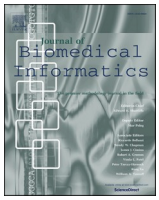
Journal of Biomedical Informatics 128 (2022) 104033

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



Original Research

A framework for the automatic description of healthcare processes in natural language: Application in an aortic stenosis integrated care process☆

Yago Fontenla-Seco [a,*], Manuel Lama [a], Violeta González-Salvado [b], Carlos Peña-Gil [b], Alberto Bugarín-Diz [a]

[a] Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[b] Department of Cardiology, University Clinical Hospital of Santiago de Compostela, SERGAS IDIS CIBERCV, Santiago de Compostela, Spain

ARTICLE INFO

Keywords:
Healthcare processes
Process mining
Process understanding
Natural language generation
Data to text systems
Fuzzy linguistic terms

ABSTRACT

In this paper, we propose a framework for the automatic generation of natural language descriptions of healthcare processes using quantitative and qualitative data and medical expert knowledge. Inspired by the demand of novel ways of conveying process mining analysis results of healthcare processes (Rojas et al., 2016), our framework is based on the most widely used Data-To-Text (D2T) pipeline (Reiter, 2007) and on the usage of process mining techniques. Backed by a general model that handles process data, this framework is able to quantify attributes in time during a process life-span, recall temporal relations and waiting times between events and its possible causes and compare case (patient) attributes between groups, among other features. Through integrating fuzzy quantification techniques, our framework is able to represent relevant quantitative process information with some degree of uncertainty present on it and describe it in natural language involving uncertain terms. A real application over the Aortic Stenosis Integrated Care Process of the University Hospital of Santiago de Compostela is presented, showcasing the potential of our framework for providing natural language descriptions of healthcare processes addressed to medical experts. Following the standards of D2T systems, manual human validation was conducted for the generated natural language descriptions by fifteen medical experts in Cardiology. Validation results are very positive, since a global average of 4.07/5.00 was achieved for questions related to understandability, usefulness and impact of the natural language descriptions on the medical experts work. More precisely, results indicate i) that the modality which conveyed the information most efficiently was natural language ii) a very clear preference of texts over the usual graphic representation of process information as the way for conveying information to experts (4.28/5.00), and iii) natural language descriptions provide relevant and useful information about the process, allowing for its improvement.

1. Introduction

Processes allow organizations to represent and structure the activities that take place within them and their information systems as well as how data and resources are managed.

In healthcare organizations, processes, commonly referred to as healthcare processes [3], structure and organize clinical and non-clinical activities aimed to diagnose, treat and prevent diseases on patient health [4]. In recent years, as a response to an increasing pressure in improving medical and organizational efficiency and effectiveness, the need for enhancing processes has risen in healthcare organizations worldwide [4,5].

Process mining is a data-driven, process-centric approach, whose aim is to exploit recorded event data. By automatically discovering the underlying process model from an event log, it can extract valuable process-related information that can be used to provide insights, determine performance and detect and identify bottlenecks, which helps to understand and improve processes [6]. Due to the nature of

☆ This research was funded by the Spanish Ministry for Science, Innovation and Universities (grants TIN2017-84796-C2-1-R, PID2020-112623GB-I00, and PDC2021-121072-C21) and the Galician Ministry of Education, University and Professional Training (grants ED431C2018/29 and ED431G2019/04). All grants were co-funded by the European Regional Development Fund (ERDF/FEDER program).

* Corresponding author.
E-mail addresses: yago.fontenla.seco@usc.es (Y. Fontenla-Seco), manuel.lama@usc.es (M. Lama), violeta.gonzalez.salvado@sergas.es (V. González-Salvado), carlos.pena.gil@sergas.es (C. Peña-Gil), alberto.bugarin.diz@usc.es (A. Bugarín-Diz).

https://doi.org/10.1016/j.jbi.2022.104033
Received 1 September 2021; Received in revised form 24 January 2022; Accepted 16 February 2022
Available online 23 February 2022

healthcare processes (highly dynamic, complex, ad-hoc and increasingly multi-disciplinary [7]), processes with a high number of distinct activities and relationships among them are the norm. Commonly called *spaghetti processes*, they often derive in highly complex process mining analysis results [6]. These results are referred to be quite difficult to understand by medical experts, the final stakeholders of the analysis, whom, in general, are not specialized in process mining nor visual analysis techniques (the common way of conveying process mining analysis results) [1,4,8] and who should invest their efforts in improving healthcare processes and not in "deciphering" process mining analysis results. Moreover, authors highlight the lack of good visualization techniques for process models and process mining analysis results in the state-of-the-art, especially, in dynamic, complex and less-structured processes such as those present in the healthcare domain. Emphasizing the need for improved visualization techniques and visual analytics to facilitate the interpretation of process mining analysis results [1]. This raises the need of proposing novel and different ways of conveying the results of healthcare process mining analysis to users in a clear, direct and comprehensible way, adapted to the real needs of medical experts both in content and form [9].

Natural Language Generation (NLG) techniques aim to provide users with natural language texts that summarize the most relevant features of some data in a way that can be easily consumed [10]. As natural language is the inherent way of communicating for humans, NLG techniques do not rely on users capabilities to extract relevant information and patterns from visual analytics. Furthermore, the use of uncertain terms and expressions, common in natural language, is very effective for the summarization and communication of data.

Added to the lack of good visualization techniques for process mining analysis results, research suggests [11] that in some domains, expert knowledge is also a complementary requirement to visual analysis techniques for users to fully understand graphical information. In particular, it has been proved that in the healthcare domain, medical experts can take better decisions when presented together with textual summaries rather than when presented with graphical displays [12]. In this sense, NLG techniques seem like a good approach to enhance the understanding of healthcare processes and its analytics for medical experts. Different methods for generating insights on data through natural language have already been applied both on healthcare [13,14] and process data [15,16]. However, healthcare processes have been neglected in the NLG state-of-the-art. This highlights the need for a new approach focused on the description of healthcare process mining analysis results in natural language.

In this paper, we present a process mining-based framework for the generation of qualitative and quantitative natural language descriptions of healthcare processes. Our proposed framework is based on the most widely used Data-To-Text (D2T) pipeline [2]. It integrates process mining techniques for the extraction of the relevant features of a process with fuzzy logic techniques as a way to handle the inherent imprecision of natural language through the modeling of uncertain expressions. Parting from event log data (recorded while a process was executed in a real-life scenario), the framework is able to automatically extract the underlying process model with the use of discovery algorithms. Through the replay of the event log over the discovered process model, information about the time perspective is gathered and added to the existing control-flow, case and resource perspectives extracted from the event log, giving place to an enriched event log (which contains the original features and the new extracted ones), to which data-driven and LDD techniques can be applied in order to generate natural language descriptions. Extracting the underlying process model from an event log and replaying it for obtaining an enriched model and event log are the cornerstone of our approach. This way, the proposed framework is able to encompass all perspectives of a process, unlike previous state-of-the-art approaches. The generation of the natural language descriptions is backed by a general model that handles both qualitatively and quantitatively process data: events and its attributes, cases (patients when talking about healthcare processes) and its attributes, resources involved in the process, temporal relations and wait times between events, etc. These descriptions are given to the user as sentences regarding each of the aspects of the process highlighted as relevant. They are constructed with the use of fuzzy logic techniques, where fuzzy terms can be used in order to generate descriptions that are able to summarize information in easy and conveniently abstracted ways e.g. *"The wait time between the heart team meeting of a patient and its intervention is expected to be lower than 30 days. However, in approximately half of cases (59.30%) the patient had a wait time between its heart team meeting and its intervention higher than 30 days"*. The generality of the model means it can be easily applied to other processes different to the one here presented (even to processes from domains other than healthcare), where attention may be drawn to other elements of the process or with other language requirements.

This framework is applied to the Aortic Stenosis Integrated Care Process (AS ICP) implemented in the Cardiology Department of the University Hospital of Santiago de Compostela. In this process, consultations and medical examination (x-rays, echocardiograms, Computed Tomography (CT) scans, etc.) are performed to patients with AS in order to decide the procedure they will undergo. Once decided, patients are intervened and followed up during their recovery and discharge from the process. Through multiple meetings with medical experts, we clarified their needs and the elements of healthcare processes they need a better understanding of in order to find the most improvement opportunities and generate descriptions that help answer the most commonly asked questions when analyzing healthcare processes [1,4,17]. This way, based on the elements of healthcare processes medical experts show most interest on, particularly on the healthcare process here presented as a use case (AS ICP), our proposed model is able to quantify case (patient) attributes in time during a process life-span (e.g. number of patients with a particular characteristic such as type of intervention a patient underwent), recall temporal relations between activities in the process extracted through process mining techniques (e.g. order and wait time between the execution of two complementary medical tests), and describe differences between care paths followed by different groups of patients (e.g. difference in care paths followed by patients with ambulatory admittance vs. emergency admittance), among other features. To the best of our knowledge, our proposal is the first one which integrates process mining, D2T and fuzzy logic techniques for the description of processes in natural language involving terms with uncertainty, with an application in a healthcare process.

The following sections are structured as follows. Section 2 gives a background in the application of NLG, Fuzzy Logic and Process Mining techniques to healthcare and process data as well as a brief analysis on the current proposals of natural language descriptions of processes. Section 3 presents the Aortic Stenosis Integrated Care Process and the problem of healthcare process understanding. Section 5 introduces the proposed pipeline of the framework and explains all the processing done from the input of the data to the generation of the final descriptions. Section 6 contains the proposed descriptions and how they were generated for the Aortic Stenosis case study. Second to last, Section 7 shows the results of the validation conducted with the medical experts in charge of the AS process. Finally, Section 8 presents some concluding remarks.

2. Background and related work

2.1. Process mining

Processes allow organizations to represent and structure the activities that take place within them and their information systems as well as how data and resources are managed. They are usually represented

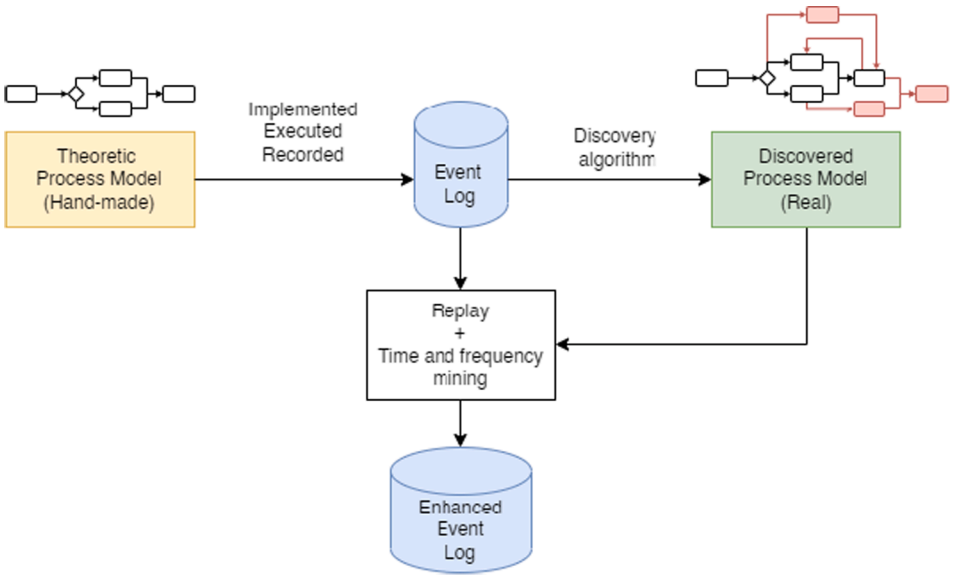


Fig. 1. Schema of the relationship between a hand-defined theoretic process model, its execution and recording, resulting in an event log and the application of discovery algorithms to discover the underlying process model. Further, replaying employs both the event log and the discovered model.

graphically as process models (in a plethora of notations), and data about their execution is recorded in what are called event logs. Process mining goal is to exploit this recorded event data in a way that can be used to understand what is really happening in a process in order to anticipate problems, streamline and improve processes by conjugating classical business process model analysis and data mining techniques. "The idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in today's systems."[6]. More specifically, process mining techniques try to provide support to a process life-cycle, in which a theoretic process model is originally designed (often an idealized version of the process), implemented, executed and monitored (while the process is executed event data is recorded in event logs), diagnosed and re-designed. However, in most organizations, the diagnosis phase is not systematically implemented, resulting in performances lower that initially planned, as many differences between the original theoretic (and idealized) process model and its execution, arise: failures in design, exceptions, changes in the process and more, trigger (unplanned) behaviours that where not originally embodied in the initial process model. This is where process mining shines, automatically discovering[1] [6] the underlying (and real as per the execution data) process model, to extract with it valuable, process related information in a meaningful way, allowing for the evaluation (comparison with the theoretic process model, reviewing of quality values defined for key performance indicators, etc.) and enhancement (re-designing the process control-flow, adding resources where the evaluation of the process indicate is needed, etc.) of processes. Fig. 1 shows the relationship between an initial theoretically defined process model, its execution and monitoring and the application of discovery algorithms (process mining) in order to extract the real process model from the event log resulting from the execution of the process.

An event log consists in a recording of executed activities α (being an activity each well-defined step in a process) that takes the form of a multi-set of cases. Being a case c a particular execution of the process and a trace $\widehat{c}$ the ordered sequence of events that characterizes it. An event e represents the execution of an activity α in a particular instant and it is defined by two mandatory attributes: the name of the executed activity α and the time of its execution (timestamp). However, events can have additional attributes such as its associated resources (e.g. medical expert in charge of the event), duration, etc. As events, cases have attributes, compulsorily its corresponding trace and an identifier. Other attributes may be its throughput time, the patient involved in the case, the origin of admission of the patient involved in the case, or in a process that takes place outside of the healthcare domain it could be the cost associated to the execution of that particular instance of the process (case), among many others.

Process mining has been applied multiple times in the healthcare domain: from discussing its applicability and challenges [5,17], to defining methodologies and reference models to increase the quality of healthcare process data and the application of process mining analysis to it [3,4]. In the last 5 years, two literature reviews have been published [1,8], showing that the application of process mining analysis in healthcare is a fast-paced evolving field.

2.2. Natural language generation

NLG systems are aimed to produce natural language texts from some underlying non-linguistic representation of information [10]. These systems use domain knowledge as well as linguistic knowledge in order to automatically generate texts or documents that summarize the most relevant aspects of some input data. They have been applied in multiple domains since their inception: from environmental information systems (weather forecasting reports [18]) to the health and related realms (textual summaries from neonatal Intensive Care Unit data [14], systems to support interaction of kids with complex communication needs [19]). Within the NLG field, D2T concerns to systems which generate text from numeric input data.

Previous to NLG and D2T systems, using knowledge from the Fuzzy Logic domain, the paradigms of Computing with Words [20] and Linguistic Descriptions of Data (LDD) [21] emerged with the aim of modelling and managing the inherent uncertainty present in natural language through the use of fuzzy sets. These paradigms are based on the fuzzy sets theory and the use of protoforms [22], predefined structures

[1] By applying a discovery algorithm i.e. a function capable of mapping an event log onto a process model such that the model is representative of the behaviour recorded on the event log.

Table 1
Type of descriptions (from a series of elements of processes the state-of-the-art highlights) each approach is able to generate.

Approach	Frequency of events	Quantify event/case attributes	Quantify relations between attributes	Temporal contextualization of attributes	Frequency and wait times of relations between events	Compliance with expected values	Comparison between groups of cases	Ordering of activities
[16,26]	✓	✓	✓	✗	✗	✗	✗	✗
Our approach	✓	✓	✓	✓	✓	✓	✓	✓

(or templates) used to generate textual summaries (fuzzy quantified statements) involving linguistic terms with some degree of uncertainty present on them.

Even having different origins and degrees of complexity, NLG and Fuzzy Logic techniques can be used together in order to develop more intelligent systems: LDD techniques provide ways of handling uncertainty and determining the most relevant characteristics of some non-linguistic data (the first stage of NLG and D2T systems), and NLG defines a pipeline able to generate well written, understandable, full natural language texts with rich semantics [23]. This way, LDD techniques are valid on the framework of a more complex NLG system, as they can be used in the context of determining an intermediate language representation in the content determination phase, aiding on the selection and summarization of the most relevant aspects of the non-linguistic input data to the system.

2.3. Related work

Different natural language generation techniques have been applied in healthcare data and process data. However, no approach has yet been proposed for the generation of natural language descriptions of healthcare processes. In the healthcare domain, multiple systems have been developed based on NLG techniques, one of the most well known is the NLG system Baby-Talk [14], developed for the automatic generation of textual summaries from neonatal intensive care data. Following the LDD approach, Almeida et al. [13] propose a system able to generate linguistic summaries of time series data for septic shock patients.

Regarding the description of processes in natural language, two different approaches can be found, each, pairing a natural language generation technique with a particular process perspective[2]: NLG and the control-flow perspective and LDD and the case perspective. Following a NLG approach [15,24,25] propose a pipeline for the description of process models (control-flow perspective) in natural language as a way of maintaining a stable representation of a process model during its life-cycle. They aim to provide ways of supporting process model validation and inconsistency detection through natural language descriptions of process models. However, by focusing only on process models and ignoring the real execution data captured on event logs, this model-centric technique cannot respond to process-related questions further than describing the structure of a process via its model. As no process discovery techniques are applied, this approach is capable of describing the theoretical ordering of activities of a process via the structure of its model and the information contained in its labeled elements (activity and gateway labels of the model). Thus, this approach is very useful when dealing with structured process models, but falls short when addressing highly unstructured processes like the ones in the healthcare domain. Healthcare processes are highly dynamic, complex, ad-hoc and increasingly multi-disciplinary. Typically, many discrepancies are found between the theoretic and the actual running processes [3,5]. Furthermore, they usually derive in spaghetti processes, with a high number of distinct activities and relationships. The description of the model of these processes is unmanageable or does not contribute to a better understanding of the process execution. Specially, medical experts show no interest in getting descriptions of theoretic process models, but of the results of its execution. In spite of this, no comparison is possible between this approach and the one presented in this paper, as they target different types of processes and pursue different objectives.

Opposite to classical business process-model analysis, data-driven techniques focus on exploiting recorded data by extracting statistics, association rules, etc. from data stored in tables. In this sense, the LDD and case-focused approach relies on the application of data mining techniques to event logs [16,26], again, without paying attention to the alignment between process model and reality, but now focusing on event log data, instead. By applying data-driven and LDD techniques to event logs, knowledge can be extracted from the process execution data and natural language descriptions can be generated. However, by focusing only on event data, the scope of the process analysis is limited, as both the control-flow and time and frequency perspectives are neglected. In this approach, only resource and case perspectives can be described due to the lack of usage of any type of process model (not hand-made nor discovered from the log). As for describing the control-flow and timing perspectives, a process model is needed in order to infer relations between activities and their wait times by replaying[3] the recorded event log over its corresponding process model. Data-driven and LDD techniques are able to recall basic case statistics, but cannot be used to analyze bottlenecks, wait times, compare behavior among groups of cases or deviations in the process flow, etc. So this approach is not able to answer frequent questions that arise when dealing with complex processes as those found in the healthcare domain.

We can conclude that existing approaches are not able to provide a complete view of what is happening in a process, as the former generates natural language descriptions of process models, ignoring the execution data, and the later generates linguistic descriptions of case and resource data from event logs, ignoring the underlying process model. Aiming to fill the gap that exists in the application of NLG techniques to processes state-of-the-art, we propose a framework for the generation of qualitative and quantitative natural language descriptions of processes. Table 1 summarizes which natural language descriptions (from the elements of processes the state-of-the-art highlights as relevant) each of the approaches is able to generate.

3. Healthcare processes: the aortic stenosis integrated care process

In this proposal we will focus on cardiology healthcare processes, more specifically on the Aortic Stenosis Integrated Care Process of the Cardiology Department of the University Hospital of Santiago de Compostela. As heart diseases (in particular ischemic cardiopathy) are still the most frequent cause of death worldwide [27], analysis of cardiology

[2] The different "dimensions" in which a process can be observed or analyzed: the control-flow perspective focuses on the ordering of activities, the case perspective focuses on properties (attributes) of cases, the resource perspective focuses on information about resources (actors) involved in the process and the time perspective is concerned with the timing and frequency of events.

[3] Process replay establishes a strong relation between a process model and the event log the process model is discovered from by relating events in the log to activities in the process. It will be further discussed in the next sections.

healthcare processes is of great interest.

Aortic stenosis (AS) is a chronic disease and the most prevalent valvular heart disease, especially in elderly patients [28]. In the absence of treatments, valve replacement is currently the only intervention with the potential to improve the prognosis of AS patients [29]. However, when intervention is indicated, it is generally not performed immediately. the paradigm of care for the patient with AS involves different professionals who at different times and places provide care that is fragmented, with little continuity of care and a high risk of lack of coordination. In contrast, Integrated Care Processes (ICP) seek to ensure the effectiveness of clinical actions through greater coordination and a guarantee of continuity of care [30]. In January 2018, an ICP of AS was implemented in the Valvulopathy Service of the Cardiology Department of the University Hospital of Santiago de Compostela, with the aim of managing the care of patients requiring intervention and guaranteeing the correct and coordinated functioning of each one of its phases[31].

3.1. Process definition

The AS ICP covers all diagnostic and therapeutic interventions, from the heart team meeting[4] (formal inclusion of the patient in the process) to the patient return to normal activity (patient exit from the process). The following stages are distinguished:

1. **Stage 0**: Prior to the inclusion of the patient in the process, all the necessary tests will be requested and evaluated for presentation at the heart team meeting with the objective of identifying severe AS patients amenable of intervention.
2. **Stage 1**: From the decision to intervene in the heart team meeting to the valve replacement procedure.
3. **Stage 2**: Surgical or percutaneous valve replacement procedure.
4. **Stage 3**: Rehabilitation and follow-up after intervention in order to facilitate the patient's recovery after surgery.

In order to monitor the results of the AS ICP, a registry of clinical data of the patients included in the process has been established. Information about the execution of all diagnostic and therapeutic interventions as well as patient management activities are recorded. Monitoring of unexpected events (e.g. mortality, unscheduled admission due to cardiovascular causes, emergency room attendance due to cardiovascular causes) is available in the registry as well.

3.2. Problem definition

Process and outcome quality indicators are defined to analyze safety, identify funnels, bottlenecks and futile actions, and implement improvements. The aim is to reduce unexpected events and unscheduled admissions, reduce overall delays in patient care, improve adherence to scientific recommendations and care protocol, and implement continuous quality improvement initiatives.

Multiple key process indicators (KPIs) are defined to evaluate the results of the AS ICP. Medical experts in the Valvulopathy Service show genuine interest in applying process mining techniques to extract valuable knowledge and relationships between the different KPIs defined. The KPIs of interest for the AS ICP, and that will be described in this work are the following: proportion of patients with emergency admittance, compliance with the standard of wait time between the heart team meeting and the intervention of a patient, relation between the type of intervention and the wait time for a patient, delay between the heart team meeting and the intervention caused by CT scanning of a patient and influence of the type of admittance in the wait time.

Due to the complexity of the process and the numerous variables and scopes defined for its evaluation, a problem arises when conveying the process mining results to the medical experts in the Valvulopathy service, as severe difficulties arise in its comprehension. The lack of good visualization techniques for process mining analysis results in the state-of-the-art, emphasizes the need of proposing novel ways of conveying the results of process mining analysis to users in a clear, and comprehensible way. At this stage, medical experts show interest in natural language descriptions of the AS ICP and its process mining analysis results, which can help in the understanding the process and the implementation of improvements. As the elements under description are common to any other healthcare process, and to any process outside the healthcare domain: relating case attributes to KPIs, analyzing the control-flow perspective (paths followed by cases in the process), getting insights about frequency and timing of events, reducing unexpected or undesired behavior (reduce unexpected events), reduce wait times (overall delays in cases), among may others. We propose a general model, able to encompass all these process elements, and therefore applicable to processes in any domain.

4. Fuzzy quantified statements

Fuzzy quantified statements can be used to extract relevant information from an input dataset and a knowledge base for summarizing knowledge about variables and their values in a given domain. These statements follow predefined formal structures or templates, that are referred to as protoforms. Fuzzy quantified sentences of type-I and type-II have been the most used ones in the literature since their inception in early 1980's. Formally, a *protoform* [22] is an abbreviation of prototypical form, an abstracted prototype (model) of a quantified proposition, which is composed of the following four elements:

- Referential X: A set of objects for which certain property or set of properties holds. For example, the set of cases from an event log.
- Summarizer A: Used to indicate some property or aggregation of properties of the referential of interest. For instance, if the referential is the set of cases on an event log, the summarizer can be any of the attributes (or aggregation of the attributes) of the cases.
- Quantifier Q: Used to express the quantity or proportion of data from the referential which fulfills the properties indicated by the summarizer (e.g. *most*).
- Truth degree T: A numerical value in the [0, 1] range, which indicates the degree of fulfillment or validity of the protoform for a given case.

This way, quantified statements like "In most cases, activity α_1 was executed" can be instantiated from an event log with type-I protoform:

$$Q\ X's\ are\ A \tag{1}$$

where quantifier Q is *most*, summarizer A is *activity* α_1 *was executed* and referential X is the set of cases from an event log. These protoforms can be used for modelling processes in any domain, as the referential can represent any set of objects, and the summarizer can make reference to any property that holds for said set of objects.

A qualifier can be added to the description to narrow the scope of the sentence, giving place to type-II protoforms:

- Qualifier B: As the summarizer, can make reference to any property or aggregation of properties of the referential. It defines a subset of

[4] The heart team meeting is a multidisciplinary clinical session in which cardiac surgeons, vascular surgeons, clinical cardiologists, and cardiac imaging experts evaluate patients with heart disease amenable to surgical treatment.

elements of the referential which fulfills the property or properties defined by B to a certain degree. This subset is taken as the referential that will be evaluated against summarizer and quantifier.

Statements like "In most cases, where activity α_2 was executed, activity α_1 was executed" can be instantiated from a type-II protoform:

$$Q\ BX's\ are\ A \tag{2}$$

Where Q, X and A are as before, and qualifier B is *activity* α_2 *was executed*. As it can be seen, *activity* α_2 *was executed* acts as a discriminant and *activity* α_1 *was executed* as the property to which the subset is evaluated against.

Instancing a protoform involves assigning values to its elements and computing its truth degree $T \in [0, 1]$, i.e. the degree to which the protoform correctly summarizes the referential. The closer it is to one the more truthful the summary is. In our approach, Zadeh's quantification model is used for computing the truth degrees [20], although any valid quantification model could be used on its behalf [32,33].

4.1. Linguistic variables

Both summarizer, qualifier and quantifier in the protoforms are usually defined as linguistic variables. Linguistic variables model the partitioning of the domain of a numeric or categorical variable into several properties. Each property is known as a linguistic value and is associated to a membership function that measures the degree in which different values of the original variable fulfill that property. A membership function $\mu_A(x)$ of linguistic value A, associates with each element x of a referential X a real number in the interval [0, 1]; with the value of $\mu_A(x)$ at x representing the "degree of membership" of x in A i.e. the degree to which x satisfies the property indicated by A. These membership functions are usually represented trapezoid functions $T[a,b,c,d]$:, defined as follows:

$$\mu_{T[a,b,c,d]}(x) = \begin{cases} 0, & (x \leqslant a)\ or\ (x > d) \\ \dfrac{x-a}{b-a} & a < x \leqslant b \\ 1, & b < x \leqslant c \\ \dfrac{d-x}{d-c} & c < x \leqslant d \end{cases} \tag{3}$$

For example, variable *waiting time* originally defined as a numerical variable ranging the interval [0, 100] can be modeled as a linguistic variable with linguistic values {*very short, short, as expected, longer than expected, really long*} over the domain [0, 100] of the original numerical variable, as Fig. 2 shows. A set of proportional quantifiers (relating a proportion over the domain [0,1]) can be defined as a series of positive, non-monotonous fuzzy sets (trapezoids) as in Fig. 3.

The use of linguistic variables allows to capture and manage the uncertainty and vagueness present inherently in human natural language and use it in the description of variables including along the time and space dimensions. In cases where more fine grained information is needed (or the original variable is categorical), crisp linguistic values can be used. The membership of an element x to a crisp linguistic value can be computed straightforward by taking $a = b$ and $c = d$ in expression (3). Resulting in an interval with crisp (instead of fuzzy) limits, the membership of an element x to these functions can only be {0,1}. When the original variable is categorical, then, a linguistic value $A := [a, d]$ is defined for each category and its membership function is defined as:

$$\mu_{[a,d]}(x) = \begin{cases} 1, & \text{if } x \in [a, d] \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

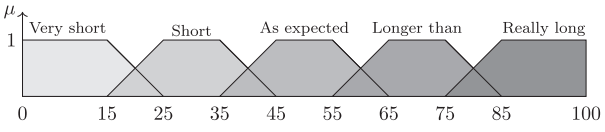


Fig. 2. Linguistic values defined for linguistic variable "waiting time" on the domain [0, 100].

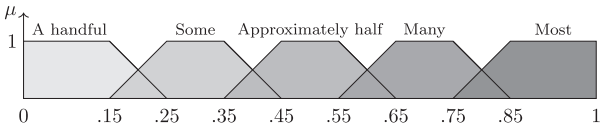


Fig. 3. Proportional quantifiers defined on domain [0, 1].

4.2. Generation process

The generation of protoforms is based on its instantiation and the evaluation of its truth degree. Once the referential and combination of summarizer, qualifier and quantifier are chosen, the first step is taking the set of values of the properties of the referential the summarizer makes reference to, i.e. taking from the dataset subject to description the column (or columns) corresponding to the variable (or variables) the summarizer makes reference to. For example, the dataset object to description is an event log, and the referential is defined as the set of cases from the log. If the summarizer relates *short waiting time*, then, the set of values of the referential that must be taken for evaluation i.e. the column of the dataset (the event log in this case) will be *waiting time*. This referential can be seen in Table 6 in column 9 and an abstraction of it in Fig. 4a, where y_i represents each of the elements on the selected column of the dataset.

The second step is evaluating this set of values over the chosen summarizer. This is done by evaluating each element of the set with the membership function of the summarizer e.g. linguistic value *short* for linguistic variable *waiting time* with its corresponding membership function defined as in Fig. 4b. This evaluation gives place to a collection of membership values on the domain [0,1], each of them indicating the membership of each of the elements of the referential to the selected linguistic value. The collection of membership values resulting from evaluating the referential Fig. 4a over the linguistic value Fig. 4b can be seen in Fig. 4c. Each y_i represents the membership value of each of the elements taken from the referential to the linguistic value.

Finally, the membership degree of the referential to the summarizer must be evaluated against the quantifier in use. To do this, the collection of membership values obtained in the previous step is aggregated. When using a proportional quantifier, a proportion of membership is computed. Then, this aggregated value is evaluated against the membership function of the quantifier and a truth value T for the protoform is obtained. Eq. (5) summarizes the complete evaluation process:

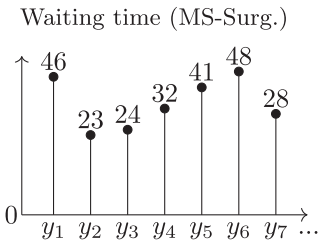
$$T = \mu_Q\left(\frac{1}{n}\sum_{i=1}^{n}\mu_A(p_i)\right) \tag{5}$$

where p_i is each of the elements in the referential, n is the number of elements in the referential and μ_A the membership function of summarizer A.

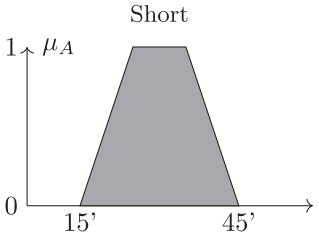
For type-II protoforms, where a qualifier is used, an additional step is necessary. This is, previous to the evaluation of the referential against the summarizer, the referential is evaluated against the membership function of the qualifier, effectively acting as a discriminant over the elements of the referential that will be evaluated against the summarizer. The resulting equation for the evaluation process is then:

$$T = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_A(p_i) \wedge \mu_B(p_i)}{\sum_{i=1}^{n}\mu_B(p_i)}\right) \tag{6}$$

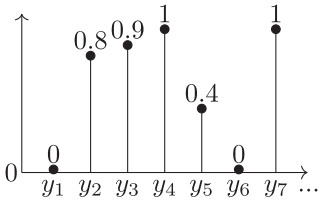
Instead of single variables, complex expressions involving, for instance, several variables, values and relationships among them can be used as



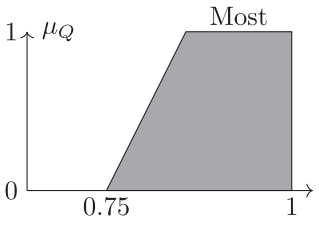
(a) Referential of *waiting times between the heart team meeting and intervention.*



(b) Summarizer "short".



(c) Membership of elements of the referential "waiting times" to summarizer "short".



(d) Quantifier "Most".

Fig. 4. Elements that constitute the protoform evaluation process.

summarizers, qualifiers or quantifiers, thus allowing flexibility and expressiveness of the protoforms. This, added to the capability of these protoforms to be applied in any domain, gives place to rich and insightful descriptions of processes. For instance, for the AS ICP, the following description can be generated *"Wait time between the heart team meeting of a patient and its intervention is expected to be lower than 30 days. However, in approximately half of cases (59.30%) patient had a wait time between its heart team meeting and its intervention higher than 30 days."* In Section 5.4 a series of extensions of the type-I and type-II protoforms for the description of processes are presented, and its instantiation over the AS ICP can be found in Section 6.

In classical LDD approaches, linguistic summaries are generated by a search (exhaustive or non-exhaustive) through the semantic space[5] guided by quality measures as the truth value. By contrast, in the D2T and NLG systems pipeline (as this proposal) expert knowledge, usually in the form of sets of rules, is used in the main stages of the generation process related to handling of data (data interpretation and document planning) to determine which messages (protoforms in our case) must be included and realized into the final text.

5. Pipeline for the generation of natural language descriptions of processes

In this section, the Process-To-Text (P2T) framework for the automatic generation of natural language descriptions of healthcare processes is presented. This framework is based on the most widely used architecture for D2T [6] systems [2], and is composed of four stages: *preprocessing*, *data interpretation*, *document planning*, and *linguistic*

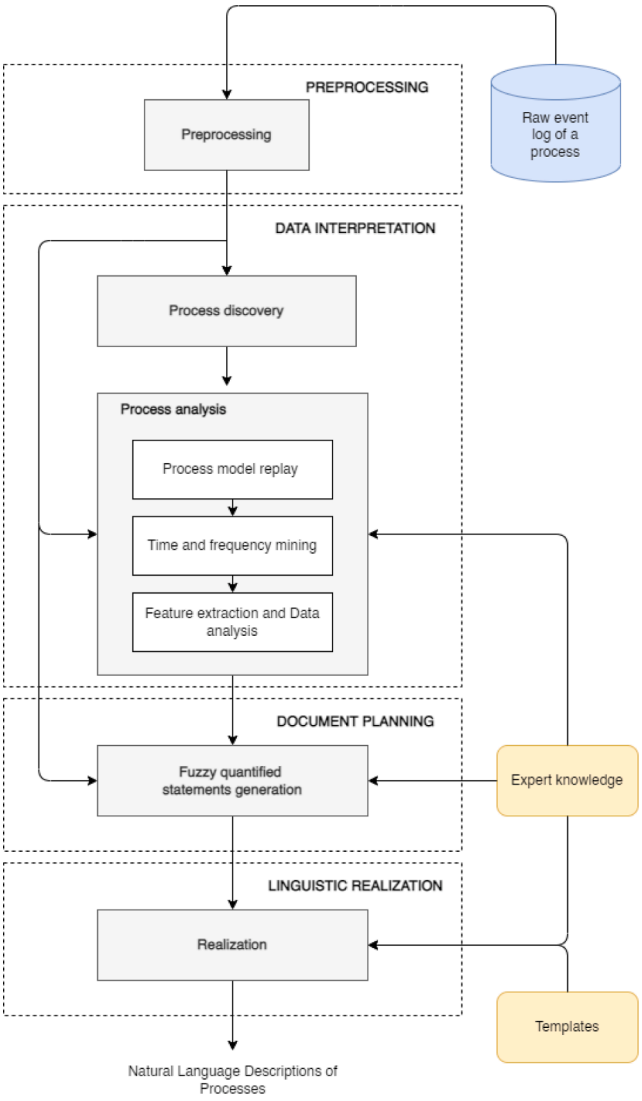


Fig. 5. The Process-to-Text pipeline for the automatic generation of natural language descriptions of processes. First the original data registry is converted into an event log. In the data interpretation phase the process model is discovered and process analysis and extraction of relevant features of the process take place. In the document planning stage protoforms are generated. Finally, protoforms are realized into natural language texts.

[5] The semantic space is the power set of all protoform instances that can be built using the defined quantifiers, qualifiers and summarizers

[6] Data-to-Text is a sub-discipline of the Natural Language Generation field aimed at generating texts from numerical input data.

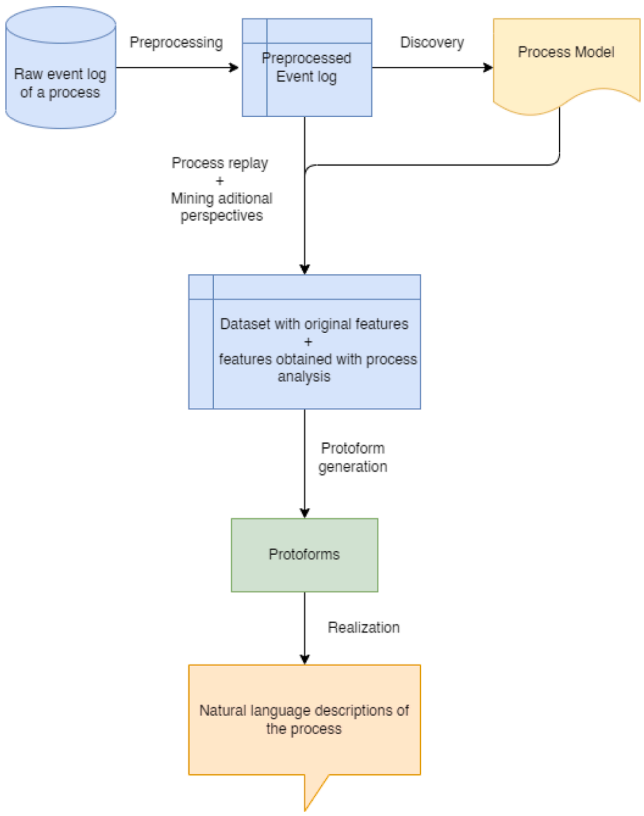


Fig. 6. Data transformations that take place within the Process-to-Text pipeline. The original raw log is converted into a preprocessed event log, after applying a discovery algorithm a process model is obtained. With the process model and the event log, process replay and mining of additional perspectives can happen, and with feature extraction techniques, a dataset for applying data mining and fuzzy quantification is obtained. Then, protoforms are generated from this dataset. Finally, the protoforms are realized into natural language text.

realization. In the preprocessing stage, the raw event log data is converted into an event log readily available for applying process mining techniques to it. Next, the data interpretation phase consists in the process model discovery, process analysis (process replay and mining of additional perspectives), and extraction of relevant features of the process that might be of interest. Second to last, the document planning stage consists in the generation of fuzzy quantified statements that summarize this information. Finally, in the linguistic realization phase, the fuzzy quantified statements are transformed into natural language texts using a hybrid template based approach and a natural language realization engine [34].

Fig. 5 illustrates the pipeline from start to end and Fig. 6 illustrates the transformations of data that take place in the pipeline, helping understand the different stages the pipeline is comprised of and what data processing takes place in each stage. In the following these stages will be described in detail.

5.1. Preprocessing

As being really ad-hoc, processes (from any domain) are recorded in many different ways by the information systems in which they are implemented, so no standard way of extracting information from processes exists. However, process mining defines a model on how process elements relate to one another, this way, parting from the raw data, and based on this model, it is possible to generate event logs to which process mining techniques can be applied. The input to this stage is the original process log. Manipulation by a process expert is required. We will use the AS ICP as an illustrative example, as commonly some preprocessing is needed in order to apply process mining techniques to healthcare data [1]. An outline of the original registry of clinical data from the AS ICP is shown in Table 2 (only some columns are shown, as the original number of variables is too large for including all). As it can be seen it does not look as a classical event log used in process mining analysis, but rather as a data table in which each tuple represents a patient, and each column one variable defined for the process. These variables include patient characteristics, such as origin of admission (Origin), results of fragility test (Fragilty), type of intervention (Interv), or sex of the patient (Sex), the dates of execution of diagnostic and therapeutic interventions, such as date of the last unscheduled admission (Date_UA), date on which a CT scan was performed on the patient (Date_CT), date of the heart team meeting of a patient (Date_MS), or date of its intervention (Date_Interv), among others.

As this registry is not suitable, as is, for applying process mining techniques to it, some preprocessing is necessary. An event log is constructed using the *patient id* as the case identifier, i.e. each patient will represent a case in the process, and the sequence of diagnostic and therapeutic interventions it goes through will represent the trace it follows. Thus, each diagnostic or therapeutic intervention will represent an event in the process. The distinct set of all events in the process represents the possible activities that can be executed, these are: the diagnostic or therapeutic interventions that can be performed on patients, patient management activities, and unexpected events. All attributes will be considered case attributes (apart from the executed activity

Table 3
Preprocessed event log extracted from the original AS ICP registry from Table 2. Columns starting with *"event_"* refer to event attributes, and columns starting with *"case_"* refer to case attributes.

case_id	event_activity	event_time	case_sex	case_age	case_fragility
1	consultation	2013-06-04 09:00	Male	84	1
1	special-consultation	2012-06-14 09:00	Male	84	1
1	echocardiogram	2012-06-21 09:00	Male	84	1
1	consultation	2012-06-21 10:00	Male	84	1
⋮	⋮	⋮	⋮	⋮	⋮
21657	echocardiogram	2012-10-25 09:00	Female	76	0
21657	consultation	2012-10-25 10:00	Female	76	0

Table 2
Extract of the original registry of clinical data from the Aortic Stenosis Integrated Care Process.

ID	Origin	Date_UA	Date_CT	Date_RX	Frailty	Date_MS	Interv	Date_Interv	Date_Rel	Sex	Birthdate
1	1	–	13/01/2018	24/11/2017	0	26/01/2018	0	07/06/2018	12/06/2018	1	04/04/1938
2	1	–	22/08/2017	04/10/2017	–	18/12/2017	0	26/06/2018	24/08/2018	1	12/04/1956
3	1	31/05/2018	15/03/2018	–	0	16/03/2018	1	03/07/2018	02/10/2018	0	13/09/1936
4	1	14/08/2018	10/04/2018	03/10/2016	1	27/04/2018	1	31/07/2018	10/10/2018	0	22/01/1934
5	1	–	09/01/2018	26/09/2017	1	02/11/2017	1	19/04/2018	30/10/2018	0	15/04/1936
6	1	04/02/2018	–	28/06/2017	1	05/12/2017	0	01/06/2018	24/09/2018	1	11/03/1942
7	1	06/02/2020	27/03/2018	28/07/2014	0	06/04/2018	0	05/09/2018	17/10/2018	1	06/05/1935
8	1	04/03/2019	15/07/2018	15/01/2018	0	06/07/2018	0	06/11/2018	19/12/2018	1	18/04/1935

and the timestamp of an event), as in the original log attributes refer to each patient, and as each patient is considered a case, all attributes are understood as case attributes. The output of this stage is an event log in the correct format for applying process mining techniques. Table 3 shows an excerpt of the event log constructed from the AS ICP original registry.

5.2. Process discovery

The next step is to discover the underlying process model from the event log by applying a discovery algorithm [6,35]. Traditionally these algorithms have followed different approaches, such as the heuristic miner [36], the inductive miner [37] or the evolutionary-based algorithms [38], among others. Here we use the inductive miner algorithm as it provides sound models with fitness 1[7], allowing in further stages to replay the whole behavior observed in the event log. Fig. 7 shows the process model of the AS ICP extracted from the event log. The input of this stage is an event log, and the outputs are both the event log and the discovered process model.

5.3. Mining additional perspectives

Process mining techniques not only allow to show what is happening with a process through process models extracted with discovery techniques (control-flow perspective), but can also generate insights on compliance, performance and efficiency: timestamps and frequencies of activities can be used to identify bottlenecks and diagnose other performance-related problems, and case data can be used to better understand decision-making and analyze differences among cases.

Process model replay establishes a strong relation between a process model and reality (as in the form of the event log the process model is extracted from) by relating events in the log to activities in the process. Using an event log and a process model as inputs, each trace is played over the model: events and relationships between events are fired over the model following the order indicated by each trace from start to finish. So, having the process model as a base, each case from the log is taken, and each of its events is fired following the control-flow indicated by the model. Fig. 8 shows an example of how replay works.

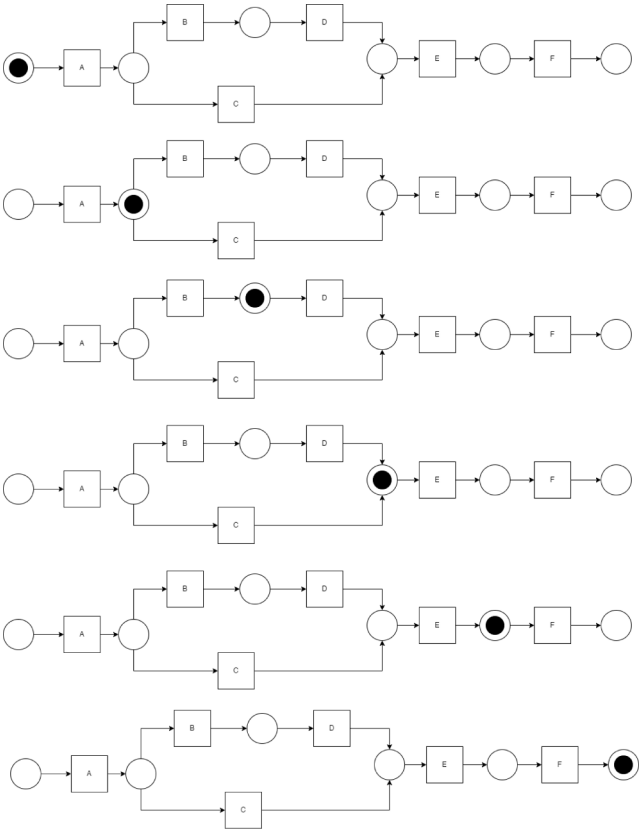


Fig. 8. Replaying of trace $= \langle a, b, d, e, f \rangle$ on top of its corresponding model (represented as a Petri Net).

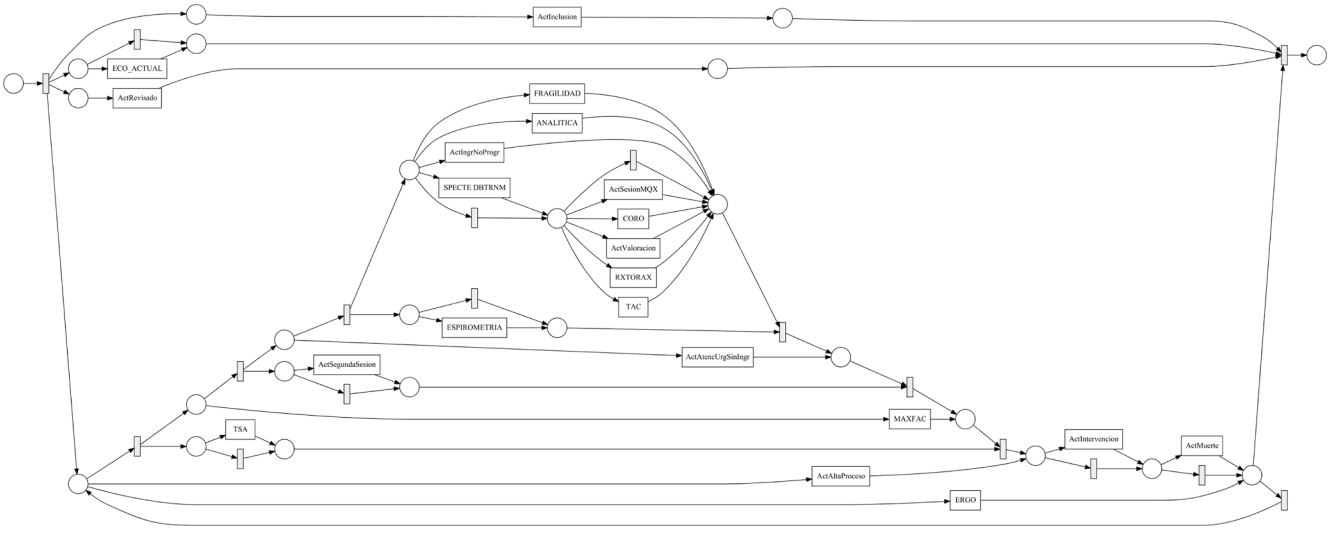


Fig. 7. Model of the valvulopathy process discovered with the inductive miner in the ProM v6.9[6] tool and represented as a Petri Net. The names of the activities are in Spanish as it is the original language of the process.

[7] Fitness refers to the quality of a process model of supporting or modeling the behavior seen in the event log, it measures "the proportion of behavior in the event log possible according to the model".

Table 4
Excerpt of durations (expressed in days) between the heart team meeting session and the intervention of a patient and the heart team meeting and the CT scanning of a patient. Duration is positive if origin event precedes destination event and negative if destination event precedes origin event.

case_id	MS-intervention	MS-CTscan
1	131	−13
2	188	−118
3	107	−1
4	86	−10
5	167	68
6	176	-
⋮	⋮	⋮
21654	120	9
21655	28	28
21656	54	−7
21657	155	49

5.3.1. Time and frequency

During the replay of the log over the model, as each activity and transition is executed, a collection that associates for each case its executed activities and relationships between activities, and how much time was spent in each of them is recorded. An example of this collection can be seen in Table 4. This collection allows to see which parts of the model are visited frequently (how many times each activity and relationship between activities has been executed) and how much time is spent in each part of the process.

Even though discovery is a necessary step, as it allows for determining how activities relate between them (this is unknown in the event log as parallels and choices are not represented in it), it is important to notice that this temporal information can not be obtained only by discovering the underlying process model. Replay is a requirement for discovering the temporal perspective, as only by relating the data recorded in the log to the control-flow perspective, we can collect frequency and timing information. The output of this stage is an enhanced event log, which now contains as new case attributes, the collected frequency and timing information.

Table 5 shows an excerpt of the output of this stage for the AS ICP case. The log is now enhanced with the collection of durations for the transition between activities *heart team meeting* and *CT scan* and the duration of the stage between the *heart team meeting* and the *intervention* of each case.

5.3.2. Feature extraction and data analysis

Collecting performance information is not enough to completely understand what is happening in the AS ICP or any process without further analysis. Once a deviance in a KPI or a delay has been detected, one of the most important questions is still unanswered: *What is the reason for the deviance or the delay?* We will refer to this step as *feature extraction and data analysis*: having as input the enhanced event log obtained in the replay stage, the goal is to describe cases in terms of a vector of variables, the features, that allows for applying multiple analysis techniques. This stage is guided by the knowledge of experts on the process domain. At the start of the pipeline, experts relate the information of the process they consider relevant and want to get insights on. Now, what features or variables are required for generating the descriptions that fulfill these requirements must be defined, selected and created (by combining existing features) when necessary. So now, we have data about cases (patients when talking about processes in the healthcare domain) in the process as a dataset in the form of a table, containing all the relevant features of each case, namely: its identifier, its corresponding trace and case attributes, including the frequency and temporal information discovered in the previous stage. Additional features may be computed at this stage e.g., for the AS ICP, a patient's age is computed through its birth date and the current date, volume variable *num_tests* is computed by the sum of all complementary test performed to a patient, volume variable *num_events* is computed by the sum of all unexpected events that took place during a patient stay in the service, etc. Table 6 shows the dataset corresponding to the AS ICP, and extracted from the enhanced log resulting from the previous stage. Only some features are shown, as the complete set of features characterizing the cases is too large.

Decision mining aims to find rules explaining choices in the process in terms of the features of a case. However, this is not limited to explaining choices in the process via rules, as any combination of predictor and response variables can be used, classification algorithms, statistical analysis, fuzzy quantification and many other techniques can be used. This way, attention can be focused on explaining deviation and

Table 5
Enhanced event log with the wait time (expressed in days) between the heart team meeting session and the intervention of a patient and the heart team meeting and the CT scanning of a patient extracted through replaying. Duration is positive if origin event precedes destination event and negative if destination event precedes origin event. Columns starting with *"event_"* refer to event attributes, and columns starting with *"case_"* refer to case attributes.

case_id	event_activity	event_time	case_sex	case_age	case_fragility	case_MS-intervention	case_MS-CTscan
1	consultation	2013-06-04 09:00	Male	84	1	46	−13
1	special-consultation	2012-06-14 09:00	Male	84	1	46	−13
1	echocardiogram	2012-06-21 09:00	Male	84	1	46	−13
1	consultation	2012-06-21 10:00	Male	84	1	46	−13
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
21657	echocardiogram	2012-10-25 09:00	Female	76	0	23	−118
21657	consultation	2012-10-25 10:00	Female	76	0	23	−118

Table 6
Dataset corresponding to AS ICP event log with its original attributes as well as the newly calculated with process mining techniques.

case_id	trace	admittance	age	sex	int_year	num_events	num_tests	HTM_INT	CT_HTM
1	$\langle incl, valor, ergo, coro, \ldots\rangle$	ambulatory	82	male	2018	0	3	46	−13
2	$\langle incl, valor, coro, CT, \ldots\rangle$	ambulatory	64	male	2017	0	2	23	−118
3	$\langle emerg, emerg\_admi, incl, valor, \ldots\rangle$	ambulatory	84	female	2018	1	2	24	−1
4	$\langle emerg, emerg\_admi, incl, valor, \ldots\rangle$	ambulatory	86	female	2018	1	3	32	−10
5	$\langle incl, valor, coro, CT, \ldots\rangle$	ambulatory	84	female	2017	1	4	41	68
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
21655	$\langle incl, valor, coro, CT, \ldots\rangle$	ambulatory	78	male	2017	0	2	48	75
21656	$\langle emerg\_admi, incl, valor, coro, \ldots\rangle$	ambulatory	85	male	2018	0	2	46	−7
21657	$\langle emerg\_admi, incl, valor, coro, \ldots\rangle$	ambulatory	85	male	2018	0	3	67	49

delays related to KPIs such as checking if the execution of certain activity influences the wait time on a case (relating control-flow and time and frequency information), how attributes affect the executed activities on a case, correlating the wait time between two activities to the result of variable of the process, etc.

A key element in this approach is the usage of statistical analysis techniques and expert knowledge to guide the generation of the natural language descriptions, i.e., statistical analysis is used to decide which descriptions are relevant and should be generated for a process. Statistical testing allows to determine whether statistically significant differences exist between expected and real values of features, the value of some feature for different groups of cases, the proportion of cases having certain feature, etc. Furthermore, as any combination of features is possible, this comparisons can be performed over newly discovered features regarding relationships between activities, allowing to describe and compare the control-flow and time and frequency perspectives for groups of cases (grouped by some other feature). This way, based on expert knowledge, hypotheses are established and statistical analysis techniques (according to the type of data under analysis in each case) are applied in order to confirm or disprove said hypotheses. Particularly, group comparison and proportion tests are used, adapted to the nature of the data: ANOVA, Mann-Whitney-Wilcoxon, Wilcoxon signed-rank test, one sample t-tests, etc. As an example, on the AS ICP, by applying the Mann–Whitney-Wilcoxon test to feature *HTM_INT*, which relates the waiting time between the heart team meeting and the intervention of a patient, and comparing groups based on feature *admittance*, it was found that patients from emergency origin have lower wait times between activities *heart team meeting* and *intervention*. In a similar manner it was found that cases in which a CT scan is performed tend to have longer throughput times that those which not.

The output of this stage is the new dataset with the additional discovered and computed features that characterize each of the cases, plus the knowledge obtained by applying statistical analysis techniques.

5.4. Protoforms for the description of processes

At this stage, utilizing the dataset constructed in the previous stage and fuzzy quantification techniques, guided by expert knowledge (included the one acquired in the previous stage), fuzzy quantified statements are generated. In this approach we propose a series of extensions and modifications of the type-I and type-II protoforms presented in Section 4 for the description of processes. Table 7 shows a summary of the proposed extensions, their associated protoform, and an example of its instantiation. In Section 6 the instantiaton of these protoforms for a specific health process (the AOS ICP) is presented.

5.4.1. Temporal contextualization of features

The first question experts try to solve when analyzing a process is how cases characterize and how their features evolve during the execution of the process. This can be done by generating descriptions that relate the value of case features, and how features relate to each other, in different time intervals of the process. This gives a first understanding of what is happening in the process and helps on detecting groups of cases (grouped by the same feature or set of features) or parts of the process that may need to be analyzed in further detail. For the generation of these statements, protoforms (1) and (2) are extended with the temporal characterization to the following protoforms respectively:

$$In\ Ti,\ Q\ cases\ had\ feature\ P \tag{7}$$

$$In\ Ti,\ Q\ cases\ with\ feature\ C\ had\ feature\ P \tag{8}$$

As in (1), Q is the quantifier and C and P are the qualifier and summarizer respectively. They can make reference to any of the features (or combination of features) of a case. Finally, Ti is the time interval in which attention on the feature (or combination of features) described by C and P is focused. It is defined as a crisp linguistic value. With (7) descriptions as the following can be generated e.g. “In year 2020, many cases had a long waiting time.” And (8) allows to check if any type of relation between features holds in a particular period of the process e.g., *“In year 2020, in most cases where resource MANAGER was involved, case had a short waiting time.”*.

Truth value of (7) can be directly calculated using [39]. For the evaluation of (8) we propose the following extension:

$$T = \mu_Q\left(\frac{\sum_{i=1}^{n}\mu_{Ti}(p_i)\wedge\mu_P(p_i)\wedge\mu_C(p_i)}{\sum_{i=1}^{n}\mu_T(p_i)\wedge\mu_C(p_i)}\right) \tag{9}$$

Where n is the number of cases, p represents a case, $\wedge$ represents the t-norm[8] minimum which is used as conjunction.

5.4.2. Frequency and temporal relations between activities

For understanding and improving wait times and control-flow, information about activity relationships is needed. Frequency and temporal relation protoforms aim to give a good understanding of the different paths cases follow (control-flow perspective), the frequency and wait times of activity relationships (frequency and time perspective) and how case features may influence case paths and wait times. These descriptions help to answer questions like “which are the most commonly followed paths and what exceptional paths are followed?”,

Table 7
Proposed extensions of type-I and type-II protoforms for the description of processes.

Extension	Protoform	Example
Temporal contextualization of features	*In T_i, Q cases had feature P*	In year 2020, many cases had a long waiting time
	In T_i, Q cases with feature C, case had feature P	In year 2020, most cases where resource MANAGER was involved had a short waiting time
Frequency and temporal relations between activities	*In T_i, in Q cases, R*	In the first semester of year 2021, in approximately half of cases, activity A takes place shortly after activity B
	In T_i, in Q cases with feature C, R	In the first semester of year 2021, in many cases where the wait time was long, activity C took place long after activity A
Compliance	*In T_i, P_1 is expected. However, in Q cases with feature C, had feature P_2.*	In year 2020, a wait time of 30 days between activity A and activity B is expected. However, in many cases where activity C is executed, wait time between activities A and B is higher than 30 days
Comparison between groups	*In T_i, Q_1 cases with feature C_1 had feature C_2. However, Q_2 cases with feature C_2 had feature P_2*	In year 2020, many cases where resource MANAGER was involved had a short waiting time between activities C and D. However, most cases where resource APPRENTICE was involved had a long waiting time between activities C and D.

[8] A t-norm is a binary operation that generalizes in fuzzy logic the conjunction operation.

"where are the bottlenecks in the process?" or "what may have caused the bottleneck?" which are some of the most frequent questions when applying process mining techniques, specially, on healthcare processes [1].

Information about frequency and timing of relationships between activities has been extracted and stored as case features when mining the time perspective as explained in Section 5.3 allowing for the generation of descriptions following the proposed protoforms:

$$In\ Ti,\ in\ Q\ cases\ R \tag{10}$$

$$In\ Ti,\ in\ Q\ cases\ with\ feature\ C,\ R \tag{11}$$

Where Ti, Q and C are as defined in (7) and R makes reference to a case feature which states whether a transition between two activities in the process takes place in a case. So for each case C_i, R_i is empty if the transition between two chosen activities does not take place (in that particular case), positive if an event representing the execution of the activity chosen as origin in the transition precedes an event representing the execution of the activity chosen as destination and negative in the opposite case. Evaluation process is similar as before, where truth degree T for (10) is computed as in [39] and for (11), as (9). Substituting P for R respectively in both cases. Following (10) descriptions as the following can be generated e.g. *"In the first semester of year 2021, in approximately half of cases, activity A takes place shortly after activity B"*. By following (11) relations between features and activity relationships can be described e.g. *"In the first semester of year 2021, in many cases where the wait time was long, activity C takes place long after activity A."*

5.4.3. Compliance

Compliance descriptions highlight deviances that happen in the process. By adding the expected value (the value originally defined by experts) of a feature to its description makes it easier for users to detect and understand deviances taking place in the process. By using a type-II protoform as in (8) descriptions of features causing deviance over the expected value of some other features can be generated e.g. describing how the participation of one particular resource in a case makes the wait time of transition between two activities longer than initially expected.

These descriptions are obtained by composing two protoforms: the first relates the expected value of some feature (defined by an expert) and the second contrasts the actual value of said feature in the process and is generated as defined in Section 5.4.2. Both statements are related through a semantic relation as in [40]. The composed protoform is structured as:

$$In\ T_i, P_1\ is\ expected.\ However,\ in\ Q\ cases\ with\ feature\ C\ had\ feature\ P_2 \tag{12}$$

This composite protoform allows for the generation of descriptions as *"In year 2020, a wait time of 30 days between activities A and B is expected. However, in most cases where activity C is executed, wait time between activities A and B is higher than 30 days"*.Its truth degree T is derived from the aggregation of the truth values of its constituents through any t-norm. For simplicity and consistency with the previously presented protoforms we propose the use of the t-norm minimum. If we refer to the expected value protoform "*In* T_i P_1 *is expected*" as $S1$ and to the contrasting protoform "*However, in* Q_2 *cases, patients with attribute C had attribute* P_2" as $S2$ the truth degree T of (12) is computed as:

$$T = T(S1) \wedge T(S2) \tag{13}$$

As the first statement is done over a known value (the value initially defined by experts), its truth value is always maximum. So, as by conjunction properties, the truth value of the deviance protoform is equal to the truth value of protoform $S2$. For this second protoform, truth value is computed as (8) or (11), depending if $S2$ describes a temporal relation between activities or other feature.

5.4.4. Comparison between groups of patients

Protoforms (8) and (11) give a first look at how two features relate: by using a feature on the qualifier and another on the summarizer one can see how the feature selected as qualifier affects that selected as summarizer[9]. Comparison descriptions aim at highlighting differences in traces, attributes case features between different groups of cases by comparing how a feature selected as summarizer behaves (which is its value) for different values of a feature selected as qualifier. Giving a comparison description, users can easily see if differences in a feature exist for different groups of cases e.g. seeing if differences for the wait time between two activities exist based on the execution of a particular activity on the process.

As compliance protoforms described in Section 5.4.3, these are composite protoforms. In this case, both protoforms relate the value of certain feature (summarizer) for a particular group of cases. The resulting composite protoform is defined as:

$$\begin{aligned} &In\ Ti,\ Q_1\ cases\ with\ feature\ C_1,\ had\ feature\ P_1.\\ &However,\ Q_2\ cases\ with\ feature\ C_2\ had\ feature\ P_2. \end{aligned} \tag{14}$$

Qualifiers C_1 and C_2 take different values of the same feature, and are used to define different groups of cases e.g. two of the possible resources in charge of a particular activity. C_1 and C_2 make reference to the same linguistic variable (qualifier) but take different linguistic values. Summarizers P_1 and P_2 work in a similar way: they convey different linguistic values of the linguistic variable under observation (summarizer) e.g., *lower than 30 days* vs. *higher than 30 days* for linguistic variable *wait time between activity A and activity B*. As before, the truth value is derived from the aggregation of the truth values of the constituents protoforms through any t-norm (we again propose the use of t-norm minimum). We refer to the first protoform "*In* T_i*, in* Q_1 *cases where patient had attribute* C_1*, patient had attribute* P_1" as $S1$ and to the contrasting protoform "*However, in* Q_2 *cases where patient had attribute* C_2 *patient had attribute* P_2" as $S2$. So the truth value of both S_1 and S_2 is computed as for (8) if P refers to a case feature or (11) if P refers to a transition between two activities. The truth value of the composite protoform is computed as (13). Following (14) descriptions as the following can be generated e.g. *In year 2020, many cases where resource MANAGER was involved had a short waiting time between activities C and D. However, most cases where resource APPRENTICE was involved had a long waiting time between activities C and D.*

5.5. Realization

The final stage of the pipeline is the linguistic realization phase. Generating fuzzy quantified statements into natural language texts which provide the final information to the users. Apart of being informative and relevant, the texts should be correct from all linguistic points of view (grammatical, morphological, lexical and orthographic).

In our framework, we follow a hybrid template-based realization approach, which integrates domain expert knowledge with text templates for generating the final texts. This approach is richer and more flexible than basic fill-in-the-gap template approaches, but simpler and quicker than full fledged NLG system implementations [41,42]. The SimpleNLG-ES [34] realization engine, (the Spanish version of the original SimpleNLG engine [43]) was used at this stage. For each of the proposed extended protoforms (temporal contextualization of features, frequency and temporal relationships between activities, compliance

[9] In (11), the summarizer specifically makes reference to a transition between two activities

and comparison between groups of patients), a base structure is defined. So even represented as separated stages in the pipeline, the generation of the fuzzy quantified statements its coupled with its realization. The basic template of each of the proposed extended protoforms is defined as a function which takes as inputs the data relative to the referential under evaluation and the quantifiers, qualifiers and summarizers (in the form of linguistic variables) which want to be used in the description. With the defined template for each protoform, the elements that constitute it, and the SimpleNLG realization engine, the protoform is evaluated and generated into a natural language text. The elements used in the protoform (referential, quantifiers, qualifiers and summarizers) can be chosen manually based on some knowledge or algorithms for choosing them can be used. As introduced in Section 5.3.2, for our approach, expert knowledge (what information is relevant for the users) and statistical analysis techniques are used for deciding which protoforms to generate.

This way, the basic syntactic structure of the description is fixed, but the terms used in it (summarizer, qualifier, quantifier) are selected and accordingly modified (number, tense, negation, etc.) for each of the natural language descriptions generated. Some additional logic exists in the templates; for example, when using an aggregation value as a feature summary (e.g. using the average value of a certain feature to give a quick insight of it instead of using a quantifier), the part of the template that relates "In *Q* cases…" gets replaced to "Cases…". Also, if multiple properties are used as a summarizer or qualifier, certain modifications need to happen, as using the correct connectives in each place is necessary for the correctness of the description.

6. Natural language descriptions of the AS ICP

Following the extended protoforms proposed in the previous section, we show how these descriptions are generated with the proposed framework, what information is used in each case and what knowledge or techniques are used for guiding the generation process of the descriptions. More specifically descriptions about temporal contextualization of case (patient) features, frequency and temporal relations between activities, compliance with clinical guidelines and comparison between groups of patients, are generated.

6.1. Temporal contextualization of features

One of the AS ICP concerns in respect of temporal contextualization of features is how many complementary tests are performed on a patient in Stage 1[10]. The following descriptions are generated:

- In year 2018, an average of 3 complementary tests are performed to a patient between the heart team meeting and its intervention.
- In year 2019, an average of 2 complementary test are performed to a patient between the heart team meeting and its intervention.

The generation process of these descriptions is as follows:

1. Feature *number of complementary tests* is created and computed by summing the number of complementary test performed between activities *heart team meeting* and *intervention* for each patient.
2. Years *2018* and *2019* are defined as crisp time intervals *Ti* and summarizer *P* is defined as the new computed feature *number of complementary tests*. Interval *Ti* works as a qualifier, defining a subset of the referential that will be evaluated against the summarizer *P*. This way, only cases starting in year 2018 are selected for the first description, and cases starting in year 2019 are selected for the second description.
3. The average number of tests for each subset defined by *Ti* is computed.
4. As an average value is given, no quantifier *Q* is needed and as this value summarizes each entire subset, therefore no evaluation is needed, the truth value of both protoforms is 1.

Experts are also interested in analyzing how the COVID-19 pandemic has affected patient characterization. By defining different years for *Ti*, experts can easily grasp if any difference in proportion of patients for some feature exists. The decision of generating these protoforms is based on the statistical analysis performed over feature *admittance*. By applying a proportion test on the data, comparing the proportion of *emergency admittance* patients in years 2019 and 2020, a statistically significant difference in proportion was found. The following descriptions are generated:

- In year 2019, in some cases (38,8%), patients had emergency admittance.
- In year 2020, in approximately half of cases (50,6%) patients had emergency admittance.

These descriptions allow experts to get some first insights and even see how the feature they are interested on, evolves during time. The generation process is as follows:

1. A set of proportional quantifiers is defined as in Fig. 3.
2. As before, *Ti* is defined as a crisp time interval and acts a qualifier. All cases are evaluated against qualifier *Ti*.
3. In this case, summarizer *P* makes reference to the type of admittance patients have in the process, this is feature *admittance* of cases (patients).
4. Feature *admittance* of all cases is selected and evaluated against membership function of summarizer *P*. As feature *admittance* only has two possible categorical values, a membership function for each category (each linguistic value) is defined as in Eq. (4). Linguistic value *emergency admittance* is taken for the evaluation.
5. Conjunction of sets of membership degrees to *Ti* and *C* is computed and aggregated into a membership proportion.
6. Membership proportion is taken and evaluated against the selected quantifier.
7. Protoform is generated.

6.2. Frequency and temporal relations between activities

On the AS ICP, attention is mainly focused on analyzing and reducing wait times in Stage 1. So medical experts show interest in descriptions that address activity relationships in this stage, mostly, how the CT scanning of a patient temporally relates to the heart team meeting and, if the CT scanning takes place after the heart team meeting, how much of a bottleneck represents for the intervention of the patient[11]. Medical expert knowledge is used to determine which protoforms to generate based on the interest in analyzing particular KPIs of the process. The following descriptions are generated:

- In approximately half of cases (52.72%), patient had a wait time between its CT scanning and its intervention lower than 30 days.

[10] Stage 1 spans from the decision to intervene a patient in the heart team meeting, to the actual valve replacement procedure

[11] Performing a CT scan is a requirement when percutaneous prosthesis implantation (TAVI) is performed, so the performance of the CT scan entails the main possible bottleneck between the decision to intervene a patient and its actual intervention.

- In approximately half of cases (56.28%) where patients were intervened via TAVI, a CT scan is performed after its corresponding heart team meeting.
- In approximately half of cases (51.57%) where patients were intervened via TAVI, patient had a wait time between its CT scanning and its intervention lower than 30 days, with a median value of 35 days.
- In most cases (90%) where patients were intervened via TAVI and to whom a CT scan was performed after its corresponding heart team meeting, patient had a wait time between its heart team meeting and its CT scanning lower that 30 days.

For the generation of these protoforms, the required columns of the dataset are: *intervention* (type of intervention performed on a patient), *CT_HTM* (wait time between the CT scanning and the heart team meeting of a patient) and *CT_INT* (wait time between the CT scanning and the intervention of a patient). These will be used as both qualifier *C* and summarizer *R* depending on the particular protoform instance. Linguistic values *before* and *after* are defined as crisp linguistic values with membership functions:

$$\mu_{\text{before}}(x) = \begin{cases} 1, & \text{if } x < 0 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

$$\mu_{after}(x) = \begin{cases} 1, & if\ x > 0 \\ 0, & otherwise \end{cases} \tag{16}$$

Where x is each of the elements of the newly computed attribute *CT_HTM* (wait time between the CT scanning and the heart team meeting of a patient) or *CT_INT* (wait time between the CT scanning and the intervention of a patient)[12]. In a similar way, summarizer *wait time between CT scanning and its intervention lower than 30 days* is defined as a crisp linguistic value with membership function:

$$\mu_{\text{wait time}}(x) = \begin{cases} 1, & \text{if } x < 30 \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

Value *30 days* is used as it is established as the maximum number of days of wait time between the heart team meeting and the intervention of a patient for providing quality care on the AS ICP clinical guideline.

Additional to the use of fuzzy quantifiers, a specific quantity is given in parentheses, this way, the granularity of the summary is chosen by the user, allowing him/her to grasp a general idea by focusing on the fuzzy quantifier, or getting the exact proportion of patients by focusing on the percentage. Same idea is followed with the median value of the wait time between the CT scanning and the intervention of the patient. With these elements protoforms are generated and evaluated as following:

1. A set of proportional quantifiers is defined as in Fig. 3.
2. Even time interval *Ti* can be used, the generated protoforms address the whole process time-span, so no need for defining or evaluating time intervals.
3. For type-II protoforms (all except the first), feature *intervention* of all cases is taken and used as qualifier. This feature is a categorical variable so, when defined as a linguistic variable, a membership function for each category (linguistic value) is defined as in Eq. (4). Linguistic value *intervened via TAVI* is selected.
4. Feature *intervention* of all cases is selected and evaluated against the membership function associated to qualifier *C*.
5. The corresponding feature used in each protoform (summarizer) is selected for all cases and evaluated against the membership function of the selected linguistic value of the summarizer *P* (wait time or temporal relation between activities).
6. Conjunction between the sets of membership degrees for qualifier *C* and summarizer *P* is computed.
7. The resulting set of membership degrees from the conjunction is aggregated and the proportion of membership is computed.
8. Membership proportion is evaluated against all possible quantifiers defined in step 1. Selecting the one with a higher truth degree.
9. Protoform is generated.

As quantity (e.g. *approximately half of cases*), temporal reasoning (e.g. *after*) and duration (e.g. *less than 30 days*) of a relationship between two activities (e.g. intervention and CT scanning) is given, this descriptions provide revealing insights to medical experts about where and how frequently bottlenecks in the process occur. Also, as the extended protoforms include a qualifier (e.g. *patients were intervened via TAVI*) relating other features of a case (patient) to the frequency and timing perspective, this allows to further analyze if any relation holds between patient characteristics (e.g. type of admittance, age of the patient, year of inclusion in the process, etc.) or volume variables (e.g. number of events suffered, number of complementary tests performed, etc.) with process variables (compliance with care protocol standards related with timing, delays and bottlenecks and paths followed by patients), thus helping answer the question “what caused the bottleneck?”.

6.3. Compliance

In the AS ICP interest is focused again on the wait time for Stage 1. By adding the expected wait time for this stage, descriptions help on detecting more easily compliance (or not) with guidelines. Also, medical experts from the Cardiology Department, but not involved in the AS ICP may need to participate in the process at some point, as healthcare processes are highly multidisciplinary [3,1,6]. In those cases, new participating experts may not know the clinical guidelines that support the process, thus, basic frequency and temporal descriptions, without further context, may not convey all the necessary information this experts need. By adding this context, all the necessary information for grasping whether a deviance in behavior or timing exists. The following description is generated:

- Wait time between the heart team meeting of a patient and its intervention is expected to be lower than 30 days. However, in approximately half of cases (59.30%) patient had a wait time between its heart team meeting and its intervention higher than 30 days.

Medical expert knowledge is used to determine which protoforms to generate based on the interest in analyzing particular variables of the process. Also, statistical testing was used to assess if statistically significant differences exist between the real and expected values of feature *wait time between heart team meeting and intervention*. In particular, a one-sample t-test was used to discover that statistically significant differences exist between the expected and real value of the feature. This knowledge also determines the usage of the semantic relation “However”, as the relevant aspect experts are trying to analyze is if deviances respect to guidelines or expected behaviors are taking place.

6.4. Comparison between groups of patients

In the AS ICP great interest is shown in analyzing if patient characterization affects output or process variables, specially how patient admittance, type of intervention and year of admittance affect wait times. The following descriptions are generated:

[12] Functions *before* and *after* are defined the same way for any temporal distance attribute computed, here is exeplified with attributes *CT_HTM* and *CT_INT*.

- In many cases (76.70%) where patient had ambulatory admittance, patient had a wait time between its heart team meeting and its intervention higher than 30 days. However, in many cases (64.64%) where patient had emergency admittance, patient had a wait time between its heart team meeting and its intervention lower than 30 days.
- In approximately half of cases (42.25%) where patient was intervened in year 2019, patient had a wait time between its heart team meeting and its intervention lower than 30 days. However, in many cases (77.50%) where patient was intervened in year 2020, patient had a wait time between its hear team meeting and its intervention lower than 30 days.
- Cases where patient was intervened via valve replacement had a wait time between its heart team meeting and its intervention of 83 days in average. However, cases where patient was intervened via TAVI have a wait time between its heart team meeting and its intervention of 49 days in average.

By comparing wait times between different years it is possible to analyze how the process has evolved and, if changes made during those years are fruitful and reducing wait times. The generation process is as follows:

1. A set of proportional quantifiers is defined as in Fig. 3.
2. Even time intervals *Ti* can be used, the generated examples refer to the whole process time-span. In the example were years 2019 and 2020 are compared, time intervals are used directly as qualifiers referring to a case feature *intervention_year*.
3. Feature used as qualifiers C_1 and C_2 (e.g. *intervention_year*) is taken for $S1$ and $S2$ respectively, and cases (patients) are classified and grouped based on the selected linguistic values (e.g. *2019* as C_1 and *2020* as C_2). This is done by evaluating the selected feature of all cases against the membership function of qualifier C_1 for S_1 and C_2 for S_2 respectively.
4. The feature used in each protoform as summarizer (e.g. *wait time between its heart team meeting and its intervention*) is taken for all cases and evaluated against the membership function of the selected linguistic values P_1 and P_2 (e.g. *lower than 30 days* and *higher than 30 days*).
5. Conjunction between the sets of membership degrees for qualifier C and summarizer P is computed for each protoform S_1 and S_2.
6. The resulting set of membership degrees from the conjunction is aggregated and the proportion of membership is computed for S_1 and S_2 respectively.
7. Membership proportion is evaluated against all possible quantifiers defined in step 1. Selecting the one with a higher truth degree for each S_1 and S_2.
8. Semantic relation "*However*" is used, as this protoform conveys differences between groups of patients.

Statistical testing is used to check if statistically significant differences exist between groups and decide which descriptions to generate. For these protoforms, proportions tests where used to assess if differences in the number of cases with a waiting time lower than 30 days was statistically significant between patients with different admittance origin and between patients intervened in years 2018 and 2019. A Mann–Whitney–Wilcoxon test was used to determine if statistically significant differences exist for the average waiting time of patients intervened via valve replacement and via TAVI. Based on medical expert hypotheses and using statistical testing accurate descriptions of the process are generated.

7. Evaluation

In this section we present the evaluation of our proposal in a real healthcare environment, the AS ICP. Following the current standards of manual human expert evaluation of the NLG field [44], this evaluation was conducted by 15 medical experts of the Cardiology Department of the University Hospital of Santiago de Compostela who did not participate in the definition of the system requirements. Manual evaluation by human experts is the appropriate evaluation methodology for this case, since automatic metric-based cannot be performed due to the lack of corpus or available datasets and to the vivid discussion on the NLG community about the validity of the metrics used. Comparison to other approaches cannot be performed since, to the best of our knowledge, this is the first system which deals with the problem of describing healthcare processes using NLG and particularly the AS ICP.

7.1. Set-up

The goal of human expert validation of a NLG system is, in general, to assess to what extent the generated natural language descriptions are understandable for its users and provide truthful and relevant information. Additionally, for the AS ICP, we are also interested in assessing if the descriptions are helpful in order to improve users work quality and, ultimately, if they give answer to any of the questions medical experts may have when dealing with and analyzing the AS ICP, or other possible healthcare processes. Following the best practices of human expert evaluation in NLG systems [44], we created a Likert-scale questionnaire where the dimensions to be assessed are asked to the human evaluators over different cases. Availability of human experts is one of the critical issues when performing manual validation in NLG, since medical experts, who are the human experts for this application, have very limited quality time for performing these tasks. This general problem is solved by focusing the assessment items and the cases included in the questionnaire on the most critical issues to be evaluated, avoiding creating a very long questionnaire involving too many different questions. As described in what follows, each human evaluator answered 9 questions about 6 different cases or scenarios, and 6 more general questions about the approach as a whole. We found this as a very good compromise between the evaluation needs and the limited availability of the medical experts.

The proposed questionnaire consists of 6 cases or scenarios regarding different elements of the AS ICP medical experts have identified as relevant and on which they have interest in get insights of. For each scenario, the natural language description generated with the framework here proposed is presented next to the corresponding graphical representation that would be used in classical process mining analysis. Fig. 9 illustrates a scenario in which a natural language description regarding the expected and real wait times between the heart team meeting and the intervention of a patient is shown next to the histogram representing the wait times (in days) for the patients between said activities on the process. For each scenario, 9 questions were asked, on a five-level Likert scale ranging from "strongly agree" to "strongly disagree". As a hybrid template-based realization approach is used, form aspects regarding linguistic quality of the descriptions (i.e., spelling and grammatical correctness) do not need to be assessed. Therefore, the evaluation process is mainly focused on content-related aspects, aiming to evaluate if natural language descriptions help on the understanding of process mining analysis results, its relevance and usefulness for the users, and if the descriptions provide interesting information for medical experts. Table 8 shows the questions asked for each scenario.

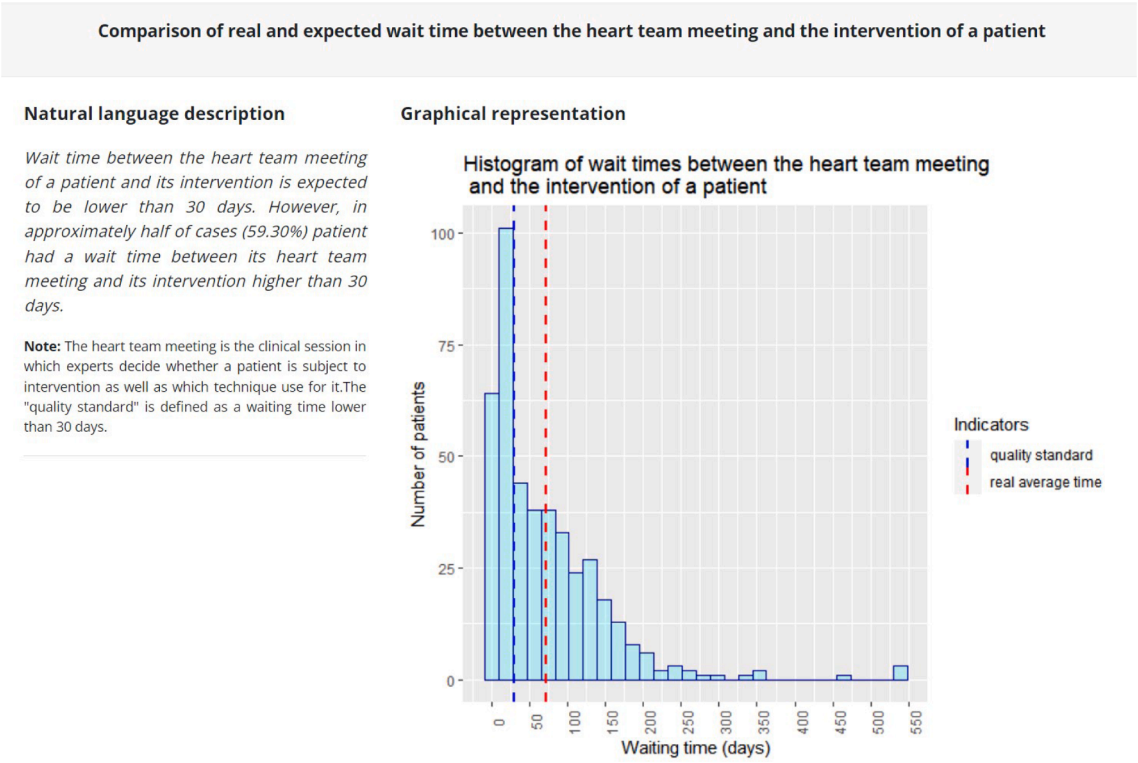


Fig. 9. One of the scenarios presented in the evaluation questionnaire of the pipeline over the AS ICP (originally in Spanish). In this scenario a natural language description regarding the expected and real wait times between the heart team meeting and the intervention of a patient is shown next to the histogram representing the wait times (in days) for the patients between said activities on the process.

Questions were grouped based on the aspect they evaluate: first two questions (*I1-I2*) assess the content presented on the scenario, helping understand if the information presented is interesting and novel for the medical experts. The second group of questions (*R1-R4*) focuses on the preferred modality of conveying information for medical experts (natural language vs. graphical displays). They aim at evaluating if natural language descriptions are preferred over graphical representations of the same information, by directly asking which representation of the information, or if the combination of both (and in which order of preference) was the most efficient in conveying the information of the scenario. It is relevant to emphasize that this is one of the main objectives of this evaluation, *assess whether medical experts can better understand process related information based on natural language descriptions rather than on graphical representations*. Next two questions (*C1-C2*) assess the quality of the natural language descriptions by asking medical experts to evaluate the ease of understanding the information presented in the natural language description and whether the information was conveyed in a comprehensible way. The last question (*F1*) regards the use of fuzzy quantification and fuzzy terms in the descriptions, by asking if the degree of concreteness in which the information is conveyed in the natural language description (by the use of average values, fuzzy quantifiers, etc.) is adequate and not too succinct to not be informative or too detailed to not give a general enough description of what is happening in the process.

For facilitating and reducing the amount of time users spend when answering of the test, question *R2* is shown only whenever the medical expert has selected a value "Neutral" or worse for *R1*. Similarly, *R3* and

Table 8
Questions asked for each scenario on the proposed evaluation test.

Id	Question
I1	The information this scenario provides is interesting for me.
I2	The information this scenario provides is new for me (I was unaware of it until now).
R1	The representation which provided me with information most efficiently has been the natural language description.
R2	The representation which provided me with information most efficiently has been the graphical representation.
R3	Natural language description coupled with the graphical representation has provided me with information in the most efficient way. Without the natural language description I would not have understood the scenario.
R4	Graphical representation coupled with the natural language description has provided me with information in the most efficient way. Without the graphical representation I would not have understood the scenario.
C1	Understanding the information this natural language description provides is easy for me
C2	I perceive information is expressed in a comprehensible way.
F1	The degree of concreteness in which the information is expressed in this natural language description is adequate (neither too detailed nor too succinct).

Table 9
General questions asked at the end of the evaluation test.

Id	Question
P1	I am familiar with the AS ICP
G1	Natural language descriptions would provide me a better understanding of what happens in my job.
G2	Natural language descriptions would allow me to complete tasks quicker.
G3	Natural language descriptions would increase the quality of my work.
G4	I would find natural language descriptions useful in my job.
G5	I find natural language descriptions easy to understand.

R4, are only shown when previous questions are answered with values "Neutral" or worse. The answer for all hidden questions in this category was set to "Strongly Disagree", as by agreeing to the previous questions user is implicitly disagreeing with them. This was done after a revision of a preliminary version of the questionnaire, since the AS ICP coordinators highlighted that users would probably consider answering to all four questions as redundant and repetitive, and would not fully complete the questionnaire because of their very limited time availability. This emphasizes the availability of human experts as a critical issue when evaluating NLG systems, as pointed out earlier. Under these challenging circumstances, having fifteen human experts is a rather unusually high number of participants, which makes the results conclusive for this application.

A simplified version of the Technology Acceptance Model [45] adapted to linguistic summarization used in [16] is used as a set of general questions asked at the end of the evaluation process in order to evaluate the natural language description of healthcare processes as a whole. This way, once evaluated whether experts prefer natural language or graphical displays, and assessing the quality of the descriptions, direct questions about perceived usefulness and intention of use are asked. Table 9 shows this set of questions. Questions *G1-G4* evaluate usefulness and intention of use and question *P1* assesses how familiar the medical expert who is performing the evaluation is with the AS ICP. Through this question, conclusions can be drawn about to what extent the descriptions are found more/less useful or more difficult/easier to understand based on the different medical expert profiles.

The questionnaire was deployed as a web page with two introductory screens followed by the six scenarios created. The first screen related the objective of the evaluation process, the second one showed instructions regarding what information is shown in each scenario, what each question is trying to assess and how to answer the questions.

7.2. Results

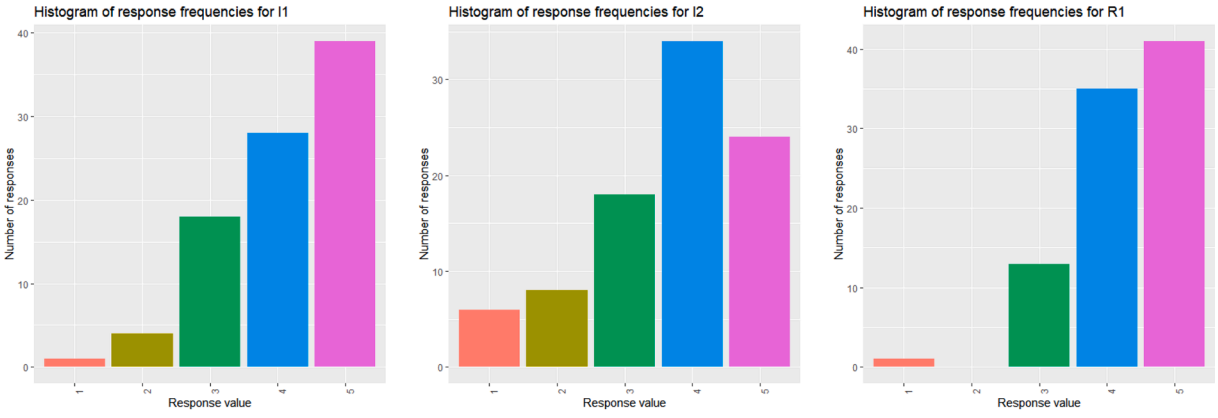
Table 10 shows the validation results for each of the proposed questions. Results are shown in a 5-level numerical scale derived from the Likert-scale used in the questionnaire, where 1 equals "Strongly disagree" and 5 equals "Strongly agree". The table shows the average value and its confidence interval at a 95% confidence (CI) (computed using the t-student distribution as the population standard deviation is not known), standard deviation (SD), mode, median and interquartile range (IQR) for each of the asked questions.

As we can see with questions *I1-I2*, in average, results show most descriptions are found to provide really interesting (4.11/5.00) and novel information (3.69/5.00). This is specially relevant, considering that 13 out of the 16 participants (81.25%) were already familiar with the AS ICP. Questions *R1-R4*, in particular question *R1*, show that natural language descriptions are (by far) the preferred way of conveying information for almost all users, who show a very clear preference for natural language descriptions (4.28/5.00) over graphical displays (1.30/5.00). Questions *C1-C2* prove information was found to be quite easy to understand (4.40/5.00) and no user has had any trouble understanding the proposed descriptions because the information was expressed in a comprehensible way (4.38/5.00). Finally, the degree in which information was presented, evaluated with question *F1* is found correct with a high agreement (4.22/5.00). So, not only natural language descriptions convey information in a more convenient way than graphical displays but also users that interact on a daily basis with the AS ICP find novel information on the descriptions. This proves the techniques proposed in this framework are helpful on extracting and conveying process mining analysis results to medical experts in a comprehensible and easy to understand way, utilizing fuzzy terms that help on the summarization of numerical data. Fig. 10 shows the histograms of response values for each of the asked questions.

Table 10
Expert validation results for the AS ICP descriptions when taking the whole answer set (6 scenarios per 15 users). Results are shown in a 5-level numerical scale derived from the Likert-scale used in the questionnaire (1 equals "Strongly disagree" and 5 equals "Strongly agree"). Average value and its confidence interval at a 95% confidence (CI), standard deviation (SD), mode, median and interquartile range (IQR) are shown for each of the asked questions.

	Average	CI	SD	Mode	Median	IQR
I1	4.11	[3.94, 4.27]	0.95	5	4	1.75
I2	3.69	[3.48, 3.89]	1.16	4	4	2.00
R1	4.28	[4.13, 4.41]	0.79	5	4	1.00
R2	1.30	[1.15, 1.44]	0.80	1	1	0.00
R3	1.26	[1.22, 1.38]	0.76	1	1	0.00
R4	1.12	[1.03, 1.20]	0.47	1	1	0.00
C1	4.40	[4.26, 4.53]	0.75	5	5	1.00
C2	4.38	[4.23, 4.52]	0.82	5	5	1.00
F1	4.22	[4.06, 4.35]	0.78	4	4	1.00

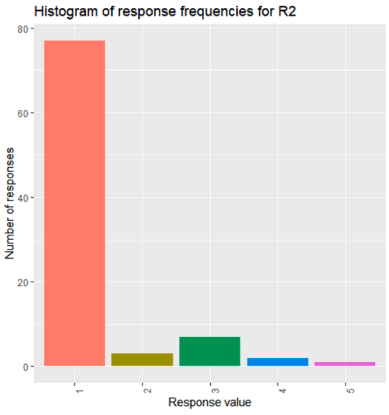
The general questions regarding the whole approach also show very positive results, with an average value higher than "agree" (4.07/5.00) for all questions. Table 11 shows the average value and its confidence interval (CI) (computed using the t-student distribution as the population standard deviation is not known), standard deviation (SD), mode, median and interquartile range (IQR) for each of the asked questions. Results prove that users are keen on incorporating natural language descriptions into their daily workflow as they find them useful (4.12/5.00) and easy to understand (4.44/5.00). Users also state that natural language descriptions provide a better understanding of what happens at their job (4.06/5.00), allow them to complete tasks quicker (4.06/5.00) and increase the quality of their work (3.69/5.00). Thus, reinforcing the conclusions extracted from the results obtained in the previous part of the evaluation test. Question *P1* is not used nor shown on Table 11 as all evaluators except three were already familiar with the process, thus, not allowing us to extract any valuable information from this question. Fig. 11 shows the histograms of response values for each of the asked questions.



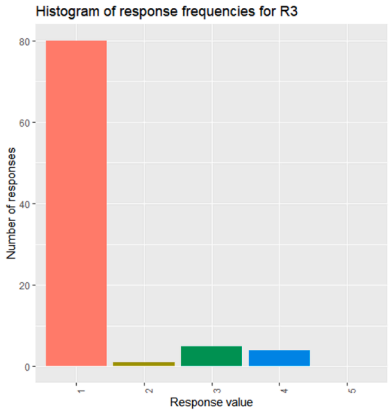
(a) Histogram of response frequencies for question *I1*.

(b) Histogram of response frequencies for question *I2*.

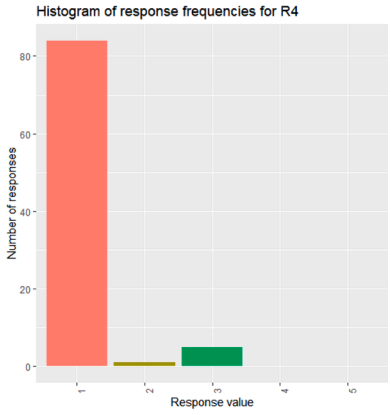
(c) Histogram of response frequencies for question *R1*.



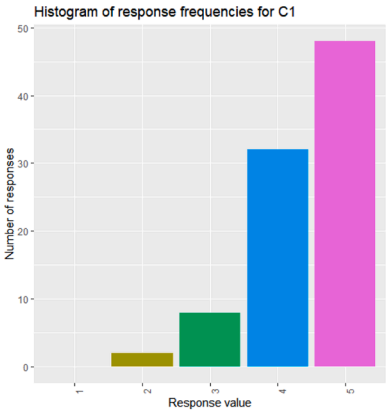
(d) Histogram of response frequencies for question *R2*.



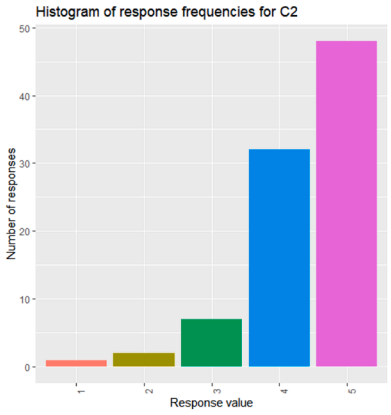
(e) Histogram of response frequencies for question *R3*.



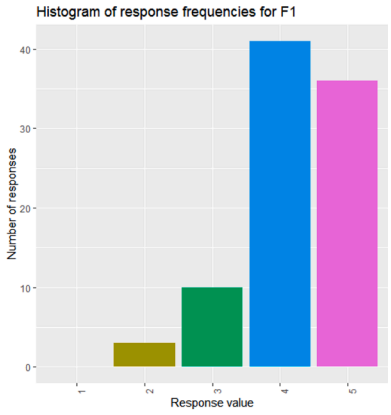
(f) Histogram of response frequencies for question *R4*.



(g) Histogram of response frequencies for question *C1*.



(h) Histogram of response frequencies for question *C2*.



(i) Histogram of response frequencies for question *F1*.

Fig. 10. Histograms of response frequencies for the questions asked in each scenario.

8. Conclusions and future work

In this paper we presented a framework for the generation of qualitative and quantitative natural language descriptions of healthcare processes addressed to medical experts. The framework is complete, since it is able to handle all stages of the generation, from the preprocessing of clinical registries to event logs, to the final generation of the natural language texts. It is based on the most widely used Data-To-Text (D2T) pipeline [2], on the usage of process mining techniques, and fuzzy quantification techniques which allow to model uncertain terms in the natural language descriptions.

The framework is able to handle relevant healthcare process data

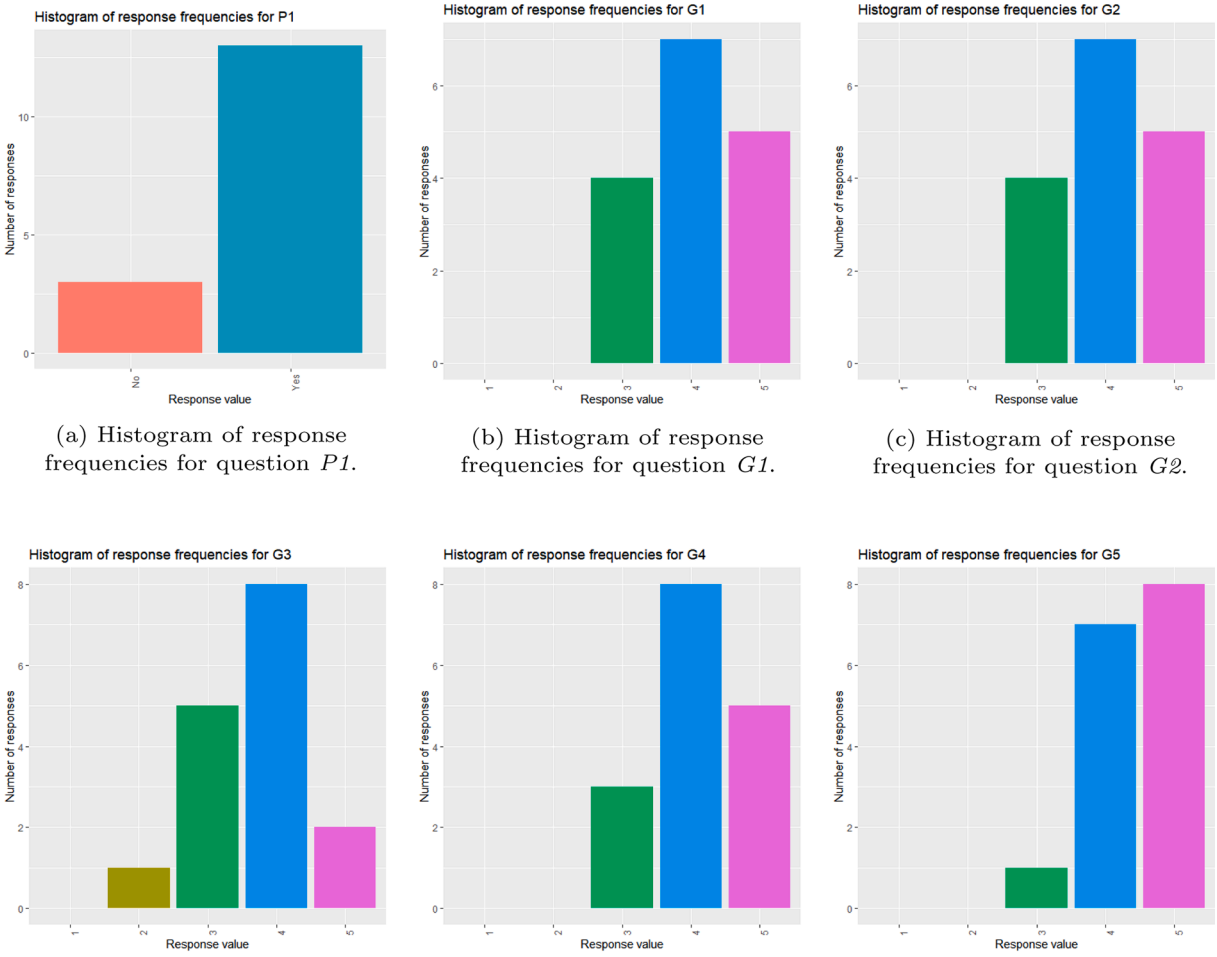
Table 11
Expert validation results for the general questions of the AS ICP. Average value and its confidence interval (CI), standard deviation (SD), mode, median and interquartile range (IQR) are shown for each of the asked questions.

	Average	CI	SD	Mode	Median	IQR
G1	4.06	[3.72, 4.40]	0.77	4	4	1.25
G2	4.06	[3.72, 4.40]	0.77	4	4	1.25
G3	3.69	[3.34, 4.04]	0.79	4	4	1.00
G4	4.13	[3.81, 4.44]	0.72	4	4	1.00
G5	4.44	[4.16, 4.71]	0.63	5	5	1.00

such as events and its attributes, temporal relations between events, patient attributes, and quantify them during process life-span, recall temporal relations and waiting times between events and its possible causes and compare patients attributes between groups, among other features.

A real application of the framework was presented and validated, over the Aortic Stenosis Integrated Care Process of the University Hospital of Santiago de Compostela. Following the usual standards of D2T systems, manual human validation was conducted for the generated textual descriptions by 16 medical experts in Cardiology. Validation results are very positive, since from a quantitative point of view a global average of 4.07/5.00 was obtained for the general questions related to the understandability and usefulness of the natural language descriptions as well as the capability of medical experts (those who validated the approach) to complete tasks easier and increase the quality of their work. Furthermore, numerical assessment for question *R1* (the representation which provided the users with information most efficiently was the natural language descriptions) was 4.28/5.00, which is much better than the assessment for graphical representations or combinations of both (1.30/5.00, 1.26/5.00, and 1.12/5.00, respectively). More specifically, they show *i)* that the modality which conveyed the information most efficiently about the process was natural language; *ii)* a very clear preference of texts over the usual graphic representation of processes information as the way for conveying information to experts; and *iii)* natural language descriptions provided relevant and useful information about the process, providing ways for its improvement.

As future work, we will aim to increase the expressiveness of the content conveyed in the descriptions and extend it to other relevant variables, attributes and indicators, obtained through applying other process mining techniques. For instance, frequent and infrequent patterns can be introduced, since they may give an easy understanding of how a process is taking place in reality and if unexpected behaviors are taking place. Also, decision mining and data mining techniques apart from statistical analysis can be included. The discovery of rules explaining associations between process elements can help in the generation of new natural language descriptions that provide insights yet unknown. Also, the discovered rules may be used in the content determination stage as an indicator of which information is may be described, helping on the automation of said stage of the framework. Finally, richer



(a) Histogram of response frequencies for question *P1*.

(b) Histogram of response frequencies for question *G1*.

(c) Histogram of response frequencies for question *G2*.

(d) Histogram of response frequencies for question *G3*.

(e) Histogram of response frequencies for question *G4*.

(f) Histogram of response frequencies for question *G5*.

Fig. 11. Histograms of response frequencies for the questions asked in each scenario.

textual descriptions will be built through a more complex and complete NLG realization stage. For instance, adding a lexicalization stage would allow for generating a greater variety of natural language descriptions, as same concepts could be realized using different terms, allowing for example to generate texts conveying the same information with different words for different medical expert profiles (e.g. a surgeon and a cardiac imaging expert). The addition of referring expression generation and aggregation phases will allow to generate better structured descriptions and combine them when needed (when two texts refer to the same subject) enhancing this way the readability of the texts and allowing to generate richer and more cohesive descriptions, giving place to whole process mining analysis result description documents.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, Process mining in healthcare: A literature review, J. Biomed. Inform. 61 (2016) 224–236, https://doi.org/10.1016/j.jbi.2016.04.007.

[2] E. Reiter, An Architecture for Data-to-Text Systems, in: Proc. 11th European Workshop on Natural Language Generation, ENLG '07, ACL, USA, 2007, pp. 97–104.

[3] Álvaro Rebuge, D.R. Ferreira, Business process analysis iacn healthcare environments: A methodology based on process mining, Inform. Syst. 37 (2) (2012) 99–116, https://doi.org/10.1016/j.is.2011.01.003, management and Engineering of Process-Aware Information Systems.

[4] R. Mans, W. Aalst, van der, R. Vanwersch, Process mining in healthcare: evaluating and exploiting operational healthcare processes, SpringerBriefs in Business Process Management, Springer, Germany, 2015. https://doi.org/10.1007/978-3-319-16071-9.

[5] R.S. Mans, M.H. Schonenberg, M. Song, W.M.P. van der Aalst, P.J.M. Bakker, Application of process mining in healthcare – a case study in a dutch hospital, in: A. Fred, J. Filipe, H. Gamboa (Eds.), Biomedical Engineering Systems and Technologies, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 425–438.

[6] W.M.P. van der Aalst, Process Mining: Data Science in Action, Springer, 2016.

[7] P. Homayounfar, Process mining challenges in hospital information systems, in: 2012 Federated Conference on Computer Science and Information Systems (FedCSIS), 2012, pp. 1135–1140.

[8] E. Batista, A. Solanas, Process mining in healthcare: A systematic review, in: 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), 2018, pp. 1–6. https://doi.org/10.1109/IISA.2018.8633608.

[9] Y. Fontenla-Seco, A. Bugarín, M. Lama, Process-to-text: a framework for the quantitative description of processes in natural language, in: B.O. Fredrik Heintz, Michela Milano (Ed.), Trustworthy AI - Integrating Learning, Optimization and Reasoning, Springer, 2020, p. 6. https://doi.org/10.1007/978-3-030-73959-1_19.

[10] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, USA, 2000.

[11] M. Petre, Why looking isn't always seeing: Readership skills and graphical programming, Commun. ACM 38 (1995) 33–44.

[12] A.S. Law, Y. Freer, J. Hunter, R. Logie, N. McIntosh, J. Quinn, A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit, J. Clin. Monit. Comput. 19 (2005) 183–194.

[13] R.J. Almeida, M. Lesot, B. Bouchon-Meunier, U. Kaymak, G. Moyse, Linguistic summaries of categorical time series for septic shock patient data, in: Proc. 2013 IEEE Int. Conf. Fuzzy Systems, 2013, pp. 1–8. https://doi.org/10.1109/FUZZ-IEEE.2013.6622581.

[14] F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, C. Sykes, Automatic generation of textual summaries from neonatal intensive care data, Artif. Intell. 173 (7–8) (2009) 789–816.

[15] H. Leopold, J. Mendling, A. Polyvyanyy, Generating natural language texts from business process models, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7328, LNCS, 2012, pp. 64–79. https://doi.org/10.1007/978-3-642-31095-9_5.

[16] R.M. Dijkman, A. Wilbik, Linguistic summarization of event logs - A practical approach, Inf. Syst. 67 (2017) 114–125.

[17] R.S. Mans, W.M.P. van der Aalst, R.J.B. Vanwersch, A.J. Moleman, Process mining in healthcare: Data challenges when answering frequently posed questions, in: R. Lenz, S. Miksch, M. Peleg, M. Reichert, D. Riaño, A. ten Teije (Eds.), Process Support and Knowledge Representation in Health Care, Springer, Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 140–153.

[18] A. Ramos-Soto, A.J. Bugarín, S. Barro, J. Taboada, Linguistic descriptions for automatic generation of textual short-term weather forecasts on real prediction data, IEEE Trans. Fuzzy Syst. 23 (1) (2015) 44–57, https://doi.org/10.1109/TFUZZ.2014.2328011.

[19] N. Tintarev, E. Reiter, R. Black, A. Waller, J. Reddington, Personal storytelling: Using natural language generation for children with complex communication needs, in the wild, Int. J. Hum. Comput. Stud. 92–93 (2016) 1–16, https://doi.org/10.1016/j.ijhcs.2016.04.005.

[20] L.A. Zadeh, Fuzzy logic = computing with words, IEEE Trans. Fuzzy Syst. 4 (2) (1996) 103–111, https://doi.org/10.1109/91.493904.

[21] R.R. Yager, A new approach to the summarization of data, Inf. Sci. 28 (1) (1982) 69–86, https://doi.org/10.1016/0020-0255(82)90033-0.

[22] L.A. Zadeh, A prototype-centered approach to adding deduction capability to search engines-the concept of protoform, in: Proc. NAFIPS-FLINT 2002, 2002, pp. 523–525.

[23] A. Ramos-Soto, A. Bugarín, S. Barro, On the role of linguistic descriptions of data in the building of natural language generation systems, Fuzzy Sets Syst. 285 (2016) 31–51, https://doi.org/10.1016/j.fss.2015.06.019.

[24] H. Leopold, J. Mendling, A. Polyvyanyy, Supporting process model validation through natural language generation, IEEE Trans. Softw. Eng. 40 (8) (2014) 818–840, https://doi.org/10.1109/TSE.2014.2327044.

[25] H. van der Aa, H. Leopold, H.A. Reijers, Detecting inconsistencies between process models and textual descriptions, in: H.R. Motahari-Nezhad, J. Recker, M. Weidlich (Eds.), Business Process Management, Springer International Publishing, Cham, 2015, pp. 90–105.

[26] A. Wilbik, R.M. Dijkman, Linguistic summaries of process data, in: 2015 IEEE Int. Conf. Fuzzy Systems (FUZZ-IEEE), 2015, pp. 1–7.

[27] World Health Organization, https://www.who.int/health-topics/cardiovascular-diseases (Online; accessed in August 2021) (2021).

[28] G. Eveborn, H. Schirmer, G. Heggelund, P. Lunde, K. Rasmussen, The evolving epidemiology of valvular aortic stenosis. the tromso study, Heart (British Cardiac Society) 99. https://doi.org/10.1136/heartjnl-2012-302265.

[29] M.B. Leon, C.R. Smith, M. Mack, D.C. Miller, J.W. Moses, L.G. Svensson, E. M. Tuzcu, J.G. Webb, G.P. Fontana, R.R. Makkar, D.L. Brown, P.C. Block, R. A. Guyton, A.D. Pichard, J.E. Bavaria, H.C. Herrmann, P.S. Douglas, J.L. Petersen, J.J. Akin, W.N. Anderson, D. Wang, S. Pocock, Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery, N. Engl. J. Med. 363 (17) (2010) 1597–1607.

[30] R. Gomis, M. Mata Cases, D. Mauricio Puente, S. Artola Menéndez, J. Ena Muñoz, J. J. Mediavilla Bravo, C. Miranda Fernández-Santos, D. Orozco Beltrán, L. Rodríguez Mañas, C. Sánchez Villalba, J.A. Martínez, Aspectos metodológicos de los procesos asistenciales integrados (PAI), Revista de Calidad Asistencial 32 (4) (2017) 234–239.

[31] V. González, C. Peña, C. Neiro, D. López, Proceso asistencial integrado de estenosis aórtica, Tech. rep., Servivio de Cardiología del Complejo Hospitalario Clínico Universitario de Santiago (2021).

[32] M. Delgado, M.D. Ruiz, D. Sánchez, M.A. Vila, Fuzzy quantification: a state of the art, Fuzzy Sets Syst. 242 (2014) 1–30.

[33] A. Cascallar-Fuentes, A. Ramos-Soto, A. Bugarín-Diz, An experimental study on the use of fuzzy quantification models for linguistic descriptions of data, in: 24th European Conference on Artificial Intelligence, 2020, pp. 267–274.

[34] A. Ramos-Soto, J. Janeiro-Gallardo, A. Bugarín, Adapting SimpleNLG to spanish, in: 10th International Conference on Natural Language Generation, Association for Computational Linguistics, 2017, pp. 142–146.

[35] A. Augusto, R. Conforti, M. Dumas, M.L. Rosa, F.M. Maggi, A. Marrella, M. Mecella, A. Soo, Automated discovery of process models from event logs: Review and benchmark, IEEE Trans. Knowl. Data Eng. 31 (4) (2019) 686–705.

[36] A. Weijters, W. Aalst, Rediscovering workflow models from event-based data using little thumb, Integr. Comput.-Aided Eng. 10 (2003) 151–162.

[37] S.J.J. Leemans, D. Fahland, W.M. van der Aalst, Discovering block-structured process models from event logs - a constructive approach, in: Application and Theory of Petri Nets and Concurrency, Springer, 2013, pp. 311–329.

[38] B. Vázquez-Barreiros, M. Mucientes, M. Lama, Prodigen: Mining complete, precise and minimal structure process models with a genetic algorithm, Inf. Sci. 294 (2015) 315–333.

[39] P. Cariñena, A. Bugarín, M. Mucientes, S. Barro, A language for expressing expert knowledge using fuzzy temporal rules, in: Proceedings of the EUSFLAT-ESTYLF Joint Conference, 1999, pp. 171–174.

[40] A. Ramos-Soto, P. Martín-Rodilla, Enriching linguistic descriptions of data: A framework for composite protoforms, Fuzzy Sets Syst. 407 (2021) 1–26.

[41] K. Van Deemter, E. Krahmer, M. Theune, Real versus template-based natural language generation: A false opposition? Comput. Linguist. 31 (1) (2005) 15–24.

[42] E. Reiter, Nlg vs. templates, in: Proceedings of the Fifth European Workshop on Natural Language Generation, 1995, pp. 95–106.

[43] A. Gatt, E. Reiter, SimpleNLG: A Realisation Engine for Practical Applications, in: ENLG 2009 - Proceedings of the 12th European Workshop on Natural Language Generation, March 30–31, 2009, Athens, Greece, The Association for Computer Linguistics, 2009, pp. 90–93.

[44] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, E. Krahmer, Best practices for the human evaluation of automatically generated text, in: Proceedings of the 12th International Conference on Natural Language Generation, Association for Computational Linguistics, 2019, pp. 355–368.

[45] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Quart. 13 (3) (1989) 319–340.