Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

AUTOMATIC POPULATION-BASED RESPONSIBILITY MODELING USING PROCESS MINING: APPLICATION TO CHRONIC OBSTRUCTIVE PULMONARY DISEASE

Ana Lucia Tula[1], Vincent Augusto[1], Xavier Boucher[1], and Marianne Sarazin[2]

[1]Center for Biomedical and Healthcare Engineering, UMR CNRS 6158 LIMOS, Saint Etienne, FRANCE
[2]Institut Pierre Louis d'Epidemiologie et de Sante Publique, Hopital Saint Antoine Sorbonne Universite, Paris, FRANCE

ABSTRACT

Population-based responsibility pursues three objectives: better health and better care at a better cost. The project aims to apply this paradigm using a process mining approach to build the clinical pathway of the population suffering certain disease to finally test if the process model well represents its clinical pathway. We asses our approach on a cohort of patients affected with Chronic Obstructive Pulmonary Disease. We use a national medico-administrative database of hospitalizations to extract our population, we stratify the disease and apply process mining. We propose different models with different rules to extract event logs and a design of experiments to compare the models using quantitative indicators: fitness, precision, generalization, simplicity and replicability through simulation. We also propose a qualitative evaluation of the best models following medical expert opinion. Our approach confirms that the models well represent the medical records and the simulation partially replicates them.

1 INTRODUCTION

The approach called "Populational-based responsibility", originally called 'responsabilité populationnelle' in French, is a new healthcare paradigm originated in Quebec (Trottier 2013). This paradigm induces that all health actors in a territory are responsible for improving the health of their population as well as the care of the patients. The intention is to develop prevention, using the territory resources in the best way, in order to prevent people with only risk factors to develop the pathology or already affected people to get their condition deteriorated.

The steps of the paradigm are:

- First, to build a stratification of a chosen population according to certain features given by the affliction selected.
- Second, the health professionals of the territory develop a clinical program adapted to this population and its resources, the indicators to measure and the actions to be implemented.
- Finally, once the program is defined, a cohort of patients that correspond to our chosen population follows the experimental protocol leading to a continuous analysis of the procedure and its results.

Following the first step, the stratification of the disease would be medico-economical as Figure 1. The model is constituted by four phases: in phase zero, the patient starts by being exposed to risk factors but may not develop the disease. If the symptoms are present, the person gets diagnosed and may start a treatment ascending to phase one. If the condition of the patient does not deteriorate, i.e. stay stable, he stays in the same phase. However, if that is not the case and the patients' health gets deteriorated irreversibly, e.g. an emergency hospitalization, they pass on to phase two and stays there, as long as they do not suffer further health deterioration in the last stage of the disease, corresponding to phase three (for example, dependency of an internal or external medical device). The last patient status after step 3 would be death. The number

of phases depend on the disease. In summary, being stable is represented by staying in the same phase, and getting sicker by moving forward into the model.

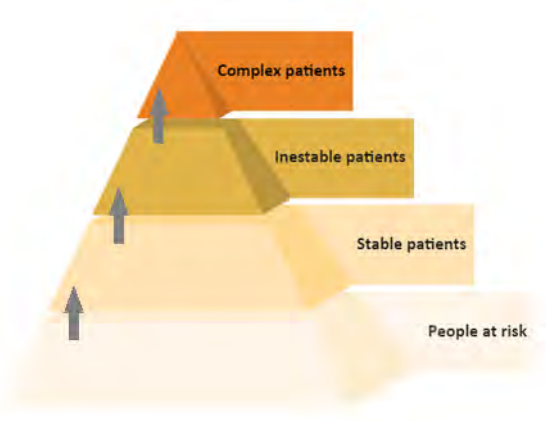

Figure 1: Populational-based responsibility stratification represented by a pyramid.

However, such an approach is difficult to apply in practice since it requires a lot of work regarding the analysis of existing data by health practitioners to model the clinical pathways. The approach is difficult to reproduce depending on (i) the cohort of patients (the pathology under study), (ii) the data availability, (iii) the availability of medical experts.

Considering such limitations, our objective is to automate the application of this paradigm to new diseases by using a process mining approach to model its clinical pathway. Process mining allows to model, analyze, and optimize processes by extracting knowledge from the process records. Applied in the context of population-based responsibility framework it can concretely help to extract information and patterns from the patients' clinical path and to convert it into a process model. Nonetheless, the added-value of process mining in this objective is very challenging due to the strong diversity of patient paths for the same disease and thus the models complexity would reduce the possibilities of knowledge extraction.

The paper proposes a methodological approach to deal with this complexity, by generating a set of alternative models, whose quality can be assessed rigorously. The approach is applied to the specific disease of Chronic Obstructive Pulmonary Disease (COPD) so we use medico-administrative data (French national hospitalization database) to feed the model. The following will be the structure of this paper: section 2 relates state of the art elements, section 2.3 and 2.4 explains key basics of process mining and healthcare concepts respectively, and section 3 explains the methodological framework proposed. Section 3.6 applies the methodological approach to COPD and presents key results in section 4 to be further discussed in section 5.

2 LITERATURE REVIEW

2.1 Populational-Based Responsibility

The approach called "Populational-based responsibility" (Trottier 2013) was launched by the Fédération hospitalière de France (FHF) in France in 2017 and integrated in the law in 2019. The first results of this paradigm were published in (Malone 2023) and are encouraging since for diabetes it has allowed to decrease the emergency admissions hospital stays caused by the illness, yet the ambulatory admission stays have increased since there is a better coordination and communication between the actors to plan hospital visits and discharges in times meaning that the patients no longer arrive through emergencies. Other used indicators show longer hospitalizations (more than five days) were lower than expected. These results should confirm the health-economic interest of the approach and thus justify its extension to other pathologies.

In (Gomez et al. 2020) and (Gomez et al. 2023) the authors propose different methods to find the interested populations using the PMSI database (French acronym for the national hospitalization database) with its four fields of activity either with a verification in the AMO (French acronym for "Compulsory Health Insurance") data or complementing it with other databases, for example the SNIIRAM (French acronym for the national claim database) database for ALD (French acronym for "Long term diseases") and medication consumption.

2.2 Process Mining

This technique consists of three forms, process discovery, conformance checking and extension of the model (Van der Aalst 2011). The first automatically discovers the relation between activities, its transitions and convert this information into a model, through certain algorithms, for example, Alpha algorithm, Heuristic miner or the fuzzy miner, the algorithm that Disco, an open source process discovery software, is based on (Günther and Rozinat 2012). For 'conformance checking', we either generate 'traces' in the model and compare them with the event log traces or we play them into the model to quantifying the difference between the model and the data by using indicators. This is applied using either token based replayability or alignments algorithm (Rozinat and Van der Aalst 2008), (Van der Aalst et al. 2012) from which are calculated fitness, precision, generalization and simplicity indicators. The third process mining form is out of the scope.

Another way to validate the generated model consists in using a simulation to generate an artificial log and compare it with the original event log (Van der Aalst 2018). Delving into the scope of using process mining and simulation, we got (Augusto et al. 2016) proposing a new methodology to perform simulation analysis of patients' clinical pathways extracted by process mining applied to cardiovascular diseases to study the impact of medical decisions. The same methodology was later applied for incisional hernia (Phan et al. 2019). Additionally, simulation in process mining has other purposes, for example, model validation, optimization and prediction as in (Van der Aalst 2018), where also several indicators are proposed. In the same field, (Fani Sani et al. 2020) overcomes the computational time that represents to calculate the alignments in big data process by proposing to simulate the process model, generate traces and calculate with those the conformance.

2.3 Basics of Process Mining

The objective of this work is to apply process mining in order to extract useful knowledge on the process discovery of the patients' path from the various stratification layers considered for Populational-based responsibility framework. For better understanding of the following sections, we introduce basics of process mining applied to healthcare.

Definition 1 Event. An event E is an element with a set of n attributes: Time stamp, activity type, case ID, etc. $E = (a,b,c,...,n)$

Definition 2 Trace. A trace T is a ordered length i sequence of events. $T = (E_1,E_2,E_3,...,E_i)$

Definition 3 (Event) Log. A log L is a set of j different traces T, at the same time, being T a set of events. $L = (T_1,T_2,T_3,...,T_j)$

Definition 4 Event class. An event class C is a size m subset of a certain attribute of the event E. $C = (a_1,a_2,a_3,...,a_m)$. Each event E over L will have an attribute that will correspond to an event class C.

Definition 5 Process model. A Process Model PsM is a visual event log representation composed of a set N of nodes representing the event class C and a set of arcs K, y transitions determined by either order or timeline between each event classes. $PsM = (L,K) = (\{n_1,n_2,\ldots,n_m\},\{k_1,k_2,\ldots,k_y\})$

2.4 Healthcare Context

Definition 6 Clinical pathway. A clinical pathway is a multidisciplinary sequence to get an specific diagnostic or a certain procedure. Is a structured care plan used to translate guidelines into local structures, detailing the steps in a course of treatment or care in a plan, pathway, guideline or protocol and it aims to standardize care for an specific clinical problem, procedure or episode of healthcare in an specific population.

Definition 7 Activity. Represents either medical acts, medical consultations or hospitalizations.

Definition 8 Medical consultation. A medical consultation is a meeting in the context of a doctor's office or any other care structure, between a doctor and a patient. It allows the doctor to express an opinion on the patients' symptoms, to establish a diagnosis, and generally to dispense prescriptions. It is represented by a two-digit code that represents the specialization of the medical consultation.

Definition 9 Medical act. Any act whose realization, physical or instrumental, is carried out by a member of a medical profession. For example: spirometry, electrocardiogram. It is represented by an alphanumeric code given by CCAM (Classification Commune des Actes Médicaux in France) that represents the classification of the medical act.

Definition 10 Hospitalization. It can be surgeries, before or after them, monitoring, examinations or stays in the hospital for one or more days. It is represented by an alphanumeric code given by ICD (International classification of diseases) that represents the disease or medical condition that caused the hospitalization.

Definition 11 Clinical pathway model. The general clinical pathway is represented as a process model obtained from the data, or event log, having as nodes the medical consultation specialization, the medical act or the hospitalization diagnosis, and as edges the temporal or ordered relation between them.

Our datasets, or event logs L, will be the set of clinical paths given by activities performed by all patients belonging to the chosen population for an specific time in clinics or hospitals. Where the traces T are each patients' path and each event E is an unique occasion where a specific patient carried out an activity on an specific date.

3 METHODOLOGY

3.1 Framework Overview

The structuring steps of this project are summarized in Figure 2. The objective of the framework is to make possible extracting knowledge in a pertinent way from the available data. The project is confronted to a knowledge extraction challenge resulting from the available data inducing extraction complexity, because it gather both meaningful and non-meaningful pieces of information and they embed a strong diversity of personalized patient paths. Thus, challenge lies in the fact that there is no a priori pertinent way to apply and to optimize process mining to such complex data sets. To address it, the approach proposed consists of applying recursively process mining on a set of alternative data inputs generating a set of alternative process models which could all contribute to distinct knowledge extraction. In the current set of our research, we propose to measure and compare the added-value of these various models by using assessment indicators formalized to test and evaluate the pertinence of these models.

In step 1 we stratify the patient hospital path into 'phases' according to the populational-based responsibility framework. Ideally, this stratification should be confirmed or made by a specialist or health authority. Step 2 simply consists of extracting a pertinent data set according to the desired population and step 3 consists of pre-processing the data in order to configure different data input files, as further explained in section 3.3. Following, the event log files will be submitted to the process mining step where it will be transformed to process models by the use of a specific tool (Disco), Step 4.

Step 5 consists of assessing and comparing the process models. The quality of a given model depends on the good quality representation of the transitions between the distinct phases, the representativeness of the most common activities for the disease as well as for each phase and the capacity to show the different

connections between activities. To make possible a systematic comparison of the various models generated, we propose a set of indicators (see section 3.5). The last step 6 consists of converting these process models into simulation models with the objective of providing complementary indicators to assess the added-value of the models and a way to generate additional knowledge concerning patient paths by the power of patient flows virtualization.

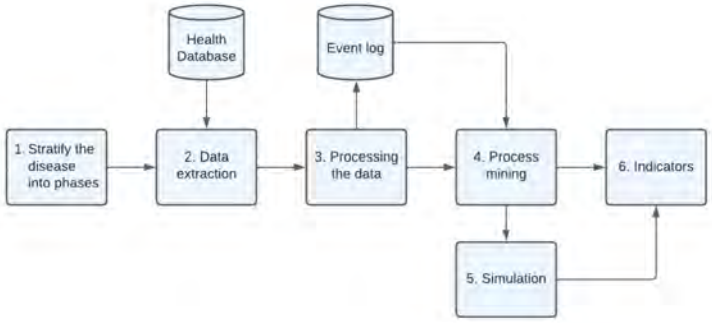

Figure 2: Project steps.

3.2 Data Preprocessing

Most of the filters we apply are meant to reduce the diversity without compromising the disease representation. The filters consist of shortening ICD or CCAM codes, for example, ICD I50.1 code (Left ventricular failure) to I5 (heart failure). The second filter consist of deleting activities according to the number of people who performed it, at least ten percent of the population should have performed it. Other filters consist of including only the disease related activities. Additionally, for some models we consider patients skipping phases (jumping from phase 0 to phase 1 for instance) since it could represent under-diagnosed diseases. At the same time, we group similar and not so relevant activities, for example all X-Ray articulations instead of showing each X-Ray articulation. Moreover, in all datasets we filter non-important and redundant activities. Thus, each rule or combination of rules will lead to an specific pre-filtered input data set available for process mining as settled in Table 1.

Table 1: Table containing the set of models with its respective rules applied.

	Codes shorten	Activities filtered	Only disease related	Skipping phases	Join similar activities	Filter paths
Model 1	X	X (15%)		X	X	X
Model 2	X	X (10%)	X	X	X	X
Model 3	X	X (15%)			X	X
Model 4	X	X (10%)	X		X	X
Model 5		X (15%)		X	X	X
Model 6		X (15%)			X	X
Model 7	X	X (10%)			X	X

3.3 Process Mining and Simulation

For the process discovery, we used Disco Software which transform each event log into a Process Model which will represent in our case one general and stratified clinical path for the entire disease. One of its advantages is that we can choose the model complexity by using two hyperparameters: how many activities

and connections between them, or *paths*, we want to show. In our case, we choose to see 100% of activities and only 15% or 10% of paths, representing the last filter in Table 1. By exporting the model to Python we proceed with the calculation of the indicators given by the library PM4PY.

As for the simulation we used the software AnyLogic(R) and a multi-agent based simulation in which we recreate a Clinical Pathway State Chart (CPSC) according to (Augusto et al. 2016). Being the input the process model PsM and the output the state chart representing the disease clinical pathway. To make the conversion, we need to generate one agent type with at least one thousand agents and to put the process model as their behavior. The nodes, at the same time the event classes, represent each state and the arcs represent the transitions between states. The probability of transitioning between activities, such as from a_1 to a_2, is calculated by dividing the number of patients who made this transition by the total number of patients who performed a_1. In other words, for each activity, we divide the number of patients who transitioned to a specific outcome by the total number of patients who started from that activity.

Once modeled the CPSC, we simulate the model and extract from its database the 'agents state chart states log' thus generating artificial traces or, clinical paths, for each one of the agents. Finally, we compare the artificial event log with the original one using the Levenshtein similarity function in python imitating the similarity indicator explained in the next section.

3.4 Indicators

As explained in 2.2, after the process discovery the conformance checking is done. It is mainly done by measuring the different indicators: fitness, precision, generalization and simplicity. There are many proposed algorithms to measure them, however, in our case we apply those already computed and implemented in the library PM4PY in Python, that at the same time are based in (Berti and Van der Aalst 2021), (Van der Aalst et al. 2012), (Munoz-Gama and Carmona 2010), (Adriansyah et al. 2014), (Buijs et al. 2014) and (Blum 2015), nevertheless, a short explanation will be addressed in this section.

A model exhibits perfect fitness when it can replay all traces in the event log capturing the *system* behavior. On the other hand, precision assesses whether a model allows behaviors not observed in the event log, lack of precision, or 'underfitting,' indicates deviations from observed behaviors. In contrast, generalization measures the model's ability to represent unobserved behaviors from the log. Simplicity mainly concerns visual clarity, ensuring that the model is easily understandable. Ideally, all these indicators should be high to accurately represent existing data and enable the model to accommodate new data effectively.

3.4.1 Token Based Replay

TBR is a technique that allows us to re-play each trace of the event log into the Petri Net model and quantify the trace representation in the model. It is based on the token, or markers, counting in the Petri net places using four counters to calculate the fitness of a trace during the replaying: being m the missing tokens (tokens needed to be added to finish the replayed trace), r the remaining token (token left in the model when the replay already finished), c the consumed token and p the produced token.

3.4.2 Alignments

The alignment consists of comparing a log trace, i.e. an ordered sequence of activities in the log, and a trace generated by the process model to calculate the differences, as *distance*, or *similarities* between them. The *distance* is calculated as the amount of missing parts, or mismatches, in both sequences whereas the *similarity* counts how many coincidences there are. To illustrate the idea, in Table 2, we got two horizontal activity sequences, **a** and **b**:

In our example the distance between these two sequences will be two (2) and the similarity will be three (3). To quantify the fitness we search in the model the most optimal alignment (the one with less

Table 2: Two rows representing activity sequences.

a	cardiologist	gastroenterologist	electrocardiogram	>>	generalist
b	cardiologist	>>	electrocardiogram	appendicitis	generalist

distance) and the worse one (only mismatches in the model sequence M part and the same for the event log sequence L) for each trace and compare.

3.4.3 Precision

As there are two ways to measure fitness, there are also two ways to measure precision. Both of them use the concept of prefix automaton in which each state corresponds to a unique prefix of the event log and the transitions correspond to the activities. For a better understanding, for a trace a = <cardiologist, gastroenterologist, electrocardiogram, generalist> its prefix will be <cardiologist>, <cardiologist, gastroenterologist> and <cardiologist, gastroenterologist, electrocardiogram>.

The TBR approach in PM4PY bases the indicator in the Escaping Edges Precision, where the automaton's next state is determined by the activities allowed by the process model. Transitions that are allowed by the process model but not observed in the event log are called escaping edges. Precision is calculated by dividing the number of escaping edges by the total number of edges.

In the alignment case, the precision is called representative-align ETC. It calculates the prefix automates based in the set of the optimal model trace alignments instead using the event log traces. The use of the automates is the same but it adds a weight in each prefix state according to the frequency with which they appear in the event log. Where for each state of the automaton, we compute its set of possible direct successor activities according to the model and then compare it with the set of activities really executed in the log.

3.4.4 Generalization and Simplicity

Generalization and simplicity are the two simplest indicators to compute. The generalization indicator is based on the alignment provided by the replay fitness and it measure how often the parts of the process model are used while replaying. And the simplicity metric, is based on the average numbers of both incoming and outgoing arcs per node in the process model. Simplicity is the only indicator that does not need to take into account the event log.

4 RESULTS AND INTERPRETATION

4.1 Case Study

COPD is a chronic inflammatory lung disease that includes abnormalities in the small airway of the lungs leading to dyspnea, airflow limitation, for example, mucus blocking the airways, inflammation, swelling of the airway lining and destruction of lung parts. Its main causes are tobacco smoke, indoor air pollution and occupational dust. Also, individual factors may be involved, e.g., genetics (Sandelowsky 2021).

The diagnosis is made by spirometry, a FVC (forced vital capacity) below or equal 70% confirms the diagnosis. In the disease curse, there are acute cases of worsen, the so-called exacerbations, in this case the condition could last weeks and patients could need extra treatment or even be admitted at the hospital for emergency care. After years, patients in very severe case of COPD oxygen therapy is prescribed for long term (Koczulla et al. 2018), (Branson 2018). People with COPD are at increased risk of developing heart disease, lung cancer, and several other conditions (Anzueto 2010). The exacerbations are related to morbidity, mortality and healthcare costs. Therefore, early detection and treatment are important to reduce the risk of exacerbations and slow the deterioration of the disease.

According to WHO (World Health Organization 2023) COPD is the third leading cause of death worldwide and the last information gathered in France, showed it affected approximately 7.5% of the adult

population, meaning 3,5 million persons, and the most important part is that there is a large proportion of cases that are not diagnosed: between 2/3 and 90%. According to the INSERM , In France in 2015, approximately 150,000 for people with a serious stage of illnesses and older than 45 years, were benefited from oxygen therapy of long duration.

4.2 COPD Context

Definition 12 Population. People at risk of contracting COPD or already affected by it.
Hypothesis. Since the database does not store the results of the spirometries we will consider the first spirometry as the spirometry with 70% or less FVC.

Definition 13 Phase 0. State 'at risk'. It is the set of activities that goes before the first spirometry (excluded), in this state we cannot find hospitalization or oxygen therapy.

Definition 14 Phase 1. State 'Diagnosed'. It is a set of activities that goes from the first spirometry, to pneumonologist, until the hospitalization for COPD (excluded), in this state we cannot find oxygen therapy and hospitalization.

Definition 15 Phase 2. State 'Hospitalized'. It is a set of activities that goes from the first hospitalization until the first oxygen therapy (excluded), in this state we cannot find oxygen therapy.

Definition 16 Phase 3. State 'Oxygen therapy'. It is a set of activities that goes from the first oxygen therapy until the death of the patient

4.3 Qualitative Results

In this section we proceed to show one of the models in Figure 3, Model 7, in which we can see all the four phases separated by 'bow ties'. However, since this model is too complex to be visually explained, we built a small and simple clinical path graph to explain in Figure 4. If we have 26 patients as population, it will mean they started their path by seeing a generalist, four of them will go directly to perform an spirometry, meaning that 22 patients have other particular path not general enough to be shown in the model, however 15 of them also arrive to perform an spirometry, reaching 19 patients. And thus, we can see which are the most frequent paths as well as the least frequents ones, remembering that the number of paths seen is a hyper-parameter to change freely in Disco software. Nevertheless, the most paths we show the most spaghetti-like our model gets. In most models, activities related to COPD are seen, for example, chest CT scan, pulmonologist consultations, plethymography, gasometry and even hospitalizations for flu, pneumopaties and respiratory failure. Regarding to the activities non disease related we can see very diverse activities as control exams, cardiologist consultation, eyes related studies or health failure hospitalizations.

Figure 3: Model 7.

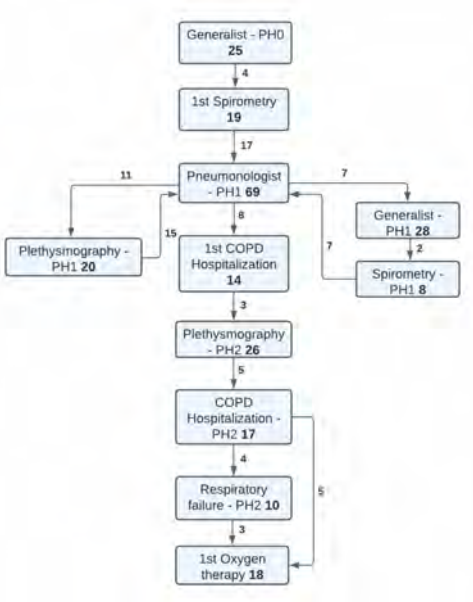

Figure 4: Example of simplified process model.

4.4 Quantitative Results

The numerical results are given in Table 3 and its graphical representation in Figures 5 and 6. The trace indicators were given to illustrate in writing the diversity of the different event logs used to build the models since the event classes will later become nodes in the process model and the trace length addresses the average amount of events that each patient performed. Concerning the graph indicators, we present precision, fitness, generalization, simplicity and similarity between the artificial traces generated by simulation and the event log traces, all of them already explained in subsection 3.4.

Table 3: Table containing the results of the measured indicators.

	Traces indicators			Graph indicators						
	Event classes	Number of patients	Traces length	Fitness		Precision		Generalization	Simplicity	Simulation similarity
				TBR	Align	TBR	Align			
Model 1	112	1118	35 (1, 602)	0.811	0.704	0.832	0.856	0.809	0.486	0.491
Model 2	62	1118	18 (1, 169)	0.825	0.760	0.863	0.863	0.819	0.538	0.674
Model 3	120	79	50 (5, 218)	0.803	0.729	0.829	0.851	0.558	0.529	0.377
Model 4	62	79	31 (5, 131)	0.808	0.776	0.853	0.853	0.600	0.552	0.564
Model 5	105	1118	27 (1, 208)	0.774	0.675	0.876	0.90	0.842	0.541	0.544
Model 6	112	79	44 (5, 244)	0.739	0.657	0.856	0.865	0.601	0.592	0.456
Model 7	130	14	61 (34, 128)	0.638	0.533	0.939	0.951	0.303	0.648	0.350

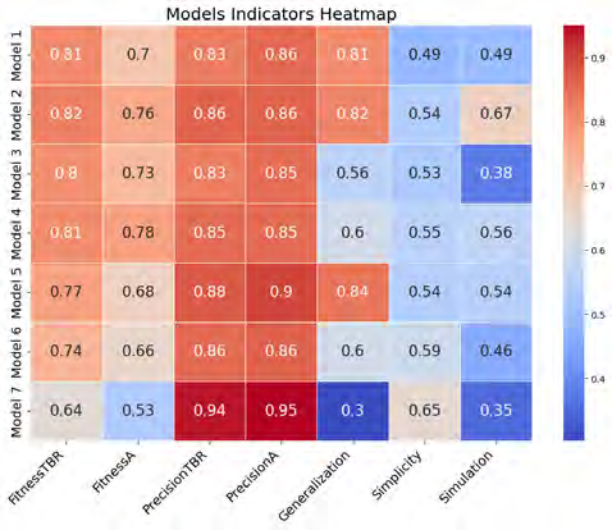

Figure 5: Indicators heatmap.

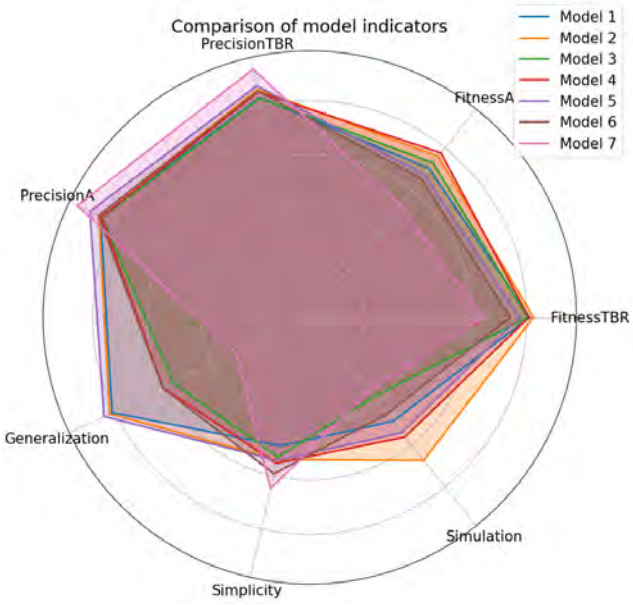

Figure 6: Indicators radial graphic.

5 DISCUSSION

As samples we took patients for a certain department in France, we took patients with COPD hospitalization (ICD code J44) for the year 2015 to see their treatment after the disease, i.e. their phase two, and another sample in 2022 to evaluate their paths before the hospitalization, i.e. phase one. In total we gathered 1118 patients where only 79 of them presented a spirometry before the hospitalization (model 3,4 and 6) and from those, only 14 patients performed oxygen therapy (model 7) after the first hospitalization.

In this study we did not study the population since the used database does not provide a complete information about the patients since it is 'economic-purposed' and only contains hospital or clinical data: it does not include data from private medical cabinets, adding some bias to the population and therefore to the models. This is shown in the datasets, we have 60 patients that only had, in 10 years, one hospitalization and we have the opposite, a person that has 1000 clinical activities, so this aggregates some errors either in the representation for the real population and the PMSI population and also in the simulation models. Is highly probable that patients we consider to have skipped phase one or three, actually may have not and have the rest of their medical history in a private medical cabinet.

From the graphic results in Figure 5, we can see that most of the models present a good level of fitness and precision above the 65%. In contrast, generalization shows a high value for those model in which we had all the 1.118 patients since in the rest of the models we only count with 79 or even 14 people, this could mean that the more samples we got the better the generalization. Furthermore, the simplicity indicator indeed does not present positive outcomes which can be visibly explained by the complexity of the disease as most of the models resemble to Figure 3.

Moreover, simulations indicators does not show better results due to two hypothesis, in one hand, the traces length explained above and in the other hand due to mathematics since there is no possible way to consider all possibilities given in the simulation model where at least we have twenty branches, i.e. decision-making stops with probabilities, that had from two until ten outputs and each with its probability. In addition, the datasets contain a very limited number of patients and exhibited significant diversity, which was insufficient to extract specific knowledge for the process mining models.

Finally, Figure 6 shows the model with the best simplicity and precision, which is model number 7. For generalization, model number 5 performs best, while model number 2 stands out in simulation and fitness through token-based replay. Lastly, model number 4 demonstrates the highest fitness with alignments.

6 CONCLUSION

Once we understood how the population-based responsibility worked and the results it gave, we tried to automatize the extraction and modeled of the clinical pathway for the population suffering certain diseases. To do this, firstly we needed to prove that the use of process mining is suitable for this application and also to show how to manage the challenges it opens. One of them was the data variability since each patients' clinical path is different. To overcome it, we applied different and more synthetic event logs. The filters went from deleting or modifying activities at the pre-processing level to filter paths in the process mining level in order to decrease the model complexity. Afterwards, with the simplest models we tested, with the use of indicators, how well they represented the event log.

We tested our methodology on COPD, with the help of healthcare professionals. We decided to separate the disease into four phases: COPD at risk, COPD diagnosed, then with hospitalization and finally performing oxygen therapy. Most of the models gave at least 65% fitness with a minimum precision of 83%, the simulation, with the mentioned bias in discussion, replicated artificial patients with an average similarity of 49% with the real patients. However, we do not discard any model since all of them contributes to provide distinct information, either with a global vision of the disease or an specific only disease model.

Future work includes generating our own algorithm for process mining discovery, using different indicators, then using the simulation part to test 'what if' scenarios.

ACKNOWLEDGMENTS

The authors would like to thank 'Fondation de l'Avenir' group for their support for this project and Mme Marion LELOUVIER, President of the management board. The Fondation de l'Avenir is a research foundation located in Paris and its objectives are clinical research and innovation in health, a bridge between fundamental medical research and all health stakeholders, understood in the broad and multidisciplinary sense: caregivers, doctors, surgeons, biologists, engineers, etc. in all establishments or at home and on all territories.

REFERENCES

Adriansyah, A., J. Munoz-Gama, J. Carmona, B. Dongen and W. Van der Aalst. 2014. "Measuring Precision of Modeled Behavior". *Information Systems and e-Business Management* 13 https://doi.org/10.1007/s10257-014-0234-7.

Anzueto, A. 2010. "Impact of Exacerbations on COPD". *European respiratory review : an official journal of the European Respiratory Society* 19:113–8 https://doi.org/10.1183/09059180.00002610.

Augusto, V., X. Xie, M. Prodel, B. Jouaneton and L. Lamarsalle. 2016. "Evaluation of Discovered Clinical Pathways Using Process Mining and Joint Agent-Based Discrete-Event Simulation". In *2016 Winter Simulation Conference (WSC)*, 2135–2146 https://doi.org/10.1109/WSC.2016.7822256.

Berti, A. and W. Van der Aalst. 2021. *A Novel Token-Based Replay Technique to Speed Up Conformance Checking and Process Enhancement*, 1–26 https://doi.org/10.1007/978-3-662-63079-2_1.

Blum, F. R. 2015. "Metrics in process discovery". Technical Report TR/DCC-2015-6, Computer Science Department, University of Chile.

Branson, R. 2018. "Oxygen Therapy in Copd". *Respiratory Care* 63:734–748 https://doi.org/10.4187/respcare.06312.

Buijs, J., B. Dongen, and W. Van der Aalst. 2014. "Quality Dimensions in Process Discovery: The Importance of Fitness, Precision, Generalization and Simplicity". *International Journal of Cooperative Information Systems* 23:1440001 https://doi.org/10.1142/S0218843014400012.

Fani Sani, M., J. J. González, S. van Zelst, and W. Van der Aalst. 2020. "Conformance Checking Approximation Using Simulation". 105–112 https://doi.org/10.1109/ICPM49681.2020.00025.

Gomez, S., S. Finkel, and A. Malone. 2020. "Des Territoires pour la Responsabilité Populationnelle : Utilisation du Programme de Médicalisation des Systèmes d'Information pour Définir des Territoires de Santé". *Revue d'Épidémiologie et de Santé Publique* 68:S55–S56 https://doi.org/10.1016/j.respe.2020.01.128.

Gomez, S., A. Malone, S. Finkel, and D. Laplanche. 2023. "Les Outils de la Responsabilité Populationnelle. Développement et Déploiement d'une Stratification Médico-Economique et Clinique de la Population Atteinte ou à Risque de Diabète de Type 2". *Revue d'Épidémiologie et de Santé Publique* 71:101486 https://doi.org/https://doi.org/10.1016/j.respe.2023.101486. Congrès national Emois 2023.

Günther, C. and A. Rozinat. 2012. "Disco: discover your processes". In *Proceedings of the Demonstration Track of the 10th International Conference on Business Process Management (BPM 2012)*, edited by N. Lohmann and S. Moser, CEUR Workshop Proceedings, 40–44: CEUR-WS.org. Demonstration Track of the 10th International Conference on Business Process Management, BPM Demos 2012 ; Conference date: 04-09-2012 Through 04-09-2012.

Koczulla, R., T. Schneeberger, I. Jarosch, K. Kenn and R. Gloeckl. 2018. "Long-Term Oxygen Therapy". *Deutsches Aerzteblatt Online* 115 https://doi.org/10.3238/arztebl.2018.0871.

Malone, A. 2023. *Tous Responsables de Notre Santé ! Bilan des 3 Premières Années de Déploiement de la Responsabilité Populationnelle*.

Munoz-Gama, J. and J. Carmona. 2010. "A Fresh Look at Precision in Process Conformance". Volume 6336, 211–226 https://doi.org/10.1007/978-3-642-15618-2_16.

Phan, R., V. Augusto, D. Martin, and M. Sarazin. 2019. "Clinical Pathway Analysis Using Process Mining and Discrete-Event Simulation: an Application to Incisional Hernia". In *2019 Winter Simulation Conference (WSC)*, 1172–1183 https://doi.org/10.1109/WSC40007.2019.9004944.

Rozinat, A. and W. Van der Aalst. 2008. "Conformance Checking of Processes Based on Monitoring Real Behavior". *Information Systems* 33:64–95 https://doi.org/10.1016/j.is.2007.07.001.

Sandelowsky, H. 2021. "COPD - Do The Right Thing" https://doi.org/.org/10.1186/s12875-021-01583-w.

Trottier, L. H. 2013. *La Responsabilité Populationnelle: Des Changements Organisationnels à Gérer en Réseau*. Institut national de santé publique du Québec.

Van der Aalst, W. 2011. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer Berlin, Heidelberg.

Van der Aalst, W. 2018. "Process Mining and Simulation: A Match Made in Heaven!" https://doi.org/10.22360/summersim.2018.scsc.005.

Van der Aalst, W., A. Adriansyah, and B. Dongen. 2012. "Replaying History on Process Models for Conformance Checking and Performance Analysis". *WIREs Data Mining and Knowledge Discovery* 2:182–192 https://doi.org/10.1002/widm.1045.

AUTHOR BIOGRAPHIES

ANA LUCIA TULA is currently a Ph.D student at the Center for Health Engineering at Mines Saint-Etienne, France. In 2023 she finished a double diploma program between Ecole des Mines Saint Etienne, France and Faculty of Exact, Physical and Natural Sciences from National University of Cordoba, Argentina receiving her masters diploma in biomedical engineering as well as her biomedical engineer diploma. Her email address analucia.tula@emse.fr.

VINCENT AUGUSTO Vincent Augusto received his Ph.D. degree from the Ecole Nationale Superieure des Mines de Saint-Étienne (EMSE), France, in 2008 and his Habilitation à Diriger des Recherches degree from the Jean Monnet University, in 2016. Currently, he is a professor of industrial engineering in the Center for Health Engineering and in the IEOR team of CNRS UMR 6158 LIMOS, EMSE. His research interests include modeling, simulation, optimization of health care systems and their supply chains. His e-mail and web addresses are augusto@emse.fr and http:/www.emse.fr/ augusto, respectively.

XAVIER BOUCHER is a professor in Industrial Management at the Ecole Nationale Supérieure des Mines de Saint Etienne (France), he is currently Research Director for FAYOL Research Center (interdisciplinary Research Center, developing activities for the Global Performance of Industrial Companies and Territories). and an active member of several scientific societies in the field of Industrial Engineering (IFIP, CIRP IPSS2, SOCOLNET). His email adress is boucher@emse.fr

MARIANNE SARAZIN is a public health doctor with a doctorate in Life Sciences from Mines Saint-Etienne, France. She is the head of the Medical Information Department of the Mutualiste Sanitary Group of Saint-Etienne and a collaborator in the Center for Health Care Engineering at Mines Saint-Etienne. She is also with the UMRS 1136 Inserm lab, which specialized in the modelling of epidemics. Her email address is marianne.sarazin@iplesp.upmc.fr.