu be a medical unit (activity) of the pathway and mr the unit medically responsible.
for each e_{x1} and $e_{x2} \in \sigma$ such as $e_{x1} = (mu, start, tx_1)$ $e_{x2} = (mu, end, t_{x2})$ **do**
if $mu \neq mr$ **then**
$e'_{x1} = ((mu, level1), start, t_{x1})$
$e'_{x2} = ((mu, level1), end, t_{x2})$
end if
end for

Rule 5: If a polyvalent unit is followed by a specialised unit (except the intensive care unit and continuous care unit) within 7 and a half days, then the stage of the polyvalent unit is labelled level 1.

Algorithm 7 Rule 5: Detection of polyvalent unit used as a buffer.

for each e_{x1} and $e_{x2} \in \sigma$ such as e_{x1} = (*mu, start,* t_{x1}) and e_{x2} = (*mu, end,* t_{x2}); the two events of stage s_x **do**
  if *mu* is a polyvalent unit and is not the last medical unit of the pathway and is not followed by an intensive care unit and $\#_{duration}(s_x) > 7\frac{1}{2}$ days **then**
    e'_{x1} = ((*mu, level*1), *start,* t_{x1})
    e'_{x2} = ((*mu, level*1), *end,* t_{x2})
  end if
end for

Rule 6: If the mutation is labelled as "awaiting bed" then the observation unit stage is labelled irrelevant (level 1). Daily and weekly hospitalisations only concern scheduled admissions. When an unscheduled patient begins his pathway in a daily or weekly hospitalisation unit, generally it means that no bed was available in the full hospitalisation unit.

Algorithm 8 Rule 6: Detection of observation unit used as a buffer.

for each e_{x1} and $e_{x2} \in \sigma$ such as e_{x1} = (*mu, start,* t_{x1}) and e_{x2} = (*mu, end,* t_{x2}) with *mu* = $\#_{activity}(e_{x1}) = \#_{activity}(e_{x2})$ = observation unit **do**
  if the mutation is labelled "awaiting bed" **then**
    e'_{x1} = ((*mu, level*1), *start,* t_{x1})
    e'_{x2} = ((*mu, level*1), *end,* t_{x2})
  end if
end for

Rule 7: If a weekly or daily hospitalisation unit is followed by the full hospitalisation unit, then the stage is labelled level 1.

Algorithm 9 Rule 7: Detection of daily and weekly hospitalisation units used as a buffer.

for each e_{x1} and $e_{x2} \in \sigma$ such as e_{x1} = (*mu, start,* t_{x1}) e_{x2} = (*mu, end,* t_{x2}) **do**
  if *mu* is a weekly hospitalisation unit and is not the last medical unit and the next activity is the associated full hospitalisation unit **then**
    e'_{x1} = ((*mu, level*1), *start,* t_{x1})
    e'_{x2} = ((*mu, level*1), *end,* t_{x2})
  end if
end for

Algorithm 10 Pathway correction specific for our dataset.

Let σ' =< e'_{11}, e'_{12}, ..., e'_{m1}, e'_{m2} > be the labelled trace that starts at t_0 and ends at t_m.
Let *r* be the rule applied at e'_x.
Let σ'' =< e''_{11}, e''_{12}, ..., e''_{m1}, e''_{m2} > be a copy of σ'.
for each $e''_x \in \sigma$ **do**
  if $\#_{activity}(e''_x)$ = (mu, level1) and *r* =rule 1 or rule 5 or rule 6 **then**
    $\#_{activity}(e''_x)$ = the next relevant activity
  else if $\#_{activity}(e''_x)$ = (mu, level1) and *r* = rule 4 **then**
    $\#_{activity}(e''_x)$ = the activity medically responsible
  else if $\#_{activity}(e''_x)$ = (mu, level1) and *r* = rule 7 **then**
    $\#_{activity}(e''_x)$ = the full hospitalisation unit of the same speciality
  else if $\#_{activity}(e''_x)$ = (mu, level0) **then** e''_x is deleted from σ''
  else $\#_{activity}(e''_x)$ = the medical unit of $\#_{activity}(e'_x)$

end if
end for

for each $e''_x \in \sigma''$ with x < m **do**
if $\#_{activity}(e''_{x1}) = \#_{activity}(e''_{x+1,1})$ **then**
$\#_{time}(e''_{x2}) = \#_{time}(e''_{x+1,2})$ with $\#_{trans}(e''_{x+1,2})$ = complete and all the events of stage *x+1* are deleted from σ''
end if
end for
Return $\sigma'' = < e''_{11}, e''_{12}, ..., e''_{p1}, e''_{p2} >$, the corrected trace.

Table S1: Correction of the pathways

Rule	Correction
Rule 1	The irrelevant ED stage is replaced by the next relevant stage.
Rule 2 and 3	The irrelevant stages are deleted.
Rule 4	The real medical unit is replaced by the unit medically responsible.
Rule 5	The activity name polyvalent unit, is replaced by the activity name of the next relevant stage.
Rule 6	The activity name observation unit, is replaced by the activity name of the next relevant stage.
Rule 7	The activity name of the weekly or daily hospitalisation is replaced by the full hospitalisation unit of the same speciality.

Appendix 2: Statistical Analysis Tables

Table S2: Bivariate Analysis for Categorical Features

		Proportion (%)		*P* value
	Features	**ED visit ≤ 5h**	**ED visit > 5h**	
ED visits	**Weekday**			< .001
	Monday	16	20	
	Tuesday	14	16	
	Wednesday	15	14	
	Thursday	15	13	
	Friday	16	14	
	Saturday	13	11	
	Sunday	11	12	
	Season			0.513
	Spring	26	25	
	Summer	25	25	
	Autumn	24	24	
	Winter	25	26	
	Arrival Period			< .001
	Morning	19	16	
	Afternoon	33	21	
	Night	34	33	
	Deep Night	14	29	
	Next Stage is Irrelevant			0.053
	True	14	13	
	False	86	87	
ED visits	**History Next Stage**			< .001
	3O Surgery	6	5	
	Cardiology	2	3	
	Cardiology ICU	2	1	
	CCU	1	1	
	Geriatric Medicine	9	13	
	ICU	1	1	
	Nephrology Endocrinology	1	1	
	Neurology	1	2	
	Neurology ICU	2	3	
	Observation Unit	22	13	
	Observation Unit Villeneuve	9	5	
	Oncology Haematology Hepatogastro-enterology	4	6	
	Orthopaedic Surgery	5	3	
	Polyvalent Medicine	14	17	
	Post-emergency Villeneuve	4	3	

	Post-emergency	8	11	
	Pulmonology	3	4	
	Rheumatology	2	3	
	Visceral Surgery	4	5	
ED visits	**Corrected Next Stage**			< .001
	3O Surgery	1	1	
	Cardiology	3	4	
	Cardiology ICU	2	1	
	CCU	1	1	
	Geriatric Medicine	11	15	
	ICU	1	1	
	Nephrology Endocrinology	1	1	
	Neurology	2	3	
	Neurology ICU	2	3	
	Observation Unit	20	12	
	Observation Unit Villeneuve	8	4	
	Oncology Haematology Hepatogastro-enterology	5	8	
	Orthopaedic Surgery	6	4	
	Polyvalent Medicine	15	17	
	Post-emergency Villeneuve	3	3	
	Post-emergency	7	9	
	Pulmonology	4	5	
	Rheumatology	3	4	
	Visceral Surgery	5	5	
		ED visit ≤ 10h	**ED visit > 10h**	
ED visits	**Weekday**			< .001
	Monday	16	25	
	Tuesday	14	19	
	Wednesday	15	13	
	Thursday	15	10	
	Friday	16	9	
	Saturday	13	8	
	Sunday	11	16	
	Season			< .001
	Spring	26	21	
	Summer	25	27	
	Autumn	24	24	
	Winter	25	28	
	Arrival Period			< .001
	Morning	20	1	
	Afternoon	32	7	
	Night	34	39	
	Deep Night	14	53	
	Next Stage Irrelevant			< .001
	True	14	10	

	False	86	90	
ED visits	**History Next Stage**			< .001
	3O Surgery	6	2	
	Cardiology	2	4	
	Cardiology ICU	2	0	
	CCU	1	1	
	Geriatric Medicine	9	20	
	ICU	1	1	
	Nephrology Endocrinology	1	1	
	Neurology	2	2	
	Neurology ICU	2	3	
	Observation Unit	21	9	
	Observation Unit Villeneuve	9	1	
	Oncology Haematology Hepatogastro-enterology	4	7	
	Orthopaedic Surgery	5	2	
	Polyvalent Medicine	14	21	
	Post-emergency Villeneuve	4	1	
	Post-emergency	8	14	
	Pulmonology	3	4	
	Rheumatology	2	3	
	Visceral Surgery	4	3	
ED visits	**Corrected Next Stage**			< .001
	3O Surgery	1	0	
	Cardiology	3	4	
	Cardiology ICU	2	0	
	CCU	1	1	
	Geriatric Medicine	11	20	
	ICU	1	1	
	Nephrology Endocrinology	1	2	
	Neurology	2	3	
	Neurology ICU	2	3	
	Observation Unit	20	8	
	Observation Unit Villeneuve	8	1	
	Oncology Haematology Hepatogastro-enterology	5	9	
	Orthopaedic Surgery	6	1	
	Polyvalent Medicine	15	20	
	Post-emergency Villeneuve	3	1	
	Post-emergency	7	12	
	Pulmonology	4	5	
	Rheumatology	3	3	
	Visceral Surgery	5	3	
		Relevant stage	**Overflow stage**	
Overflow Beds	**Weekday**			< .001
	Monday	15	15	
	Tuesday	16	16	

	Wednesday	15	15	
	Thursday	15	13	
	Friday	16	13	
	Saturday	13	14	
	Sunday	10	13	
	Season			0.782
	Spring	27	27	
	Summer	25	25	
	Autumn	23	22	
	Winter	25	26	
	Arrival Period			< .001
	Morning	10	6	
	Afternoon	37	27	
	Night	31	36	
	Deep Night	22	30	
		Prompt Discharge	**Delayed Discharge**	
Delayed Discharge	**Discharge Destination**			< .001
	Home	70	63	
	Nursing home or home hospitalisation	4	5	
	Rehabilitation centre	15	27	
	Psychiatry	1	2	
	Transfer	2	1	
	Death	7	2	
	Last Stage			< .001
	Season			< .001
	Spring	18	16	
	Summer	16	57	
	Autumn	33	13	
	Winter	34	15	

Table S3: Bivariate Analysis for Numerical Features

	Features	**Mean (Median)**	**(Q1, Q3)**	**Mean (Median)**	**(Q1, Q3)**	***P* Value**
		ED visit ≤ 5h		**ED visit >5h**		

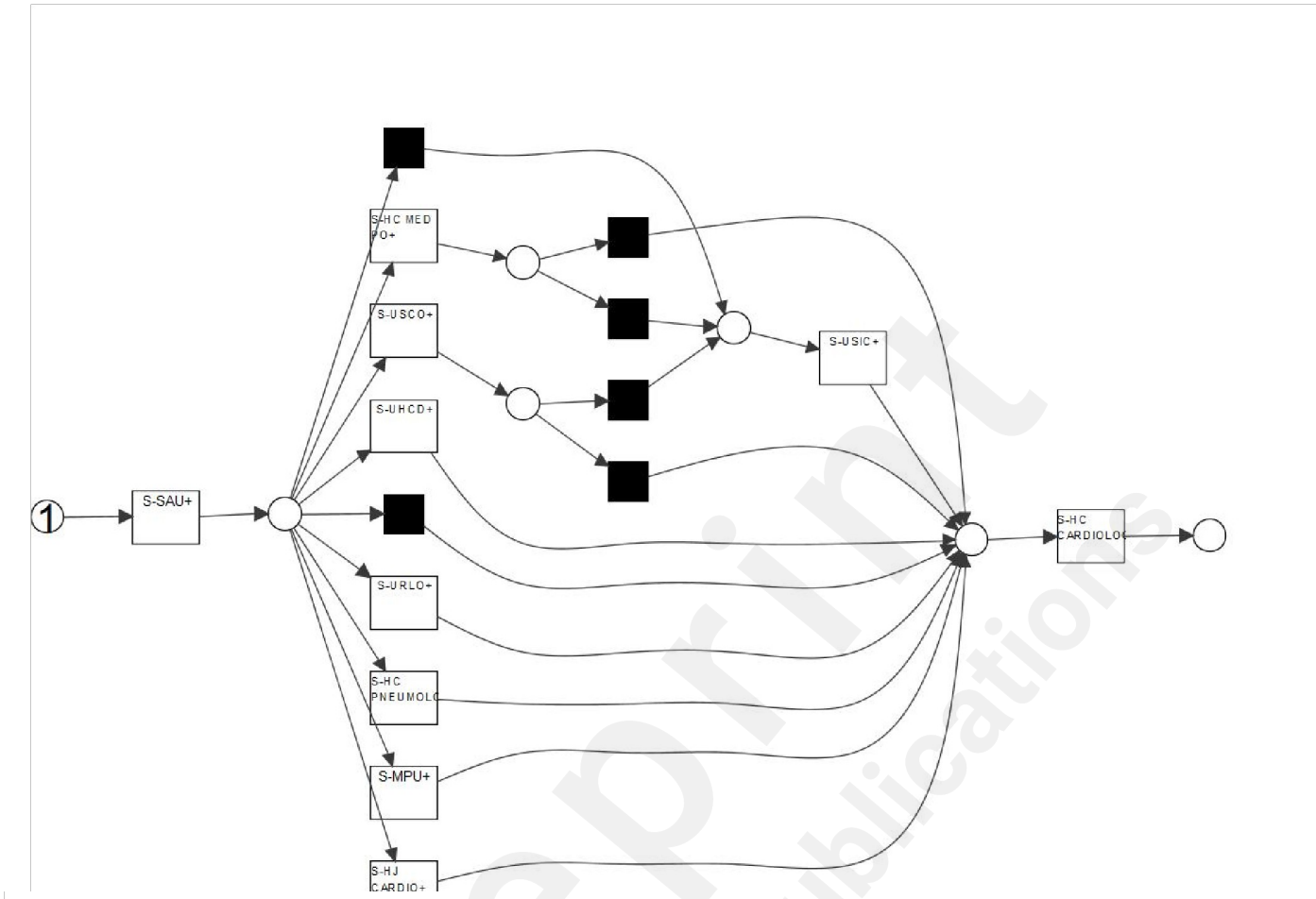
		n = 16784 (83 %)		n = 3512 (17 %)		
ED Visits [a]	Age	68.5 (73)	(56, 85)	70.7 (75)	(60, 85)	<.001
	ED crowds	25.2 (25)	(18, 32)	23.9 (24)	(16, 31)	<.001
	Occupation Rate History Next Stage	85.7 (90.6)	(73.55, 100)	91.0 (96)	(83, 104)	<.001
	Occupation Rate Corrected Next Stage	87.3 (92.8)	(78.4, 100)	92.1 (97.6)	(88, 104)	<.001
		ED visit ≤ 10h		**ED visit >10h**		
		n = 18888 (93 %)		n = 1408 (7 %)		
ED Visits [a]	Age	68.6 (73)	(56, 85)	72.6 (76)	(64, 86)	<.001
	Occupation Rate History Next Stage	86.0 (91.2)	(74, 100)	94.5 (100)	(91.2, 106)	<.001
	Occupation Rate Corrected Next Stage	87.6 (93)	(79.2, 100)	94.8 (100)	(92.8, 106)	<.001
Overflow Beds		**Overflow Phase**		**Relevant Phase**		
		n = 2800 (11.28 %)		n = 22014 (88.72 %)		
	Arrival Hour [b]	13.9 (16)	(10, 19)	13.8 (15)	(11, 18)	0.365
	Occupation Rate Corrected Unit [b]	87.0 (95.2)	(85.6, 100)	85.4 (90)	(75.2, 98.6)	<.001
	ED crowds [c]	n = 1971 (13.18 %)		n = 12979 (86.82 %)		
		25.5 (25)	(19, 32)	24.4 (25)	(17, 31)	<.001
	Age [d]	n = 538 (3.65 %)		n = 14205 (96.35 %)		
		65.3 (68)	(52, 83)	67.2 (71)	(54, 84)	0.033
		Delayed Discharge		**Prompt Discharge**		
		n = 3550 (25.32 %)		n = 10471 (74.68 %)		
Delayed Discharge	Age	70.5 (74)	(60, 85)	70.4 (74)	(60, 85)	0.722

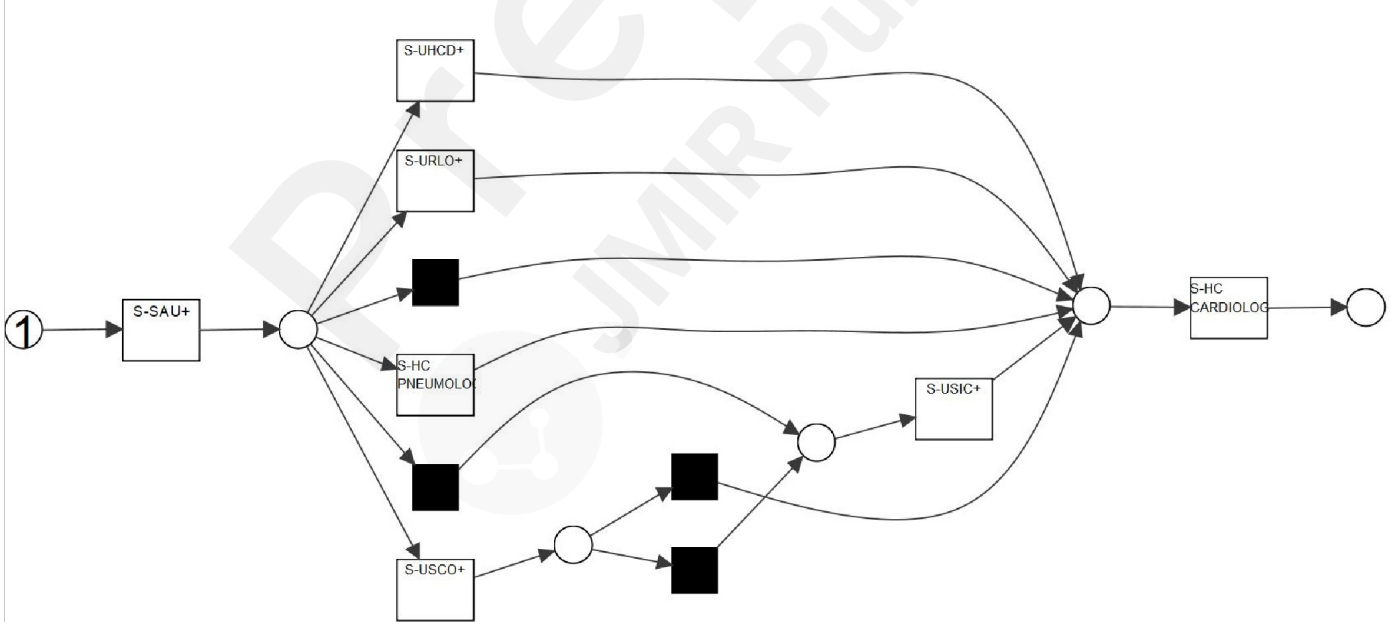
a. The analysis was done only on the data of the principal site (Scorff) because ED crowds and age are very different between the principal site and the smaller one (Villeneuve).
b. We compared the set of relevant stages and irrelevant stages.
c. In each pathway, the ED crowds was only computed for the first medical unit subsequent to the ED stage.
d. We compared the set of pathways without overflow stage and the set of pathways with at least one overflow stage.

Appendix 3: Process Models

Figures legend:
S-HC 3 O(ORL-OPH-ORT): ENT, Ophthalmology, Orthopaedic Surgery
S-HC CARDIOLOGIE: Cardiology
S-HC CHIR ORTHO: Orthopaedic Surgery
S-HC CHIR VISCERALE: Visceral surgery
S-HC HGE: Hepato-Gastro-Enterology
S-HC MED PO: Polyvalent Medicine
S-HC MGG: Geriatric Medicine
S-HC NEPHRO ENDOC: Nephrology Endocrinology
S-HC NEUROLOGIE: Neurology
S-HC ONCO-HEMATO: Oncology Haematology
S-HC PNEUMOLOGIE: Pulmonology
S-HC RHUMATO-MI: Rheumatology Infectious Diseases
S-HC UNV: Neuro-vascular
S-HJ CARDIO: Cardiology Daily Hospitalisation
S-MPU: Post-Emergency Care
S-SAU: Emergency Department
S-UHCD: Observation Unit
S UNITE SAISON URG: Seasonal Unit
S-URLO: ICU
S-USCO: CCU
S-USIC: Cardiology ICU
S-USINV: Neurology ICU

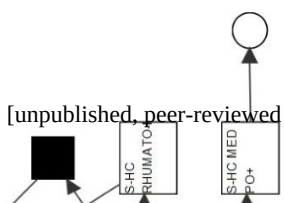
(a) History

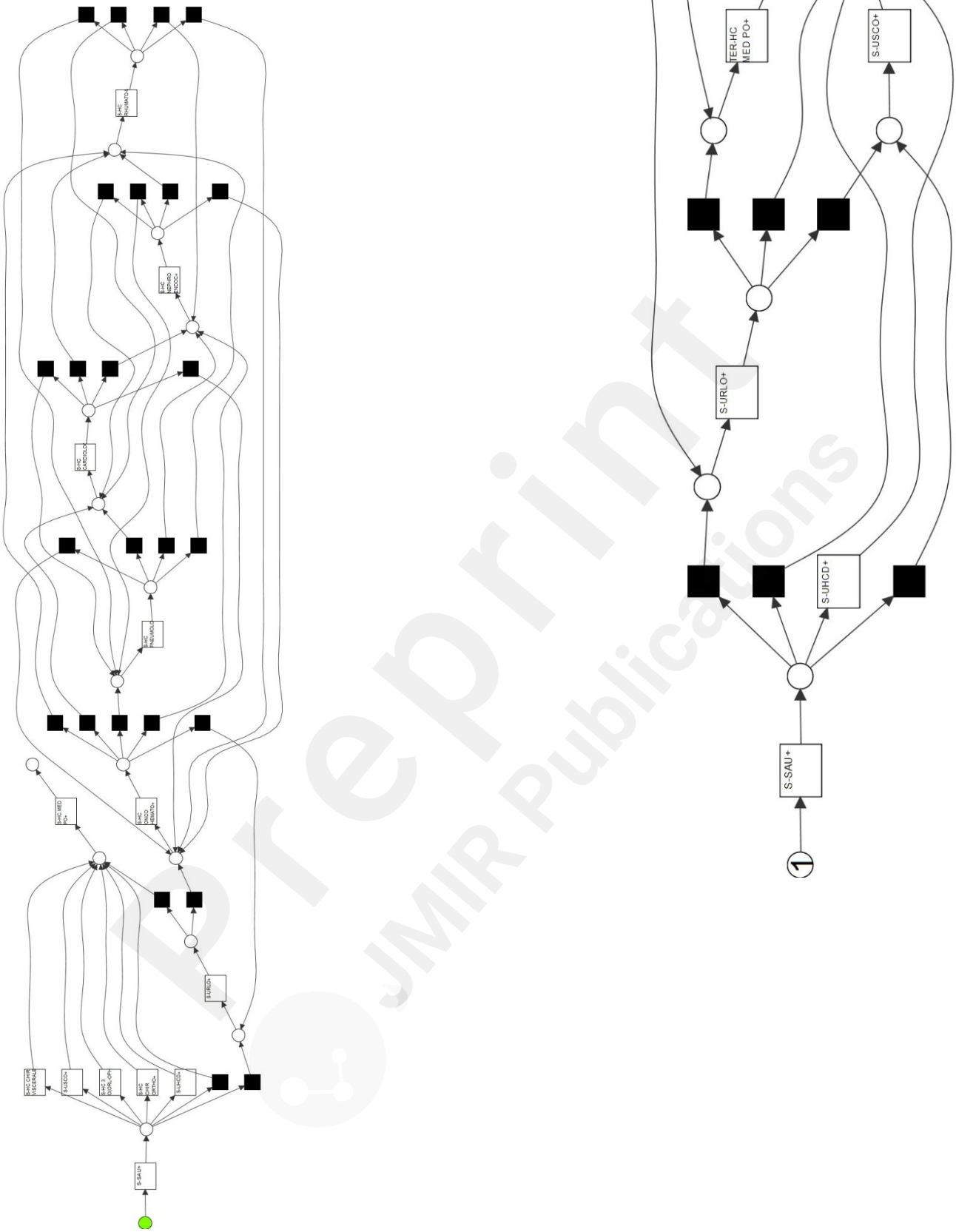




(b) Corrected

Figure S1: Cardiology graphs

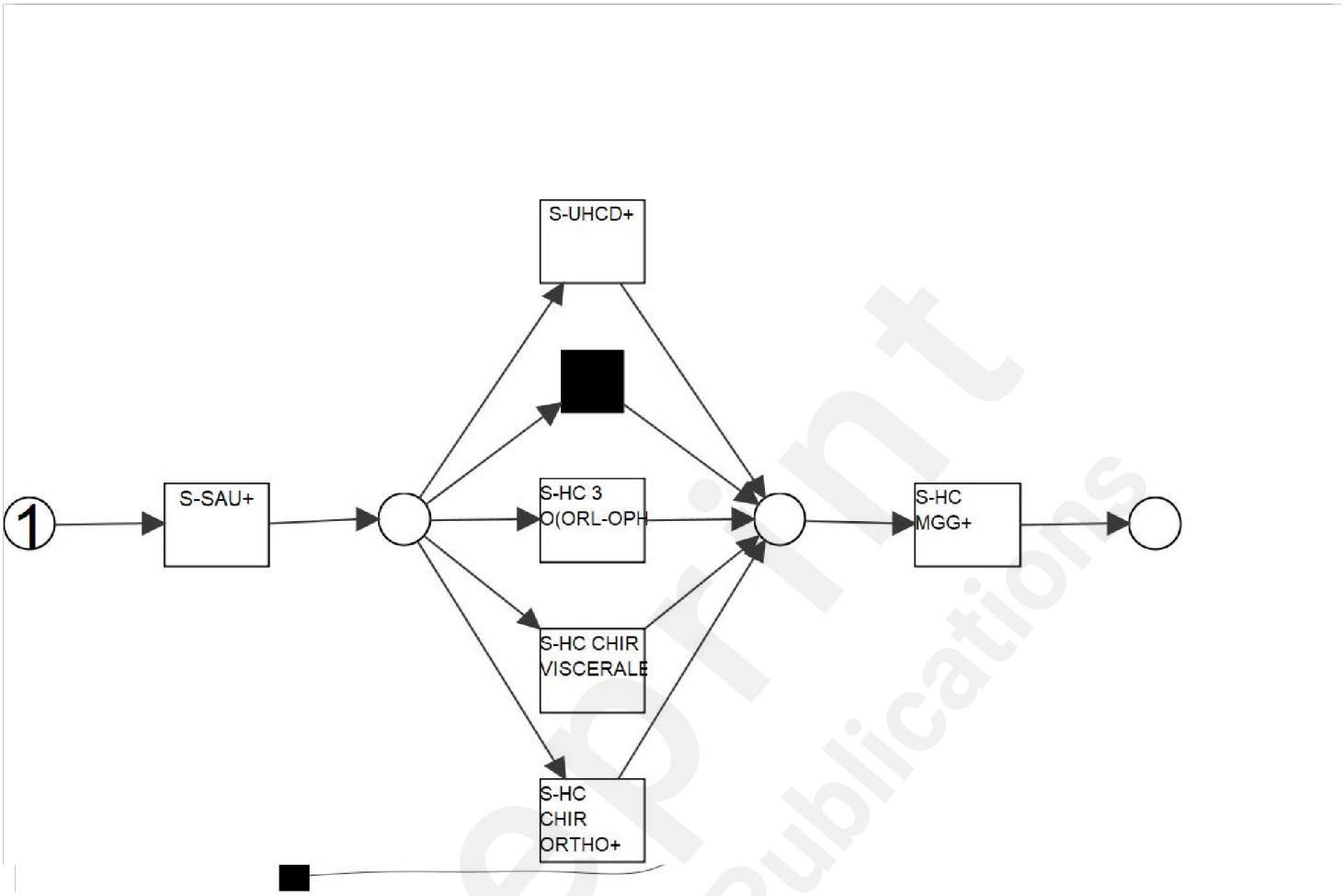




(a) History (b) Corrected

Figure S2: Polyvalent Medicine graphs

(a) History
(b) Corrected



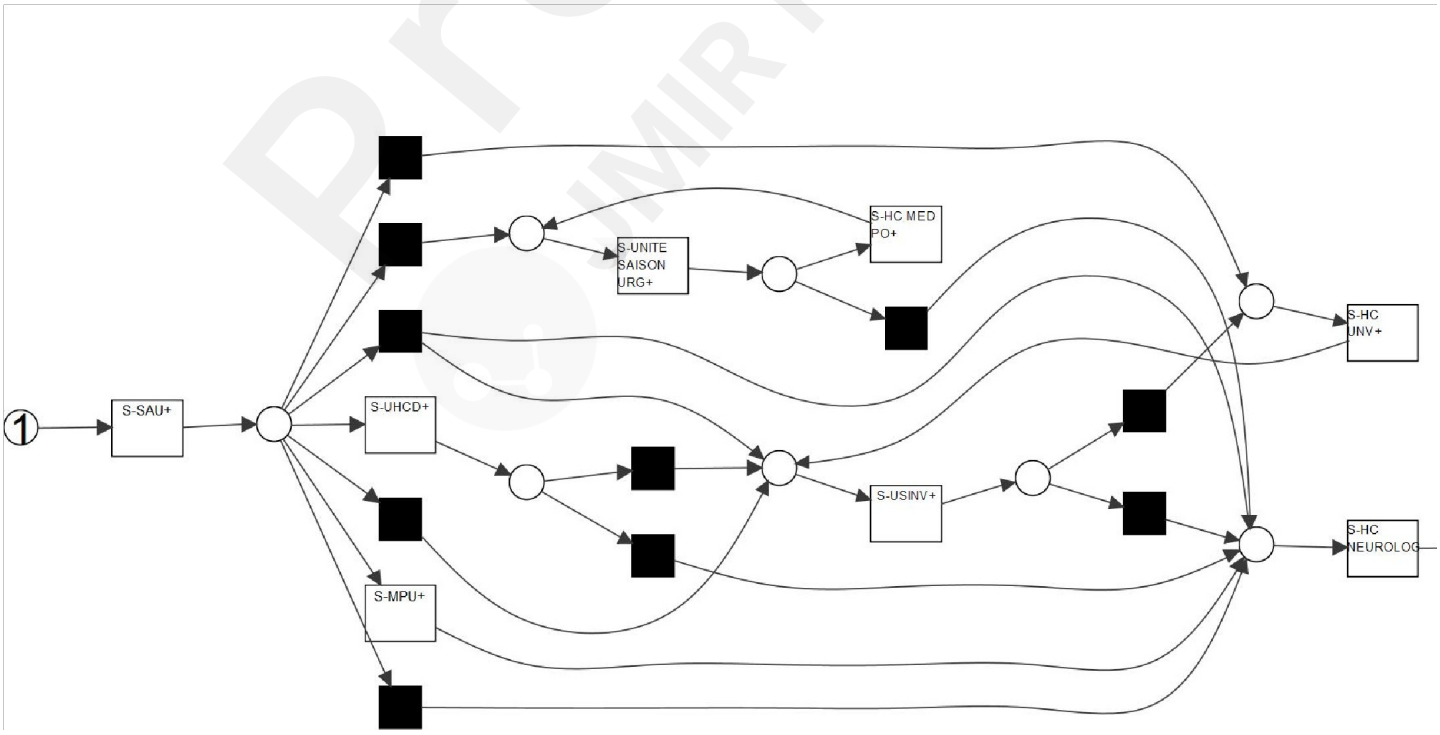
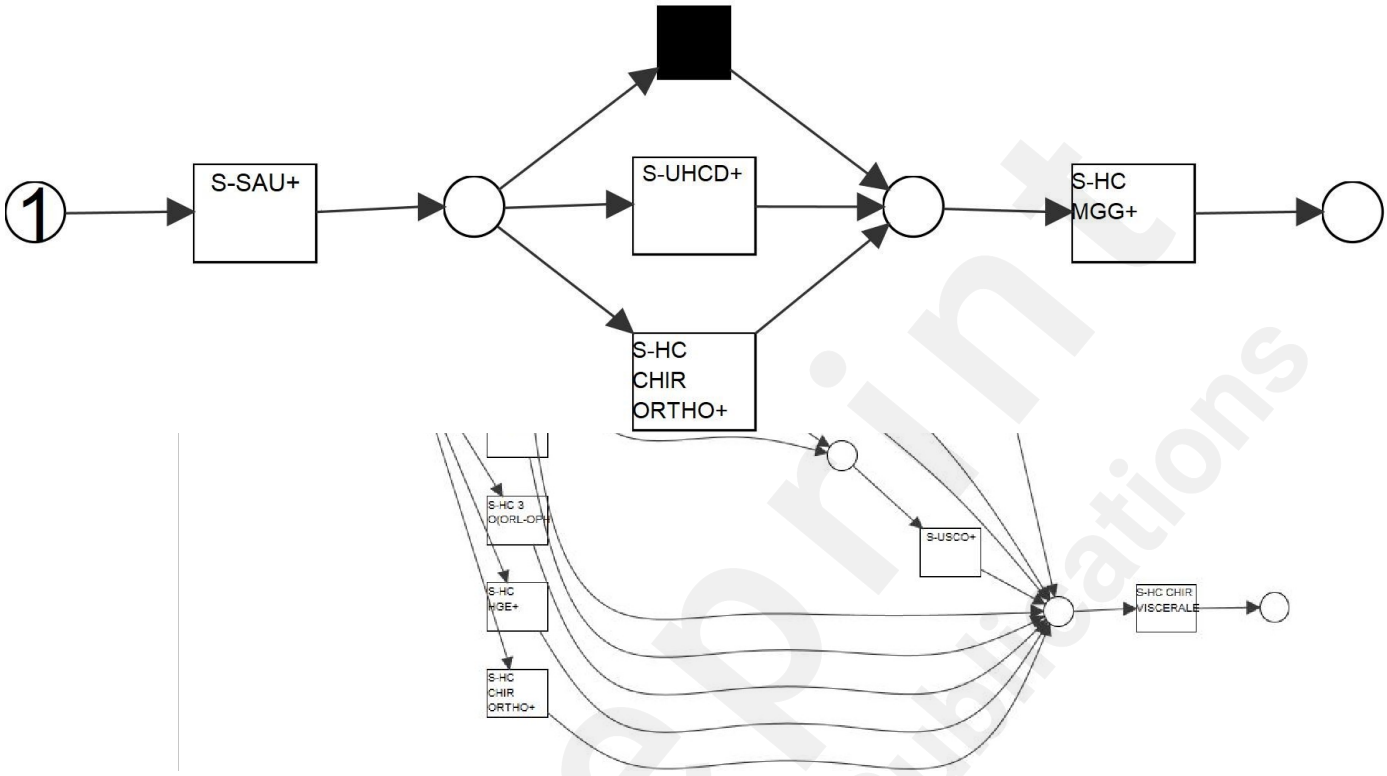


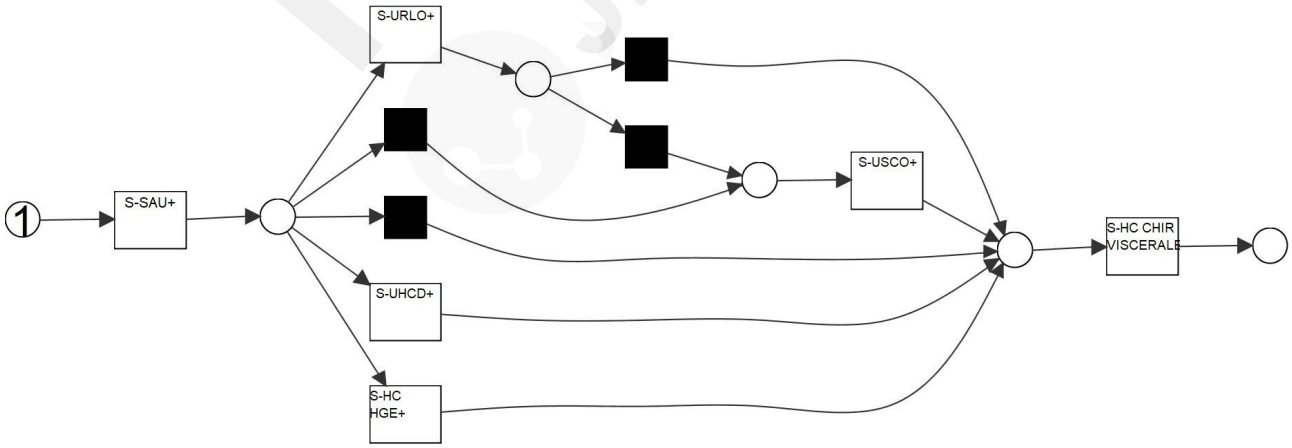
Figure S3: Neurology graphs
(a) History



(b) Corrected

Figure S4: Geriatrics graphs

(a) History



(b) Corrected

Figure S5: Visceral Surgery graphs

Appendix 4: Simulation parametrisation

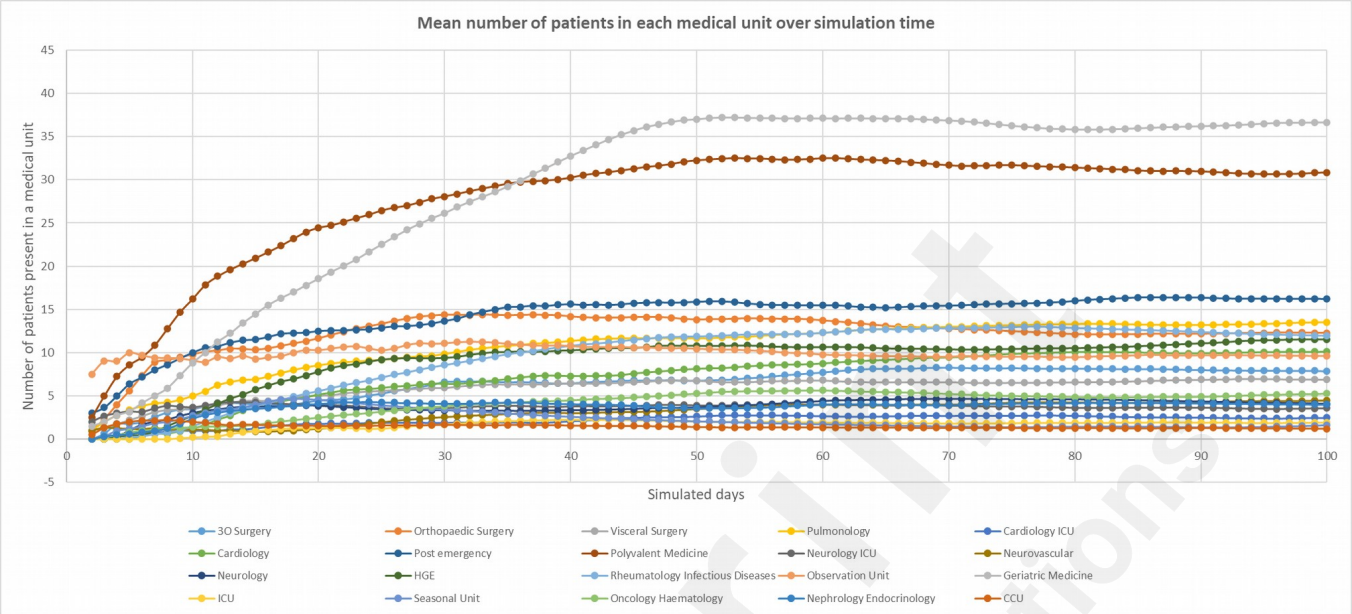


Figure S6: Mean number of patients in each medical unit over 100 simulated days

Table S4: Mean length of stay in days in each medical unit: from the real dataset and the simulation.

	Simulation		**Real dataset**		
Medical unit	**Mean**	**CI95**	**Mean**	**CI95**	**Absolute error**
3O Surgery	2.95	(2.75, 3.16)	3.06	(2.84, 3.28)	0.11
Cardiology	6.04	(5.57, 6.51)	5.72	(5.34, 6.1)	0.32
Orthopaedic Surgery	4.94	(4.57, 5.31)	5.04	(4.69, 5.4)	0.10
Visceral Surgery	3.49	(3.18, 3.79)	3.93	(3.59, 4.28)	0.45
Hepatogastroenterology	6.88	(6.44. 7,32)	6.98	(6.48, 7.48)	0.10
Polyvalent Medicine	10.00	(9.57, 10.42)	10.70	(10.13, 11.26)	0.70
Geriatric Medicine	13.50	(12.99, 14.02)	14.94	(14.1, 15.78)	1.44
Nephrology Endocrinology	8.25	(7.23, 9.26)	10.21	(8.87, 11.55)	1.96
Neurology	6.11	(5.15, 7.08)	7.59	(6.33, 8.85)	1.48
Oncology Haematology	11.71	(10.34, 13.07)	13.42	(11.26, 15.57)	1.71
Pulmonology	9.60	(8.89, 10.31)	9.59	(8.91, 10.27)	0.01
Rheumatology - infectious diseases	9.05	(8.35, 9.76)	9.18	(8.36, 10.0)	0.13
Neurovascular	6.51	(5.59, 7.44)	8.57	(6.81, 10.33)	2.06

Post emergency	5.13	(4.85, 5.42)	4.60	(4.35, 4.84)	0.54
Emergency Department	0.29	(0.28, 0.29)	0.26	(0.25, 0.26)	0.03
Observation Unit	0.67	(0.66, 0.68)	0.67	(0.65, 0.68)	0.00
Seasonal Unit	7.12	(5.94, 8.31)	6.51	(5.25, 7.77)	0.61
ICU	5.49	(4.48, 6.49)	4.98	(3.96, 6.0)	0.50
CCU	2.24	(1.95, 2.54)	2.43	(2.08, 2.78)	0.19
Cardiology ICU	3.05	(2.78, 3.32)	2.91	(2.7, 3.11)	0.14
Neurology ICU	3.54	(3.35, 3.74)	3.37	(3.18, 3.56)	0.17
Mean absolute error					0.61

Appendix 5: Additional definitions

Definition (*Medical Unit*)**.** A medical unit provides treatment and care to inpatients and outpatients. We distinguish three main disciplines:

- Medicine, Surgery and Obstetrics units (MCO): These units provide acute treatments and care in medicine, surgery, obstetrics, odontology, and oncology.
- Rehabilitation units (Rehab): These units provide care for patients who need rehabilitation of a wound organ or who need to be readjusted to limited capacity.
- Psychiatry Units (PSY): These units provide care and treatment for patients with mental health disorders.

A medical unit is defined by a name, an id, a location (the hospital site where the medical unit is) and a discipline as follows:

mu = (name, id, site, discipline)

Definition (*Hospital*)**.** A hospital ensures the diagnosis, monitoring and care of sick people, injured people, and pregnant women.

Definition (*Hospital Group*)**.** A hospital group is a legal institution that includes different hospitals and residential care facilities. Therefore, a hospital group has multiple sites, and a site corresponds to a geographical location.

Definition (*Hospital Stay*)**.** A hospital stay is the period of time an inpatient has spent in the same hospital (site) and in medical units of the same discipline. In the healthcare records each stay is identified by an id. Hence, each time a patient is transferred from one site to another site or hospital, the stay id changes. Similarly, if a patient is transferred from one discipline (e.g. MCO) to another discipline (e.g. PSY), the stay id changes.

Definition 5 (*Length of Stay (LoS*))**.** The length of stay is the duration of a hospital stay. The LoS is generally expressed in days.

Definition 6 (*Discharge from Hospital*)**.** The discharge is the official release from hospital care or from a medical care facility. The discharge disposition is the place where patients go after discharge.

The following definitions are specific to our framework.

Definition (*MCO-stay*). In this work, we need to enlarge the notion of a hospital stay to the hospital group. We considered the stays of patients through all the MCO units of the entire hospital group. Therefore, a stay begins when a patient is admitted to one of the MCO units of the hospital group and ends when the patient is discharged from the last MCO unit or transferred outside the hospital group.

Definition (Discharge from MCO dispositions)**.** We consider seven main types of discharge disposition after an MCO stay:

- Return home
- Return home with the help of community nursing
- Home hospitalisation: this service of home health care service provides complex and medical care at home by preventing patient from being at the hospital
- Transfer to a rehabilitation facility (rehab)
- Transfer to a psychiatric unit or a psychiatric hospital (PSY)
- Admission to a long-term care facility such as a nursing home
- External transfer: transfer to an MCO unit of another hospital group.

Some dispositions require the development of an individual health care plan; and therefore discharge needs to be planned by medical and social staff.

Example. One hospital group had two general hospitals (sites), hospital A and hospital B. Patient Y

visited the ED of hospital A. Then patient Y was admitted to the cardiology unit of the same hospital. After he or she was admitted to the geriatric unit of hospital B. Patient Y went to the rehabilitation unit of hospital B. The patient visited 3 MCO units, 1 Rehab unit, and two hospitals of the same hospital group. The MCO-stay is the period of time from admission to the ED to discharge from the geriatric unit: ED → Cardiology → Geriatrics.
Table S5 shows the match between the concepts of the framework and the example.

Table S5. An example of framework.

Features	**Values**
GH	GHBS
Site	Scorff and Villeneuve
Medical units	ED (1323, Scorff, MCO), observation unit (1321, Scorff, MCO), geriatric ward (1030, Villeneuve, MCO)
Discharge destination	community nursing
Hospital stay 1	stay at Scorff hospital between the 4th of January and the 5th of January
Length of stay 1	1 day
Hospital stay 2	stay at Villeneuve hospital between the 5th of January and the 12th of January
Length of stay 2	7 days
MCO stays	stay in the GHBS between 4th of January and the 12th of January
Patient identifier	0000056098
MCO stay identifier	4560000632
Patient pathway	(0000056098, 4560000632, σ, community nursing)

References

1. Abohamad W, Ramy A, Arisha A. A hybrid process-mining approach for simulation modeling. In: 2017 Winter Simulation Conference (WSC); 2017:1527-1538. doi:10.1109/WSC.2017.8247894
2. Andrews R, Wynn MT, Vallmuur K, et al. Leveraging Data Quality to Better Prepare for Process Mining: An Approach Illustrated Through Analysing Road Trauma Pre-Hospital Retrieval and Transport Processes in Queensland. Int J Environ Res Public Health. 2019;16(7):1138. doi:10.3390/ijerph16071138
3. Bernard A, Boichut MA, Descouts A, et al. Bed manager : mission, profil et activité ?. EHESP; 2019:66.
4. Bernardi FA, Alves D, Crepaldi N, Yamada DB, Lima VC, Rijo R. Data Quality in Health Research: Integrative Literature Review. Journal of Medical Internet Research. 2023;25(1):e41446. doi:10.2196/41446
5. Bose RPJC, van der Aalst WMP. Trace Clustering Based on Conserved Patterns: Towards Achieving Better Process Models. In: Rinderle-Ma S, Sadiq S, Leymann F, eds. Business Process Management Workshops. Springer Berlin Heidelberg; 2010:170-181.
6. Bose RPJC, van der Aalst WMP. Context Aware Trace Clustering: Towards Improving Process Mining Results. In: Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics; 2009:401-412. doi:10.1137/1.9781611972795.35
7. Bose RPJC, Mans RS, van der Aalst WMP. Wanna improve process mining results? In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM); 2013:127-134. doi:10.1109/CIDM.2013.6597227
8. Demir E, Gunal MM, Southern D. Demand and capacity modelling for acute services using discrete event simulation. Health Systems. 2017;6(1):33-40. doi:10.1057/hs.2016.1
9. De Roock E, Martin N. Process mining in healthcare – An updated perspective on the state of the art. Journal of Biomedical Informatics. 2022;127:103995. doi:10.1016/j.jbi.2022.103995
10. De Weerdt J, vanden Broucke S, Vanthienen J, Baesens B. Active Trace Clustering for Improved Process Discovery. IEEE Transactions on Knowledge and Data Engineering. 2013;25(12):2708-2720. doi:10.1109/TKDE.2013.64
11. Delias P, Doumpos M, Manolitzas P, Matsatsinis N. Supporting healthcare management decisions via robust clustering of event logs. Knowledge-Based Systems. 2015;(84):203-213. doi:10.1016/j.knosys.2015.04.012
12. Dixit PM, Suriadi S, Andrews R, et al. Detection and Interactive Repair of Event Ordering Imperfection in Process Logs. In: Krogstie J, Reijers HA, eds. Advanced Information Systems Engineering. Lecture Notes in Computer Science. Springer International Publishing; 2018:274-290. doi:10.1007/978-3-319-91563-0_17
13. El-Bouri R, Eyre DW, Watkinson P, Zhu T, Clifton DA. Hospital Admission Location Prediction via Deep Interpretable Networks for the Year-Round Improvement of Emergency Patient Care. IEEE Journal of Biomedical and Health Informatics. 2021;25(1):289-300. doi:10.1109/JBHI.2020.2990309
14. Elghazel H, Deslandres V, Kallel K, Dussauchoy A. Clinical pathway analysis using graph-based approach and Markov models. In: 2007 2nd International Conference on Digital Information Management. Vol 1. ; 2007:279-284. doi:10.1109/ICDIM.2007.4444236
15. Ethik-IA. La Garantie Humaine dans le projet de règlement sur l'IA de la Commission européenne ! DSIH. Published April 22, 2021. Accessed March 24, 2022. https://www.dsih.fr/article/4215/la-garantie-humaine-dans-le-projet-de-reglement-sur-l-ia-de-la-commission-

europeenne.html
16. Fahland D, van der Aalst WMP. Simplifying Mined Process Models: An Approach Based on Unfoldings. In: Rinderle-Ma S, Toumani F, Wolf K, eds. Business Process Management. Vol 6896. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2011:362-378. doi:10.1007/978-3-642-23059-2_27
17. Franck T, Bercelli P, Aloui S, Augusto V. A generic framework to analyze and improve patient pathways within a healthcare network using process mining and discrete-event simulation. In: K.-H. Bae, B. Feng, S. Kim, et al., eds. Proceedings of the 2020 Winter Simulation Conference; 2020:12.
18. Greco G, Guzzo A, Pontieri L, Sacca D. Discovering expressive process models by clustering log traces. IEEE Transactions on Knowledge and Data Engineering. 2006;18(8):1010-1027. doi:10.1109/TKDE.2006.123
19. Holm LB, Lurås H, Dahl FA. Improving hospital bed utilisation through simulation and optimisation: With application to a 40% increase in patient volume in a Norwegian general hospital. International Journal of Medical Informatics. 2013;82(2):80-89. doi:10.1016/j.ijmedinf.2012.05.006
20. Huang Z, Dong W, Duan H, Li H. Similarity Measure Between Patient Traces for Clinical Pathway Analysis: Problem, Method, and Applications. IEEE Journal of Biomedical and Health Informatics. 2014;18(1):4-14. doi:10.1109/JBHI.2013.2274281
21. Karakra A, Lamine E, Fontanili F, Lamothe J. HospiT'Win: a digital twin framework for patients' pathways real-time monitoring and hospital organizational resilience capacity enhancement. In: Proceedings of the 9th International Workshop on Innovative Simulation for Healthcare (IWISH 2020). CAL-TEK srl; 2020:62-71. doi:10.46354/i3m.2020.iwish.012
22. Lassen KB, van der Aalst WMP. Complexity metrics for Workflow nets. Information and Software Technology. 2009;51(3):610-626. doi:10.1016/j.infsof.2008.08.005
23. Ly LT, Indiono C, Mangler J, Rinderle-Ma S. Data Transformation and Semantic Log Purging for Process Mining. In: King R, ed. Advanced Information Systems Engineering. CAiSE 2012. Vol 7328. Lecture Notes in Computer Science. Springer International Publishing; 2012:238-253. doi:10.1007/978-3-642-31095-9_16
24. Martin N, Martinez-Millana A, Valdivieso B, Fernández-Llatas C. Interactive Data Cleaning for Process Mining: A Case Study of an Outpatient Clinic's Appointment System. In: Di Francescomarino C, Dijkman R, Zdun U, eds. Business Process Management Workshops. Vol 362. Lecture Notes in Business Information Processing. Springer International Publishing; 2019:532-544. doi:10.1007/978-3-030-37453-2_43
25. Mendling J. Testing Density as a Complexity Metric for EPCs. Vienna University of Economics and Business Administration; 2006. https://www.researchgate.net/publication/228347008_Testing_density_as_a_complexity_metric_for_EPCs
26. Munoz-Gama J, Martin N, Fernandez-Llatas C, et al. Process mining for healthcare: Characteristics and challenges. Journal of Biomedical Informatics. 2022;127:103994. doi:10.1016/j.jbi.2022.103994
27. Ordu M, Demir E, Tofallis C, Gunal MM. A comprehensive and integrated hospital decision support system for efficient and effective healthcare services delivery using discrete event simulation. Healthcare Analytics. 2023;4:100248. doi:10.1016/j.health.2023.100248
28. Prodel M, Augusto V, Jouaneton B, Lamarsalle L, Xie X. Optimal Process Mining for Large and Complex Event Logs. IEEE Trans Automat Sci Eng. 2018;15(3):1309-1325. doi:10.1109/TASE.2017.2784436
29. Prodel M, Augusto V, Xie X, Jouaneton B, Lamarsalle L. Stochastic simulation of clinical pathways from raw health databases. In: 2017 13th IEEE Conference on Automation Science and Engineering (CASE). IEEE; 2017:580-585. doi:10.1109/COASE.2017.8256167

30. ProMTools. Prom process mining workbench. URL https://promtools.org/, accessed on 2022-11-22.
31. Rogge-Solti A, Mans RS, van der Aalst WMP, Weske M. Repairing Event Logs Using Timed Process Models. In: Demey YT, Panetto H, eds. On the Move to Meaningful Internet Systems: OTM 2013 Workshops. Vol 8186. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2013:705-708. doi:10.1007/978-3-642-41033-8_89
32. Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: A literature review. Journal of Biomedical Informatics. 2016;61:224-236. doi:10.1016/j.jbi.2016.04.007
33. Selker HP, Beshansky JR, Pauker SG, Kassirer JP. The Epidemiology of Delays in a Teaching Hospital: The Development and Use of a Tool That Detects Unnecessary Hospital Days. Medical Care. 1989;27(2):112-129.
34. Song M, Günther CW, van der Aalst WMP. Trace Clustering in Process Mining. In: Ardagna D, Mecella M, Yang J, eds. Business Process Management Workshops. Springer Berlin Heidelberg; 2009:109-120.
35. Uhl L, Augusto V, Lemaire V, et al. Progressive prediction of hospitalisation and patient disposition in the emergency department. In: 2022 IEEE International Conference on Big Data (Big Data); 2022:1719–1728, doi:10. 1109/BigData55660.2022.10020777
36. van der Aalst W. Process Mining. Data Science in Action. second. Springer Berlin Heidelberg; 2016. doi:10.1007/978-3-662-49851-4
37. van Eck ML, Lu X, Leemans SJJ, van der Aalst WMP. PM²: A Process Mining Project Methodology. In: Zdravkovic J, Kirikova M, Johannesson P, eds. Advanced Information Systems Engineering. Vol 9097. Lecture Notes in Computer Science. Springer International Publishing; 2015:297-313. doi:10.1007/978-3-319-19069-3_19
38. van Zelst SJ, Mannhardt F, de Leoni M, Koschmider A. Event abstraction in process mining: literature review and taxonomy. Granul Comput. 2021;6(3):719-736. doi:10.1007/s41066-020-00226-2
39. Vanbrabant L, Martin N, Ramaekers K, Braekers K. Quality of input data in emergency department simulations: Framework and assessment techniques. Simulation Modelling Practice and Theory. 2019;91:83-101. doi:10.1016/j.simpat.2018.12.002
40. vanden Broucke SKLM, De Weerdt J. Fodina: A robust and flexible heuristic process discovery technique. Decision Support Systems. 2017;100:109-118. doi:10.1016/j.dss.2017.04.005
41. Veiga GM, Ferreira DR. Understanding Spaghetti Models with Sequence Clustering for ProM. In: Rinderle-Ma S, Sadiq S, Leymann F, eds. Business Process Management Workshops. Springer Berlin Heidelberg; 2010:92-103.
42. Verhulst R. Evaluating Quality of Event Data within Event Logs: An Extensible Framework. Master. Eindhoven University of Technology; 2016.
43. Wood RM, Murch BJ. Modelling capacity along a patient pathway with delays to transfer and discharge. Journal of the Operational Research Society. 2020;71(10):1530-1544. doi:10.1080/01605682.2019.1609885

Abbreviations

DES: Discrete-Event Simulation
ED: emergency department
EHR: electronic health records
GHBS: Groupe Hospitalier Bretagne Sud

ICU: Intensive Care Unit
LoS: Length of Stay
MCO : Medicine (Médecine), Surgery (Chirurgie), Obstetrics (Obstétrique) and Odontology (Odontologie)
OU: observation unit

Supplementary Files

Figures

Hospital group structure.

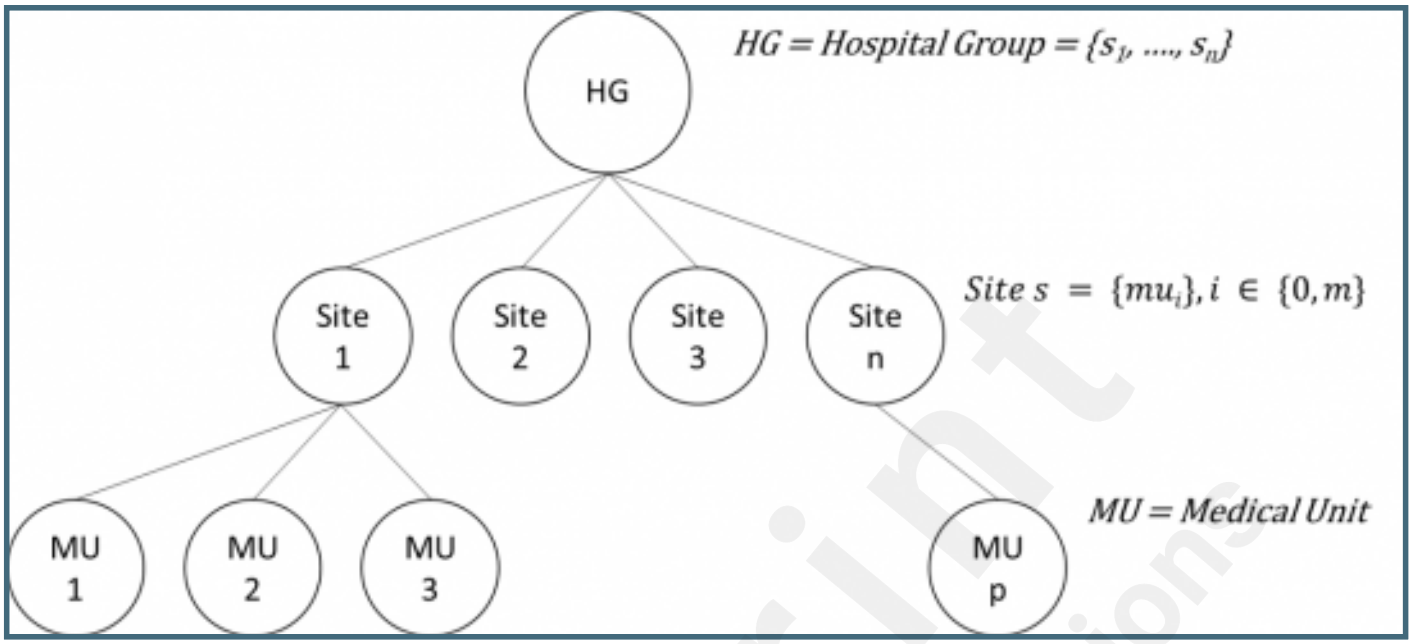
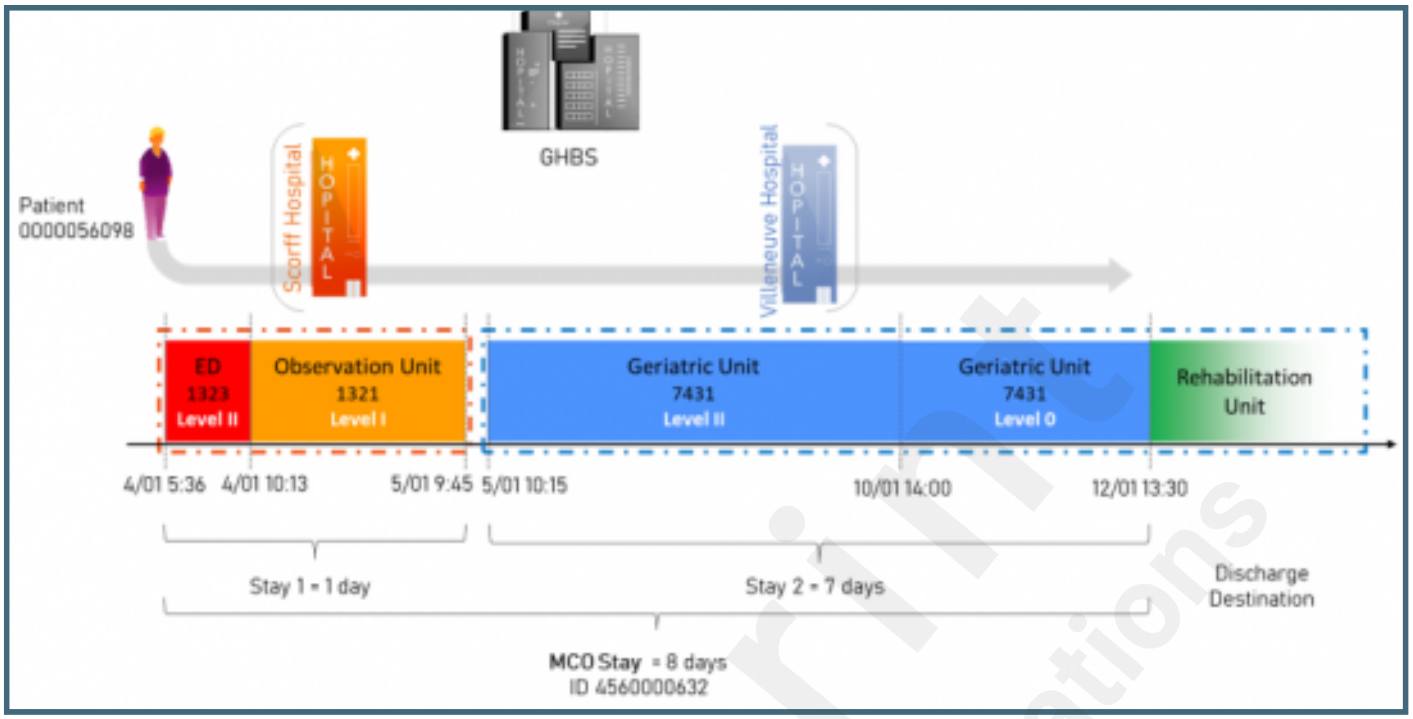
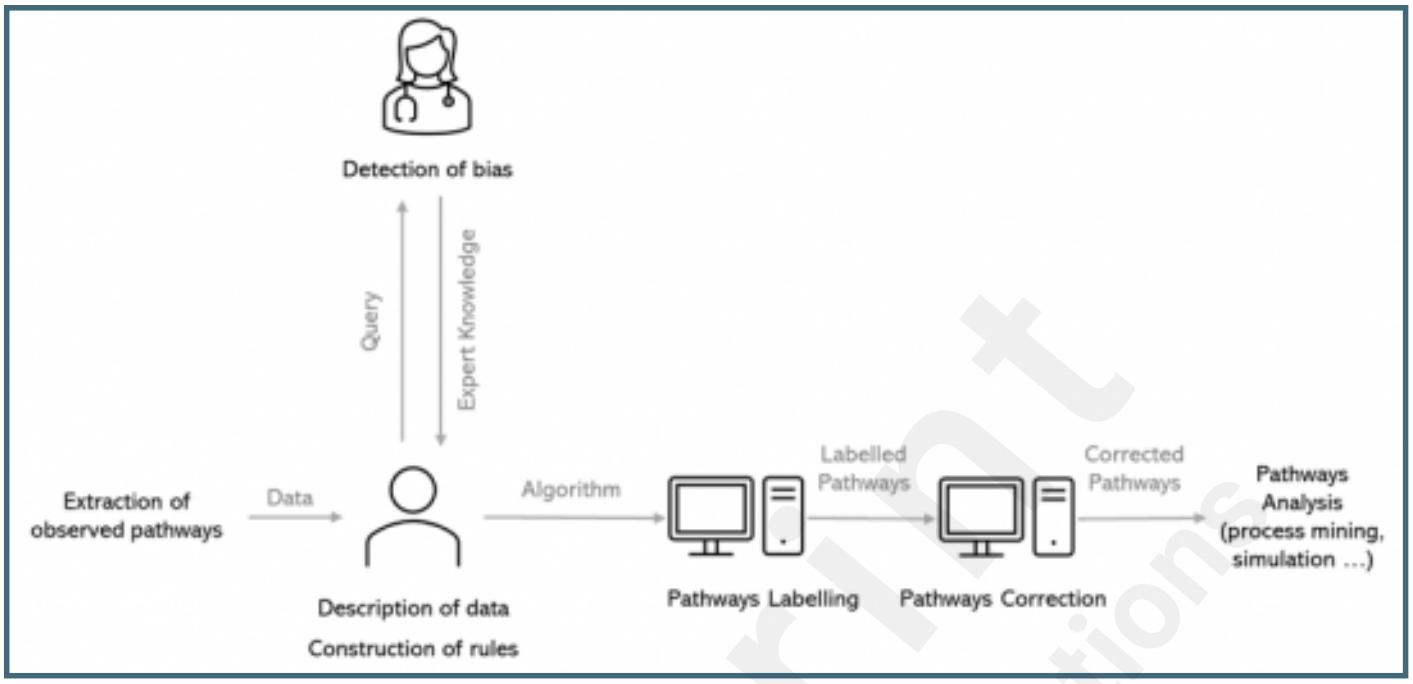


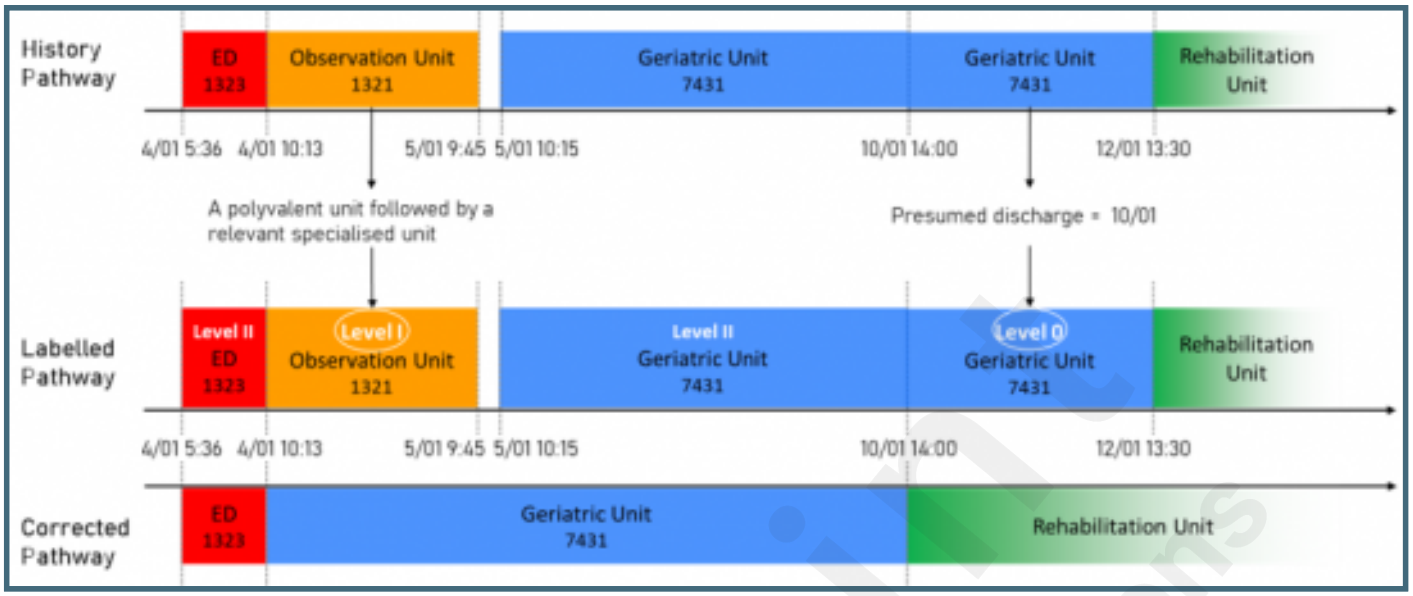
Illustration of the framework with a fictive pathway and patient.



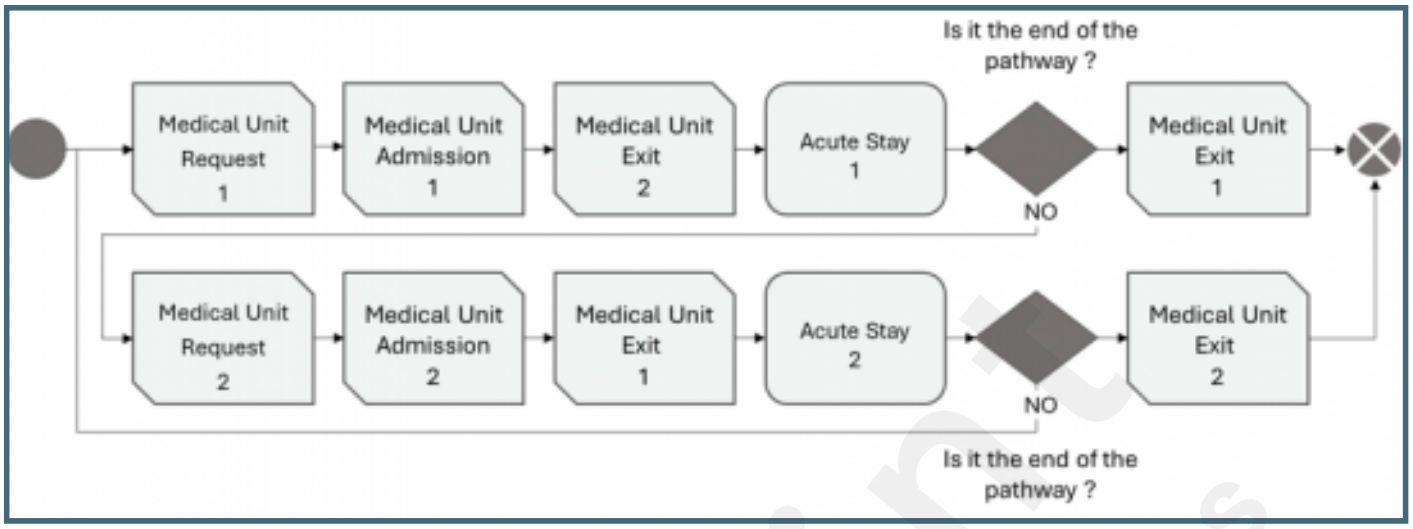
Pathways Labelling Method.



Example of the correction of a pathway.



Modelling of Patients Flows through Medical Units.



Multimedia Appendixes

Algorithms of the rules.
URL: http://asset.jmir.pub/assets/fbeba3a9ab9f6d1a9106dea6e3971051.pdf

Satistical analysis results.
URL: http://asset.jmir.pub/assets/8a50a0de0f0c19a07e905a2dbe63fcc6.pdf

Process models.
URL: http://asset.jmir.pub/assets/1ff011900ed5e549bcebb67d0ee9dcb5.pdf

Simulation parametrisation.
URL: http://asset.jmir.pub/assets/132af0595c3964aa2ae0f2b4a7bb3aca.pdf

Additional definitions.
URL: http://asset.jmir.pub/assets/4da7c2a13ce054cbf7e0be225f72d75e.pdf