Modular Petri Net Modeling of Healthcare Systems - draft -

Cristian Mahulea, Liliana Mahulea, José Manuel García Soriano and José Manuel Colom

April 25, 2017

Abstract

This paper presents a modular approach for modeling healthcare systems using Petri Nets. It is shown that a healthcare system can be constructed by different modules whose inputs and outputs are connected according to their geographical location. Each module can be modeled in two phases: (1) obtain the sequences of treatments and cares received by a patient in the case of a particular disease/condition, and (2) add the resources necessary to perform the previous sequences. The global model is obtained by fusion the inputs and outputs of the modules and by adding information on the patients. The constructed modules together with the resources are Petri nets belonging to a new subclass called *Healthcare Petri nets* that is proved to have equivalent behavior with S^4PR nets, a well-known class of Resource Allocation Systems. This allows us to apply the structural results already existing in the literature for S^4PR to the context of healthcare systems. In order to illustrate the results, a case study of a Public Healthcare Area in Zaragoza is considered as a use case.

Published as: C. Mahulea, L. Mahulea, J.M. Garcia-Soriano and J.M. Colom "Modular Petri Net Modeling of Healthcare Systems," Flexible Services and Manufacturing Journal, Special Issue on Analysis, Design and Management of Health Care Systems, 2017.

1 Introduction

Healthcare system (HS) is a (public) service of high relevance in all countries being one of the most complex services because of the large number of stakeholders and the nonlinear and unpredictable behavior of element interactions. Small variations in the interaction among elements usually affect the behavior of the system as a whole. This paper proposes to analyze the management of HS through the behavior of the deployed medical protocols and the available resources used for their execution. In general, HS contain a huge amount of medical protocols to describe the activities and treatments that the medical staff should know and follow for many (almost all) medical problems that a patient may have. The design of a medical protocol needs a phase of verification and validation before it becomes operative into the real system. Verification and validation activities aim to check the correctness of the design (or at least the satisfaction of the requirements stated at the beginning of the design). If risky situations are detected, hints for alternate solutions must be provided. Risky situations are, for example, the presence of blocking steps (i.e., deadlocks) or sets of non-progressing activities (i.e., livelocks) within a medical protocol due to the absence of required resources. In real HS, deadlocks are solved "in situ" by medical personnel detecting the situation, identifying the affected actors and resources, and recovering the situation using ad-hoc solutions that require to be studied a priori in order to take informed decisions.

This paper considers the construction of formal models to reason about the sequences of treatments that patients suffering a particular disease should follow, and the influence of the resources needed to carry out these sequences. The set of sequences of treatments is first described by mathematical models such as Petri Nets, while in a second step, resources that are necessary for each activity will be added to the model. Then a particular class of nets is obtained with a well-defined structure that can be used for formal analysis. This methodology has been successfully applied to the modeling of manufacturing systems with Petri nets, in particular Resource Allocation Systems (RAS) in manufacturing [30]. Besides correctness analysis, medical managers need to assess non-functional properties. They would like to consult information on "how well" a medical protocol works by focusing on the assessment of quantitative properties related to performance, dependability and survivability. Examples of performance metrics are the length of waiting lists, the patient waiting time for a determined procedure, morbidity rate, mortality rate, satisfaction rate, etc. Performance analysis entails evaluating such metrics, detecting bottlenecks and, possible, removing them by finding better design solutions. The Petri net model that we propose can assist the medical manager to evaluate the performance of medical protocol by using simulations of the model and exploitation of the structural information of the model itself.

The number of medical protocols that should be implemented and modeled depends on the properties to be studied. However, to reduce the complexity of the model three medical structures of the HS are identified: *primary care team*, *specialized attending center* and *hospital*. It is shown that each of these structures can be seen as a *module* with some inputs and some outputs and the global model is obtained by composing these modules based on the common inputs and outputs.

The main contributions of this work are the followings:

- Introduction of a methodology for the analysis and management of a HS based on the intensive use of formal models (in this case Petri nets) in the different phases of the life cycle of this kind of systems: specification, design, synthesis, implementation, monitoring, observation, etc.
- Proposal of a systematic procedure for the construction of the formal models for HS based on three pillars:
  - *Abstraction of the HS.* Retention of characteristics and elements of the HS relevant for the understanding and study the behavior and control of the flow of patients in the HS.
  - *Reduced repertoire of constructive components of HS models.* Identification of abstraction objects of the HS around the concept of medical protocol (active part that follows the flow of patients) and resources (as facilities needed to carry out an activity of the medical protocol for a patient).
  - *Modular composition of components.* The modular construction of the formal model allowing to manage the intrinsic complexity of this kind of systems, and proceed in an incremental way in the understanding and studying of the system. The modules will be, for example, the medical protocols of parts of the HS.
- The transposition of the rich body of concepts, methods and tools available in the context of PN to the application domain of the management and control of the flow of patients in the HS. This requires the translation of properties, characteristics, parameters and concepts of the HS domain to the context of Petri nets; and on the other hand, the solutions obtained from the formal model need to be translated to the HS domain. In this sense, one of the main contributions is the interpretation of the structure of these systems and the analysis and synthesis of well-behaved structures beyond the actual values of the flow of patients.

Note to practitioners: The work developed in this article arose from the specifications obtained from the healthcare system, where the protocols discussed here are deployed. From these specifications, the methodology presented has been developed, and the formalisms have been adapted to capture the aspects whose study has been object of the work. The results have been contrasted with the ones observed by the professionals of the healthcare system. Nevertheless, the implantation of these techniques within the medical centers (to be used by the professionals involved in the execution of these protocols, as well as the managers dedicated to the decision making in the healthcare systems) requires the development of a User Interface that saves the semantic gap between the world of healthcare systems and the formal models used to represent and control these systems. We believe that end users should use Graphical User Interfaces or modeling languages as *Unified Modeling Language* (UML), whose semantics can be easily expressed through the Petri nets presented here. In the same way, the results obtained must be interpreted in terms of the domain of healthcare systems. This is one of the future work that we plan to develop.

The paper is organized as follows. Sec. 2 presents related works. In Sec. 3 the basic components for the construction of the Petri net in a modular way are introduced:

the medical protocol and the resources that are needed (shared between protocols). Sec. 4 defines the Petri net models and provides a proof that the resulted model has an equivalent behavior with S^4PR nets. Sec. 5 interprets the PN models in the health domain while Sec. 6 applies the methodology to a HS. Some properties of PN are interpreted in healthcare domain in Sec. 7 while in Sec. 8 some conclusions are given.

2 Related works

In this section, some of the main methods dealing with HS are described. However, this is not an exhaustive list since "*undertaking a review of modeling and simulation in health care is without doubt a Herculean task*" [6]. Moreover, it is described the relation of the contribution of this work with respect to the state of the art in HS.

As it was mentioned, HS are complex and nonlinear dynamic systems making that the one of the most used approach for their analysis and design to be *dynamic simulation*, being the most used methods:

- *System Dynamics* [15] is a simulation modeling methodology that can be applied to HS (see [17] and the references herein). It is very useful when the number of patients is big since they are not treated individually but assuming flows of patients. It can be used for policy analysis and design.

- *Discrete event simulation* models [10, 19, 23, 29] have been extensively used for HS analysis and evaluation. These models are mainly used to characterize and analyze *specific* processes and the use of resources.

- *Agent-based modeling* [24] is a simulation method for modeling dynamic, adaptive, and behavioral systems such as HS. They have been applied to model natural disasters, such as epidemic infectious diseases [16, 22, 27].

Statistical modeling and analysis are also used in HS [7] as a method to study different properties and in particular for taking clinical decisions, e.g., study the effectiveness of a drug. Also, these methods can be used for management being possible to evaluate statistically the evolution of a designed policy.

Models used for *software systems* have been also adapted, as for example UML [2, 4, 12] and Business Process Model and Notation (BPMN) [25]. However, these are semi-formal models being used manly to obtain executable specifications of the system. In some cases, for analysis purposes, the semantics of UML is defined in terms of (colored) Petri nets by using transformation techniques. Nevertheless, analysis techniques for Colored Petri Nets [18] are based on state space exploration/construction that, in general, is computationally intractable. In [31] a colored PN model has been proposed to model medical protocols while in [3] the same type of models is used to model the spread of directly transmitted infectious diseases. But the main analysis techniques for colored Petri nets are simulations and state-space exploration.

Operational Research have also been used for the analysis and simulation of different problems in the HS. For example, in [20, 21] a stochastic mathematical programming model for planning surgeries is presented. The solution is obtained using

the Monte Carlo simulation and Mixed Integer Programming problems. In [8] Integer Linear Programming is used for operation planning of elective patients.

The main drawback of the previous methods is the fact that the results of the evaluation are valid for a specific scenario and it is difficult to extend the methodology to extract conclusions for families of scenarios or, in other words, conclusions derived from the architecture of the HS or its structure. We propose the use of PN, but this is not new because have been used in HS domain. For example, a three level strategy to design at the tactical level hospital departments is proposed in [13], an architecture for an Integrated Systems to manage the operational level at healthcare at home is presented in [14] while continuous Petri nets are used in [11] to optimize an emergency department by relaxing the integer variables to real ones. The contributions of the approach we propose with respect to those that use PN is the abstraction method and the primitive components to be identified that allows to construct models belonging to a subclass with a powerful theory that can be used for analytical purposes. Another advantage is derived from the modularity of the methodology that makes easier the construction of models and its change under an analysis based on the question *what if*?

3 Abstraction of the Healthcare System

The new models introduced in this work permit to study a particular facet of the HS: the flow of patients throughout the HS and the influence of the resources needed to activate the protocols over this flow. To obtain a reasonable model we need to abstract the HS retaining only those aspects that are relevant for the study of these properties. So, for example, from a particular medical treatment we are only interested in aspects such as: phases of the treatment, causal dependences between the phases, resources needed to realize each phase, number of patients that can be attended, time consumed in each phase.

Subsection 3.1 presents the structure of a HS that will be considered as case study while the next two subsections present the two basic pillars used in this abstraction process previously to the construction of the model. The first one is the flow of patients receiving the corresponding treatments while the second one is related to the resources that a patient is using in order to progress inside the HS.

3.1 High-level Structure of a Healthcare System

This subsection describes, as an example, the structure of the Public Spanish National HS being used as a case study for the proposed modeling methodology. Its general organization consists in *Primary Healthcare* and *Specialized Healthcare*. The Primary Healthcare is the basic and, in general, the main entrance of a patient in the HS. Moreover, the Specialized Healthcare is accessed by referral from Primary Healthcare or from an Emergency Department. In general, this Specialized Healthcare consists in complex and costly diagnostic and therapeutic resources.

Basic health zones are the smallest units of the organizational structure of the HS. They are usually organized around a single *primary care team* (PCT) which exercises the gatekeeper function. Regarding patient choice, the possibility to choose a specialist

and a hospital is relatively less developed, compared to the PCT. In any case, access to specialist healthcare requires referral from a primary doctor.

Each *Healthcare Area* of the HS is covering a population of around 200,000 inhabitants and it is composed by many basic health zones and one hospital. In general, the hospital has an Emergency Department where the patients can be directly attended in the case of an emergency, without passing through the PCT. In order to not overload the hospital, each healthcare area has one or more Specialized Attending Centers (SAC). To these centers arrive the patients that should be consulted by a specialist but, in principle, should not be hospitalized and don't need urgent attention.

Finally, an ambulance service is used to attend home emergencies, treat and transport patients. This service includes basic support (Conventional Ambulance mainly used to transport the patients) and advanced support (Emergency Medical Unit that can diagnose, treat and stabilize a patient before it's transport to a hospital). The differences between these two ambulances is the human and equipment resources.

3.2 Treatment of Flows of Patients by Medical Protocols

The flow of patients throughout the set of the sequential treatments where in each step a patient receives a care in order to obtain at the end their cure. The set of sequential treatments that can be followed for a particular pathology are embedded in the so-called *medical protocol*. Internally, a protocol can have alternative cares for the same disease, but a patient receives sequentially the cares when choices are solved.

A group of patients affected by the same disease follow the same protocol. When we are studying a flow of patients in order to take decisions in the management of the HS, the parameters that characterize this flow are, for example, the maximum number of patients that can be simultaneously served by the established protocol, the waiting time to obtain a treatment, the mean size of the waiting queues to obtain a treatment, the resources needed to support the treatments of a given population, and so on. Then, we can offer to the managers of the HS real indicators with data about how can a protocol be optimized.

Our abstraction process retains only those aspects of the medical protocol concerning the HS to be analyzed. This means that we will represent the situations of patients when they receive a treatment inside the structure of the HS or when they should be transferred between different locations inside the HS. Therefore, a sequence in a protocol will be the interleaving of treatments and transfers between treatments expressing the sequence of activities applied to the flow of patients belonging to the HS under study.

On the other hand, the treatments in a protocol respect the principle of locality. This means that we can group sets of treatments for different diseases in modules that can be identified with the different subsystems of the HS presented previously: PCT, SAC and Hospital. Each one of these subsystems has a well defined interface representing the different input and output points for the patients accessing to one of these subsystems for receiving a part of their treatments. The connection of these modules will be realized by a transport activity allowing to transfer patients between different parts of the HS structure.

3.3 Resources shared between treatments in the medical protocols

Two types of resources needed to provide a treatment of a given protocol can be identified: human and material. In principle, all treatments belonging to a medical protocol require the use of some shared resources for which patients have to compete because of the limited number of copies available for each resource type. The resources represented in our models have three main characteristics: 1) *exclusivity* - they are needed to carry out the treatment and used by a single patient in an exclusive way, 2) *limited availability*, and 3) *conservativity* - they cannot be created nor consumed when they are used by the patients following a protocol.

For example, we can consider as a resource a x-ray machine for making x-ray radiographs. This test is very common in HS and appears in many medical protocols. We can have several identical copies (functionally equivalent) of x-ray machines and each one can be interchangeable used to satisfy a request of a patient. Another example of resources can be found in almost all PCT with a continuing care service (similar to an Emergency Department for level IV-V conditions, i.e., easy to diagnose and treat) where a group of medical professionals (doctors, nurses, etc.) are considered as resources necessary for the implementation of the protocol treatments, and are shared by patients who are following this protocol.

If the analysis of a protocol is done in an isolated way then is possible only to determine consistency, soundness or causal relations of the protocol itself. If one wants to study the interrelationships between different protocols of the same HS or the viability of a protocol itself, this requires the explicit representation of resources. Therefore, the analysis/optimization to be done concerns complete subsystems or interaction of the different subsystems via shared resources. For example, the response of the HS in the case of an abnormal situation is very important and should be studied before the situation appears (e.g., stability of the system under an epidemic flu). In this case, stress tests can be used on a model where the resources should be explicitly represented: medical personnel and stuff.

In this paper, the resources that can be used are private (specific) to each module, i.e., they can be used only by one module. Each module will be composed by a number of medical protocols modeling operations that should be done sequentially in the case of a particular situation. After defining all important protocols, the resources will be assigned to the treatments that require them. For example, a visit with the primary doctor will require a human resource (in this case a doctor). This methodology is very similar to those used in modeling manufacturing systems [9] where they are considered/viewed as Resource Allocation Systems (RAS).

4 Petri net models for RAS view of HS

This section introduces the basic notions of Petri nets (see [26, 28] for a gentle introduction). The PNs are the formal models that will be used to describe the structure and the behavior of the HS.

A *Petri net* (PN) is a tuple $\mathcal{N} = \langle P, T, \boldsymbol{Pre}, \boldsymbol{Post} \rangle$ with P and T two non-empty disjoint sets of places and transitions; $\boldsymbol{Pre}, \boldsymbol{Post} \in \mathbb{N}^{|P| \times |T|}$ the pre and post inci-

dence matrices defining the arc weights from places to transitions and from transitions to places, respectively.

For $h \in P \cup T$, the sets of its input and output nodes are denoted as ${}^\bullet h$ and $h^\bullet$, respectively. Let p_i, $i = 1, \ldots, |P|$ and $t_j, j = 1, \ldots, |T|$ denote the places and transitions. Each place can contain a non-negative integer number of tokens, this number represents the marking of the place. The distribution of tokens in places is denoted by $\boldsymbol{m}$ while $\boldsymbol{m}(p_i)$ or simply m_i is the marking of place p_i. The initial token distribution, denoted by $\boldsymbol{m}_0 \in \mathbb{N}^{|P|}$, is called the initial marking of the net. A PN with an initial marking is a PN system $\langle \mathcal{N}, \boldsymbol{m}_0 \rangle$. A PN $\mathcal{N}$ is called pure if ${}^\bullet p_i \cap p_i{}^\bullet = \emptyset$ for all places $p_i \in P$. $\mathcal{N}$ is called strongly connected if there exists a path from each node $h_1 \in P \cup T$ to every other node $h_2 \in P \cup T$.

A transition $t_j \in T$ is enabled at $\boldsymbol{m}$ iff for all $p_i \in {}^\bullet t_j$, $m_i \geq \boldsymbol{Pre}(p_i, t_j)$. An enabled transition t_j can fire leading to a new state $\boldsymbol{m'} = \boldsymbol{m} + \boldsymbol{C}(\cdot, t_j)$, where $\boldsymbol{C} = \boldsymbol{Post} - \boldsymbol{Pre}$ is the *token flow matrix*, or simply *incidence matrix*, and $\boldsymbol{C}(\cdot, k)$ is its k^{th} column. It will be said that $\boldsymbol{m'}$ is a *reachable marking* that has been reached from $\boldsymbol{m}$ by firing t_j. If $\boldsymbol{m}$ is reachable from $\boldsymbol{m}_0$ through a finite sequence of transitions $\sigma = t_{i1} t_{i2} \ldots t_{ik}$, the state (or fundamental) equation is satisfied: $\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{C} \cdot \boldsymbol{\sigma}$, where $\boldsymbol{\sigma} \in \mathbb{N}^{|T|}$ is the *firing count vector*, i.e., its j element is the cumulative amount of firings of t_j in the sequence σ.

In order to study quantitative properties of the system, time should be introduced into the model. There are different possibility in PN, the most common one is to assign time delay to transitions. The set of transitions is partitioned into two subsets: (a) *immediate transitions* (represented as filled rectangles) that fires in zero time units; and (b) *timed transitions* (represented as white rectangles). If such a transition is enabled then it is fired after a random time delay. For simplicity, in this paper we consider random delays with negative exponential probabilistic density functions (pdf) but any pdf can be used.

The support of an integer vector $\boldsymbol{v}$, denoted by $||\boldsymbol{v}||$ is the set of all indices i such that $\boldsymbol{v}(i) \neq 0$. A left annuler of the incidence matrix (i.e., a vector $\boldsymbol{y} \in \mathbb{N}^{|P|}$, $\boldsymbol{y} \gneq 0$, such that $\boldsymbol{y} \cdot \boldsymbol{C} = \boldsymbol{0}$) is called a P-semiflow. A P-semiflow allows to obtain a weighted token conservation law defined on the places belonging to the support of the P-semiflow. A right annuler of the incidence matrix (i.e., a vector $\boldsymbol{x} \in \mathbb{N}^{|T|}$, $\boldsymbol{x} \gneq 0$, such that $\boldsymbol{C} \cdot \boldsymbol{x} = \boldsymbol{0}$) is called a T-semiflow. The existence of repetitive behaviors in a Petri Net, i.e., to recover a marking $\boldsymbol{m}$ after the firing of sequence of transitions σ, means that there exists a T-semiflow $\boldsymbol{x} = \boldsymbol{\sigma}$.

The PNs can be classified according to their structures. A *state machine Petri net* is a Petri net where each transition has at most one input and at most one output place, i.e., for all $t \in T$, $|{}^\bullet t| \leq 1$ and $|t^\bullet| \leq 1$. A well known class used in Resource Allocation Systems is S^4PR [30]. Inspiring on S^4PR class, a new class of PN for modeling HS is defined. First, a healthcare module PN is given and after that the definition of the full model is obtained by fusion of the common nodes.

Definition 4.1 *A* basic health Petri net module *(BHM) is a pure PN system* $\langle \mathcal{N}, \boldsymbol{m}_0 \rangle$ *with* $\mathcal{N} = \langle P, T, \boldsymbol{Pre}, \boldsymbol{Post} \rangle$ *where:*

1. $P = P^{in} \cup P^A \cup P^R \cup P^{out}$ *is a partition such that:*

(a) [input places] $P^{in} = \{p_1^{in}, p_2^{in}, \ldots, p_{|P^{in}|}^{in}\}$ *such that* $|P^{in}| > 0$ *and* $|^\bullet(p_i^{in})| = 0$ *for all* $i = 1, \ldots, |P^{in}|$*;*

(b) [output places] $P^{out} = \{p_1^{out}, p_2^{out} \ldots p_{|P^{out}|}^{out}\}$ *such that* $|P^{out}| > 0$ *and* $|(p_i^{out})^\bullet| = 0$ *for all* $i = 1, \ldots, |P^{out}|$*;*

(c) [activity places] $P^A = P^T \cup P^W$ *where (i)* P^T *is the set of treatment places* $P^T = \{p_1^t, p_2^t, \ldots, p_{|P^T|}^t\}$ *such that* $|P^T| > 0$ *and (ii)* P^W *is the set of waiting places* $P^W = \{p_1^w, p_2^w, \ldots, p_{|P^W|}^w\}$*. All activity places are initially empty, i.e.,* $\boldsymbol{m}_0(p_i^a) = 0$ *for all* $p_i^a \in P^A$*;*

(d) [resource places] $P^R = \{r_1, r_2, \ldots, r_{|P^R|}\}$, $|P^R| > 0$.

2. [protocol subnet] $\langle \bar{P} = P \setminus P^R, T, \boldsymbol{Pre}(\bar{P}, \cdot), \boldsymbol{Post}(\bar{P}, \cdot), \boldsymbol{m}_0(\bar{P}) \rangle$ *is a forest of acyclic state machines, i.e.,* $\forall t \in T$, $|^\bullet t \cap \bar{P}| = 1$ *and* $|t^\bullet \cap \bar{P}| = 1$*; such that* P^{in} *are the unique source places (without input transitions) and* P^{out} *are the unique sink places (without output transitions); and all places of* $\bar{P}$ *are unmarked,* $\boldsymbol{m}_0(\bar{P}) = 0$*;*

3. *For each* $r \in P^R$ *there exists a unique minimal P-semiflow,* $\boldsymbol{y}_r \in \mathbb{N}^{|P|}$*, such that* $\{r\} = ||\boldsymbol{y}_r|| \cap P^R$*,* $\boldsymbol{y}_r(r) = 1$*,* $\left(P^{in} \cup P^{out} \cup P^W\right) \cap ||\boldsymbol{y}_r|| = \emptyset$*, and* $P^T \cap ||\boldsymbol{y}_r|| \neq \emptyset$*. For all* $r \in P^R$*,* $\boldsymbol{m}_0(r) \geq \max\{\boldsymbol{y}_r(p) | p \in ||\boldsymbol{y}_r|| \setminus \{r\}\}$

4. $P^T = \bigcup_{r \in P^R}(||\boldsymbol{y}_r|| \setminus \{r\})$

5. [resource allocation] *For any* $t \in T$*, if* $\{t\} \cap \left(P^R\right)^\bullet \neq \emptyset$ *then* t *is an immediate transition.* ■

In the previous definition we specify that the model of a module is composed by a set of acyclic state machines representing the protocols inside a healthcare module that can share some resources that are private to the healthcare module. The initial marking of the state machines is zero because at the beginning there are no patients receiving treatments and the waiting queues are emptied. The other important specification is that a resource place belongs to a P-semiflow containing the resource itself, plus some places belonging to P^T (treatment places) that are the holders of the resources if a patient is receiving a treatment in these places. This represents the conservative use of resources because the P-semiflow gives rise to a marking invariant saying that the resources are available or they are used by the holders but its number is constant. Observe that the coefficient of the holder place in the P-semiflow of the resource r represents the number of copies of r that a patient, receiving the treatment modeling by this place needs. Finally, the initial marking of the resource places is the minimum one needed to guarantee that each protocol can be followed in isolation. That is, if we put a lower number of resources than the one specified in the definition, we can guarantee that there exists a protocol that cannot be completed because it needs resources that will never be available. The last condition ensures that execution time of an activity is independent to resource allocation.

Remark 4.2 *Let us highlight three main differences between a* BHM *(Def. 4.1) and a* S^4PR*:*

- *The process subnet is composed by strongly connected state machines in the case of S^4PR, each cycle contains one idle place belonging to a set P_0. The protocol subnet of a BHM is composed by a forest of state machines, that are not strongly connected and without cycles. However, the forest will be converted into a strongly connected component with the inclusion of a place modeling the population;*

- *In S^4PR, each place belonging to the process subnet minus the idle places belongs to the support of at least one P-semiflow associated to a resource. In the case of BHM, the set of waiting places P^W (that belong to the protocol subnet) contains places which are not belonging to the support of any P-semiflow associated to the resources;*

- *In BHM, resources are allocated through immediate transitions while in S^4PR the model has no time information.* ■

During the modeling of a medical protocol may be necessary to execute different operations in order to continue with the protocol. For example, the doctor can ask for different medical tests to be executed in different physical locations. It may seem that these are concurrent operations. In PN methodology, concurrent operations are modeled with a fork[1] modeling the starting of the $|t^\bullet|$ treatments in parallel followed by a join[2]. If these constructions are used, the model will not be anymore a BHM since the subnet will not be anymore a state machine. However, in the case of the medical protocol of a patient with a particular disease the interleaving semantics could be considered. This happens because the patient can do only one operation at a time moment and the decision on which one is executed first depends only on the existence of the corresponding resource.

Let us consider a small example where the doctor, in order to put a diagnostic, is sending the patient for a blood analysis and also for an ECG. Obviously, if the order in which these analysis should be performed is specified, the sequential operations are easy to model. Let us assume that the patient should perform first the analysis for which the required resource is available. The PN model is shown in Fig. 1(a). Assume that for blood analysis a nurse is necessary for extraction. This resource is modeled by place r_1 where n_1 is the number of nurses available. Assume also that for the ECG it is necessary the ECG machine modeled by place r_2. Notice that the transitions t_2 and t_3 are enabled, i.e., can be fired, depending on the marking of r_1 and r_2. If $\boldsymbol{m}(r_1) = 0$, transition t_2 cannot be fired while if $\boldsymbol{m}(r_2) = 0$, transition t_3 cannot be fired. The sequences $p_2 \to p_4 \to p_6$ and $p_3 \to p_5 \to p_7$ model the two possible orders in which patient arrived at p_1 can take the analysis.

In order to get the global model, it is necessary to obtain the modules corresponding to its structure and then, connect the inputs and outputs by fusing the common places. Finally, the information of the healthcare area population is necessary to be added. Formally,

[1] t is a fork transition if $|t^\bullet| > 1$, i.e., t has more than one output place.
[2] t is a join transition if $|{}^\bullet t| > 1$, i.e., t has more than one input place.

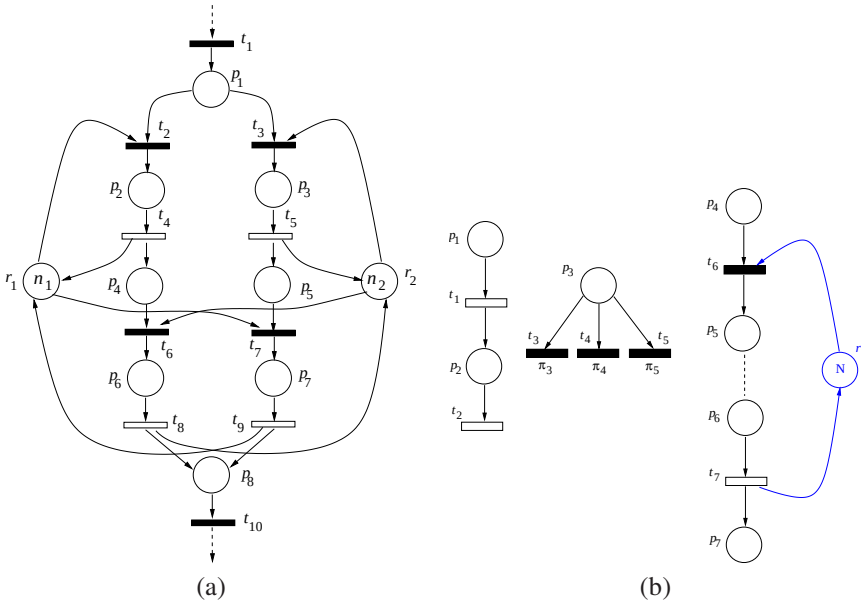

Figure 1: (a) Interleaving semantics. (b) Basic Petri net elements of a protocol.

Definition 4.3 *A* healthcare Petri net *(HPN)* $\mathcal{N} = \langle P, T, \boldsymbol{Pre}, \boldsymbol{Post} \rangle$ *is a PN system composed by:*

1. [Set of BHM modules] $\mathcal{N}_i = \langle P_i^{in} \cup P_i^A \cup P_i^R \cup P_i^{out}, T_i, \boldsymbol{Pre}_i, \boldsymbol{Post}_i \rangle$, $i = 1, \ldots, q$;
2. [Population place] p_P *such that* $\boldsymbol{m}_0(p_P) = n_p$;
3. [Node integration] $P = \{p_P\} \cup \left(\bigcup_{i=1,\ldots,q} \left(P_i^{in} \cup P_i^A \cup P_i^R \cup P_i^{out} \right) \right)$; $T = \bigcup_{i=1,\ldots,q} T_i$;
4. [Terminal places] *All places* $\left(\bigcup_{i=1,\ldots,q} P_i^{out} \right) \setminus \left(\bigcup_{i=1,\ldots,q} P_i^{in} \right)$ *are merged with the place* p_P.
5. [Entrance places] *For each* $p_e \in P_i^E = \left(\bigcup_{i=1,\ldots,q} P_i^{in} \right) \setminus \left(\bigcup_{i=1,\ldots,q} P_i^{out} \right)$ *there exists a transition* t_e *such that* $t_e \in p_P{}^\bullet$, ${}^\bullet t_e = \{p_P\}$, $t_e{}^\bullet = \{p_e\}$ *and* ${}^\bullet p_e = \{t_e\}$.
6. [Protocol subnet] *The subnet generated by* $\langle P' = P \setminus \left(\bigcup_{i=1,\ldots,q} P_i^R \right), {}^\bullet P' \cup P'^\bullet \rangle$ *is a strongly connected state machine such that every cycle contains* p_P.

■

Therefore, a HPN is composed by a set of BHM modules which are connected by fusing the common inputs and outputs (step 3) and by connecting them using a place p_P

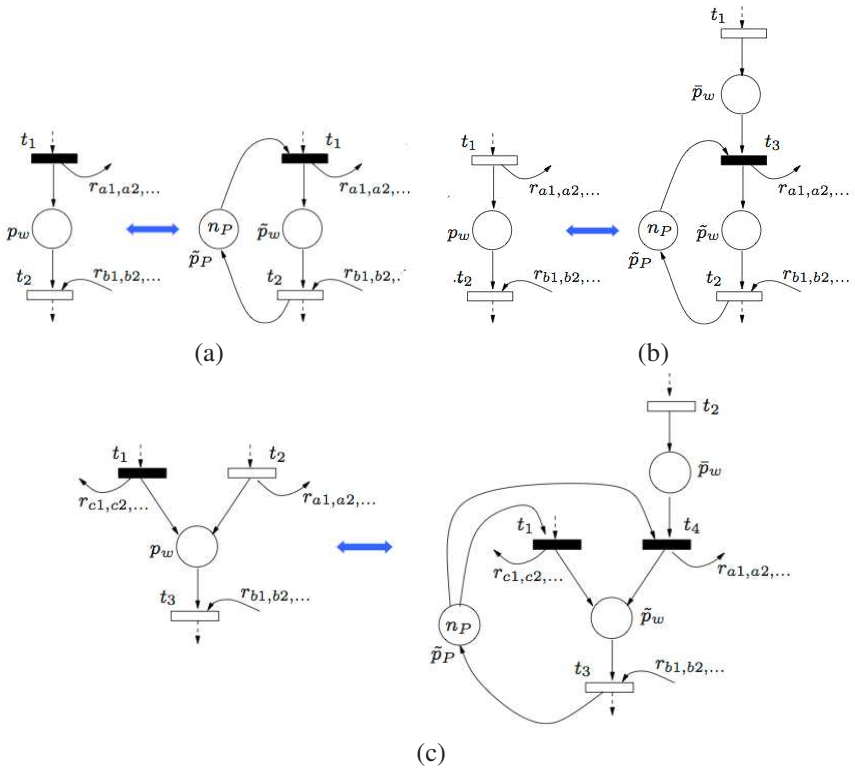

Figure 2: Eliminating waiting place p_w in a HPN.

which is modeling the healthcare area population (step 2). The outputs of the modules that are not inputs into other modules are merged with place p_P (step 4) modeling a patient that has been recovered or has been released home. Finally, the protocol subnet is a strongly connected state machine (step 6) and the modules are connected in a such way that no internal cycles exist, otherwise condition in step 6 is not satisfied.

It is easy to prove that a HPN system has an equivalent behavior with a S^4PR net system.

Proposition 4.4 *Let $\mathcal{N}$ be a HPN system composed by q BHM modules. The behavior of $\mathcal{N}$ is equivalent to an S^4PR net system.*

Let $\mathcal{N}$ be a HPN and we construct the S^4PR denoted $\mathcal{N}'$ with an equivalent behavior.

- *The set of idle places of $\mathcal{N}'$ is the place p_P, i.e., $P'_0 = \{p_P\}$;*
- *The set of resources places is $P'_R = \left(\bigcup_{i=1,\ldots,q} P_i^R\right) \cup \{\tilde{p}_P\}$ where P_i^R is the set of resource places of module i while $\tilde{p}_P$ is a* virtual resource place *being also an* implicit place[3]. *The initial marking of $\tilde{p}_P$ will be equal with the initial marking*

[3]An implicit place is a place that by its removal the behavior of the net is not changed.

of population place p_P, i.e., $\boldsymbol{m}_0(\tilde{p}_P) = \boldsymbol{m}_0(p_P) = n_P$;

- *The set of process places is $P'_S = \bigcup_{i=1,\ldots,q} \left(P_i^T \cup P_i^E \cup \bar{P}_i^W \cup \tilde{P}_i^W\right)$, where: (a) P_i^T is the set of treatment places of module i hence are holders of resource places of module i, i.e., P_i^R; (b) P_i^E is the set of entrance places of module i that can be easily done holder of the virtual resource $\tilde{p}_P$; and (c) sets $\bar{P}_i^W$ and $\tilde{P}_i^W$ are obtained from the set of waiting places of module i, i.e., P_i^W, by using the following transformation.*

Let $p_w \in P_i^W$ be a waiting place of module i. Fig. 2 shows the corresponding places $\tilde{p}_w$ and/or $\bar{p}_w$ that should be added in the S^4PR net depending on the time assigned to its input transitions. In Fig. 2(a), p_w has only one immediate input transition t_1 and will be transformed into the place $\tilde{p}_w$ holder of the virtual resource $\tilde{p}_P$. If p_w has a timed input transition, Fig. 2(b) shows both places $\bar{p}_w$ (holder of resources $r_{a1,a2,\ldots}$ that are released at t_1 in the original net) and $\tilde{p}_w$ holder of virtual place $\tilde{p}_P$. Moreover, if p_w has one immediate and one timed transition, transformation in Fig. 2(c) should be used. It is easy to see that the procedure can be generalized for any set of input transitions in p_w.

After this strategy, all waiting and entrance places are transformed to holders of at least one resource.

- *The set of transition is $T' = T$ together with transitions that eventually have been added when waiting places are transformed into process places;*

- *The protocol subnets of S^4PR will corresponds to the protocol subnets of HPN (where the waiting places are transformed).*

Notice that the last two conditions in both definitions are similar and the conclusion holds. □

5 Interpreting PN models in the HS framework

Any medical center has its own medical protocols (some may be common with other centers, inspired from national or international guides) but, in principle, adapted to the resources of the center. It is logical that at certain moment the same resource may be needed in different protocols at the same time. This locality principle in the application of the protocols leads to group them in PCTs, in hospitals or in the SACs that constitute the basic modules composing a Healthcare Area. A different model is proposed for each structure in this section.

In each module several medical protocols are (partially) deployed. A protocol is associated to a disease and it is a sequence of activities (medical tests, treatments or waiting queues) a patient should follow to be treated. A patient will be modeled by means of a token that flows through the net representing the protocol. A given activity will be modeled by means of a place where the tokens contained in that place represent the patients that are performing the activity (receive the treatment or wait for the

next treatment). The patient starts an activity when an input transition of the place representing the state *following the activity* occurs and moves the token representing the patient inside the place. The patient ends an activity when an output transition of the place representing the activity occurs and moves the patient to the following activity (the output place of the transition has just occurred). In Fig. 1(b), the leftmost PN represents two sequential activities by means of the places p_1 and p_2. The activity in place p_1 ends when the transition t_1 occurs and moves the token from the place p_1 to the place p_2. Activity in place p_2 ends when transition t_2 occurs. Observe that inside the places p_1 and p_2 one can have simultaneously several patients performing concurrently the activity. The number of patients in a given activity place p_1, is the marking of that place denoted by $\boldsymbol{m}(p_1)$. The set of all operation places is denoted by P^A in Def. 4.1.

Each transition in a PN representing a protocol has an unique input activity place (from P^A) and an unique output activity place (from P^A), because the transition represents the movement of a single patient from an activity in the protocol to the next one (we do not model the death or the birth of a patient or a translation to a different health area). This means that the net representing the protocol belongs to the subclass of Petri Nets named State Machines. Therefore, if T is the set of transitions of the net of a protocol, the following holds: $\forall t \in T, |{}^\bullet t \cap P^A| = |t^\bullet \cap P^A| = 1$. These transitions can have associated a time delay. This time represents a duration of the occurrence of the transition that we will interpret as the time consumed by the patient in the activity associated to the input place of P^A to the transition.

In many cases, the next activity is not unique and the choice is based on the results of the current operation. For example, after a visit to the primary doctor (p_3 in fig. 1(b)) depending on the diagnostic, the doctor could send the patient home (occurrence of t_3), send the patient to the SAC to be seen by a specialist (occurrence of t_4), or send it to the Emergency Department (occurrence of t_5). In our model we do not represent, in an explicit way, the conditions to take one of the options, instead of this we associate an occurrence probability to each one of the transitions. These probabilities are obtained from a statistical analysis of the historical application of the protocol. Obviously, the sum of the probabilities of all output transitions of an activity place should be equal to one, in Fig. 1(b), $\pi_3 + \pi_4 + \pi_5 = 1$, where π_i represents the probability that transition t_i occurs when the 3 transitions of the conflict are enabled. Note that in this case the transitions are represented by filled rectangles instead of white rectangles since these transitions will occur instantaneously, i.e., they will have zero time delay associated.

Another important basic element is the resource that is being used, either human or material. Each type of resource is modeled by a place, e.g., r in Fig. 1(b). The marking of r, denoted by N represents the number of available instances/copies of that resource type. In this work we only consider resources that are used in a *conservative* way (it is not possible to create or to consume a resource in the execution of protocols) and they should be correctly assigned and released in the PN model. Let us assume that a resource of type r is needed during the treatments from p_5 to p_6. This is simply modeled by connecting the place r with the input transition of p_5 (the resource is assigned/allocated) and an arc connecting the output transition of p_6 with r (the resource is released). The set of resource places will be denoted as P^R.

During the modeling process we adopt the abstraction of the *patient flow*. We can have several of them in the same module, each one corresponding to a different

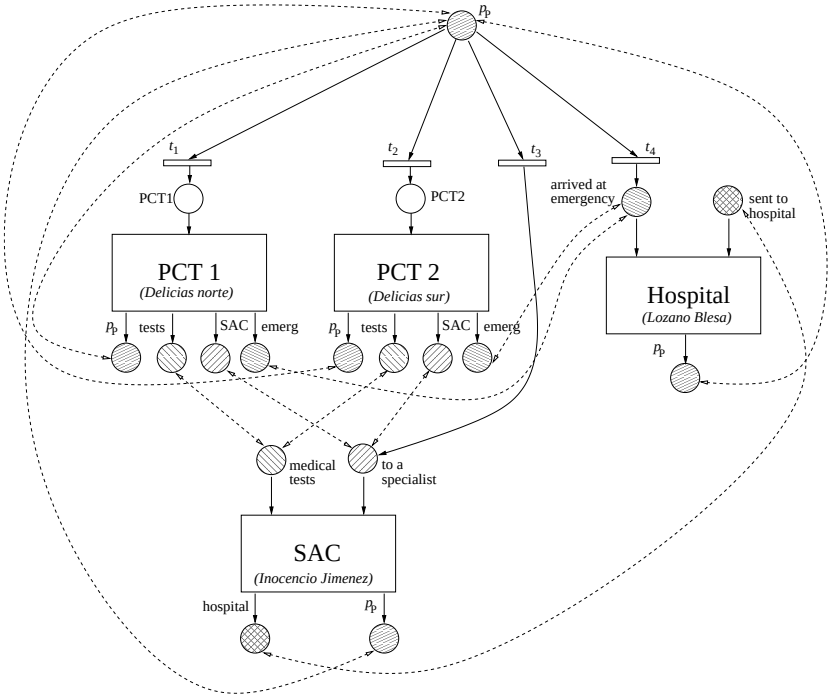

Figure 3: A schematic model of the Zaragoza III health area.

disease and possibly sharing resources of the same center. From the point of view of the patient, the protocol that the patients would follow should be applied in general, in several centers/module of the Healthcare Area. Therefore, the modules must be composed via fusion of places representing the output flow of patients of a module, incoming the next module where the protocol continues. In the final model, a complete protocol will be identified by a strongly connected state machine crossing one or more modules.

Once identified all protocols that should be implemented in a module and the input and output places of the module, the second step of the modeling methodology requires the addition of the resources. In general, these resources will be shared between the operations of different protocols but always are private to the module.

6 Modeling the Zaragoza Healthcare Area III

In this section we will apply the previous modeling methodology to the Spanish HS described in subsection 3.1 and in particular to Aragonese HS, which is structured in 8 healthcare areas. We will consider *Zaragoza III* healthcare area which includes: (1) one hospital (named *Hospital Clínico Universitario "Lozano Blesa"*); (2) one SAC (named: *Inocencio Jiménez*); and (3) 22 basic health zones each one with its own PCT.

6.1 Global model

For simplicity, only two basic healthcare zones are considered: (1) *Delicias Norte* with a population of 21,844 inhabitants for which the PCT has 15 primary doctors and 17 nurses; (2) *Delicias Sur* with a population of 24,508 inhabitants for which the PCT has 16 primary doctors and 20 nurses.

The Fig. 3 presents a scheme of the model to be constructed. In that picture we have represented modules with their input and output places used to connect these modules. A connection is represented by a dotted arc that, in the final model, will result in the fusion of the connected places. Additionally, we have explicitly represented the idle place, its output transitions and the connectivity with the final ones resulting entrance and terminal places according to Def. 4.3. Idle place p_P, modeling the population, initially contains a number of tokens equal to $n_p = 21,844 + 24,508 = 46,352$.

The first two modules are the PCT (PCT1 - Delicias Norte and PCT2 - Delicias Sur) each one with one input and four outputs. The marking of the input place represents the number of patients having appointments. An appointment can be with a doctor, with a nurse, with both of them or for doing medical tests, etc. Regarding the outputs, after an exploration, the primary doctor can send the patient to the Emergency Department, to SAC to be visited by a specialist, to make different medical tests or send him home with medical prescriptions and/or recommendations.

The second module is the SAC where a patient can come to be visited by a specialist or for do medical tests required by the primary doctor. These inputs are modeled by two places and their markings tell us the number of patients waiting for a specialist or waiting for medical tests. As outputs, the patients from SAC can go home with a prescription/recommendation or can be sent to the hospital in the case of a more complicated situation or condition.

Since the hospital has an Emergency Department, the Hospital module has two inputs, one for Emergency Department and another one for the hospital divisions. After the visit to the hospital, the patient goes home recovered or with a medical prescription/recommendation.

Notice that a patient goes to the PCT1 (occurrence of t_1), to the PCT2 (occurrence of t_2), to the SAC (occurrence of t_3) or to the Emergency Department (occurrence of t_4) according to its own evaluation or based on its citation. These transitions are the one introduced in the global model at step 5 of Def. 4.3. We assume uniform distribution of the diseases between both basic healthcare areas.

6.2 Modeling the modules

Once having the global model, the next step is to refine the modules. In the following, two medical protocols used in both PCT are described as examples. We assume that both PCT modules describe only two medical protocols: (1) for a patient with anginal chest pain (see Fig. 4); and (2) for a patient with low blood pressure. The PN models are sketched in Figs. 5 and 6 while the BHM model of the PCT is obtained by fusing resource places which are common in both protocols subnets and common outputs.

An anginal chest pain protocol used in primary healthcare. The first protocol considers patients with anginal chest pain that are going to the PCT to be visited by their

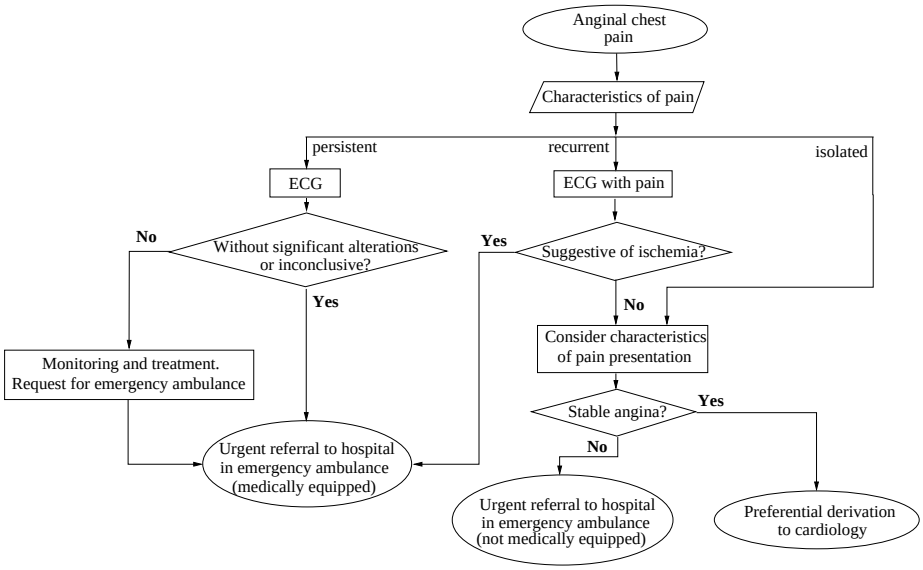

Figure 4: A medical protocol for a patient with anginal chest pain in primary care team adapted from [5].

primary doctors. The primary doctor should first do an anamnesis (case history) to the patient and see the characteristics of his pain. If the pain is persistent (could be of hours or days with fluctuation of intensity), the primary doctor asks for an ECG in the PCT with the help of a nurse. Based on the results, the primary doctor could ask urgently for a medically equipped emergency ambulance and should start controlling and care him until the emergency ambulance arrives. If the pain characteristic is recurrent (could be minutes but repeating in time) and the ECG with pain is suggestive of ischemic lesion the primary doctor asks for the emergency ambulance as well. If the ECG with pain is not suggestive of ischemia or the characteristic of the pain is isolated then the primary doctor should consider the pain presentation. If the angina is stable then the patient is sent to the specialized attention center to be seen by a specialist. Otherwise, he is sent to the hospital in an emergency ambulance but not necessary medically equipped.

The medical protocol in Fig. 4 can be modeled by the Petri net in Fig. 5. Place p_0 is modeling the waiting queue in PCT where the number of tokens in p_0 represents the number of patients with anginal chest pain waiting to be seen by a primary doctor. The initial diagnostic by the primary doctor is modeled by place p_1 and its markings is saying the number of patients that are seen in parallel by the primary doctors of the considered PCT. The time delay associated to output transition t_2 (exponentially distributed with an average of 5 minutes) means the time necessary to take the initial diagnostic by the medical doctors. Depending on this initial diagnostic, the doctor decides to follow the protocol for the *persistent pain* (transition t_3), *recurrent pain* (transition t_4) or *isolated pain* (transition t_5). If the pain is persistent or recurrent, the first step is to make an ECG (p_5) or an ECG with pain (place p_6). Notice that places p_3

and p_4 model the patients waiting for taking the ECG test. When this test is finished, the results are interpreted by the primary doctor (p_7 and p_8).

The protocol of persistent pain follows with a choice in place p_9. If the ECG is without significant alterations or inconclusive (transition t_{13}) the patient is waiting for the ambulance (place p_{12}) to be transfer urgently to the hospital in an medically equipped emergency ambulance. On the contrary, (transition t_{12}) the primary doctor starts to care him (place p_{11}) and asks for an emergency ambulance. In both situations, the patients are transferred to the hospital in the ambulance where they are cared and stabilized following the corresponding protocol of the hospital.

In the case of recurrent pain and based on the results of the ECG with pain interpreted by the primary doctor there is a choice (in place p_{10}). If the ECG with pain is suggestive of ischemia the primary doctor requests an urgent transfer to the hospital in a medically equipped emergency ambulance (transition t_{14}) and starts to care him (place p_{11}). If the ECG with pain is not suggestive of ischemia (transition t_{15}) or the characteristic of pain is isolated (transition t_5) the primary doctor should consider the characteristics of pain presentation (place p_{13}) and will chose two possible future evolution. If the angina is stable (transition t_{19}) the patient is sent to be visited by a specialist (output place p_{16}) in a SAC. If the angina is unstable (transition t_{18}), the patient is sent to the hospital in an emergency ambulance but not necessary medically equipped.

Patients with low blood pressure follow a protocol modeled by the PN in Fig. 6. They arrive to the PCT (place p'_0) and are first attended by a nurse that measures their blood pressure (p'_1) and then they wait to be visited by the doctor (p'_2). According to the value of the blood pressure and after the examination of the doctor (p'_3), the primary doctor could decide (p'_4) to ambulatory monitor the blood pressure (t'_5) or send the patient for an ECG (t'_6). If the ambulatory monitoring of the blood pressure is necessary, the patients wait in p'_5 the availability of a special equipment (resource r_5) and they are monitored in p'_6. If the ambulatory monitoring is negative (the patient has a normal blood pressure during the 24 hours of monitoring) the patient is sent home ($t'_9 \rightarrow p'_8$). Otherwise, in p'_9 waits for a doctor to be more investigated (p'_{10}) and in p'_{11} is waiting for an ECG. This test is performed in p'_{12} and after that (in p'_{13}), if the primary doctor detects arrhythmia or heart failure (with signs of gravity) the patient is transferred to the hospital (t'_{16}) in an emergency ambulance medically equipped. If from the ECG the primary doctor cannot detect alterations the patient is sent for additional medical tests such as blood exam (p'_{14} and p'_{18}). Notice that the nurse should take a blood extraction from the patient. The results of this analysis are evaluated by the primary doctor (p'_{20}) and he decide in p'_{21}: (a) if the potassium is high and the sodium low (t'_{24}) with the probability of adrenal gland disorder the patient is referred to the endocrinology in the specialized attending center to be evaluated by a specialist (p'_{22}). If the blood test shows normal results (t'_{23}) the primary doctor make to the patient the orthostatic test - that is a measure of the blood pressure in decubitus and then orthostatic - (p'_{22}) and if it is negative send the patient home with medication (t'_{26}). Otherwise (t'_{27}) (positive test - means that the blood pressure decreases in orthostatism between certain values) the doctor is evaluating the possibility of cardiac syncope or carotid syncope (p'_{25}). If the test is positive (t'_{29}) then the patient is referred to the cardiology (p'_{20}). Otherwise, (t'_{30}) neurogenic causes should be considered by the doctor (p'_{27}). In the

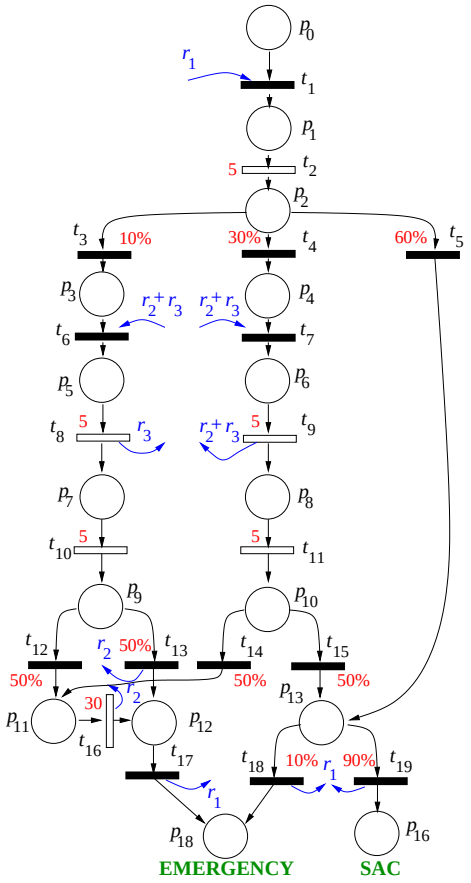

Figure 5: Petri net model of the medical protocol of Fig. 4. The resource places are not represented for sake of clarity. The allocations and releases of the resources are shown by blue arcs while the average time duration in minutes and the probabilities of choices are shown in red. The initial marking of the resources is given in Fig. 6.

negative case (t'_{33}), the patient is sent home with medication (treatment in primary care team) while in the positive case - neurological causes (t'_{32}) the doctor should evaluate if the hypotension is incapacitating (p'_{29}). If this result is negative (hypotension well tolerated) then the patient is sent home with medication (t'_{35}) while in the positive case is sent to the the neurologist (p'_{36}).

All places belonging to both medical protocols are (1) treatment places modeling treatments or cares that should be performed or (2) waiting places where the patients are waiting for the resource availability or a decision should be taken. In order to receive a treatment or care, one or more resources could be necessary. The PCT module is obtained by place fusing and the following BHM is obtained:

- Input places: $P^{in} = \{p_0, p'_0\}$;
- Output places: $P^{out} = \{p_{16} == p'_{22}, p_{18} == p'_{17}, p'_8\}$ where '==' means that the places are the same;
- Treatment places: $p^T = \{p_1 - p_{13}, p'_1, p'_3, p'_4, p'_6, p'_7, p'_{10}, -p'_{14}, p'_{18}, p'_{20}, p'_{21}, p'_{23} - p'_{30}\}$;
- Waiting places: $P^W = \{p'_2, p'_5, p'_9, p'_{19}\}$;
- Resource places: $P^R = \{r_1 - r_5\}$.

7 Exploiting the PN model

In this section some structural properties of the Petri Nets are interpreted in medical terms to illustrate some properties of the system that can be studied using the formal model.

Conservation of the resources. The healthcare systems we consider assume that the resource are conservative. For example, the number of doctors should be constant because we cannot *create* or *loose* doctors during the execution of a medical protocol. The same happens with the other resources, such as the nursers, medical equipments, etc. In the Petri net terminology, a conservation law corresponds to a P-semiflow.

Considering only the two medical protocols implemented in both PCT, the PN model is *conservative* having 11 P-semiflows, each one corresponding to the conservation of a resource: one P-semiflow corresponding to the population of the healthcare area, ten corresponding to the resources in each PCT (primary doctors, nurses, ECG machine, medical laboratories, pressure monitoring machine). For example, in the PN model of a PCT composed by the both protocols in Fig. 5 and Fig. 6 the P-semiflow corresponding to the conservation of the ECG machine (resource r_3) is: $\boldsymbol{m}(p_5) + \boldsymbol{m}(p_6) + \boldsymbol{m}(p'_{12}) + \boldsymbol{m}(p'_{13}) + \boldsymbol{m}(r_3) = 1$.

Repetitiveness of a medical protocol. A medical protocol corresponding to a given disease should be repetitive since all patients with the same symptoms should follow the same medical protocol. As stated in Sec. 4, a T-semiflow is associated to a repetitive behavior. Therefore, a medical protocol corresponds to one or more (in the case of choices) T-semiflows. For example, in the case of the protocol modeled in Fig. 6, any path from p'_0 to p'_8 together with the input transition in p'_0 (transition

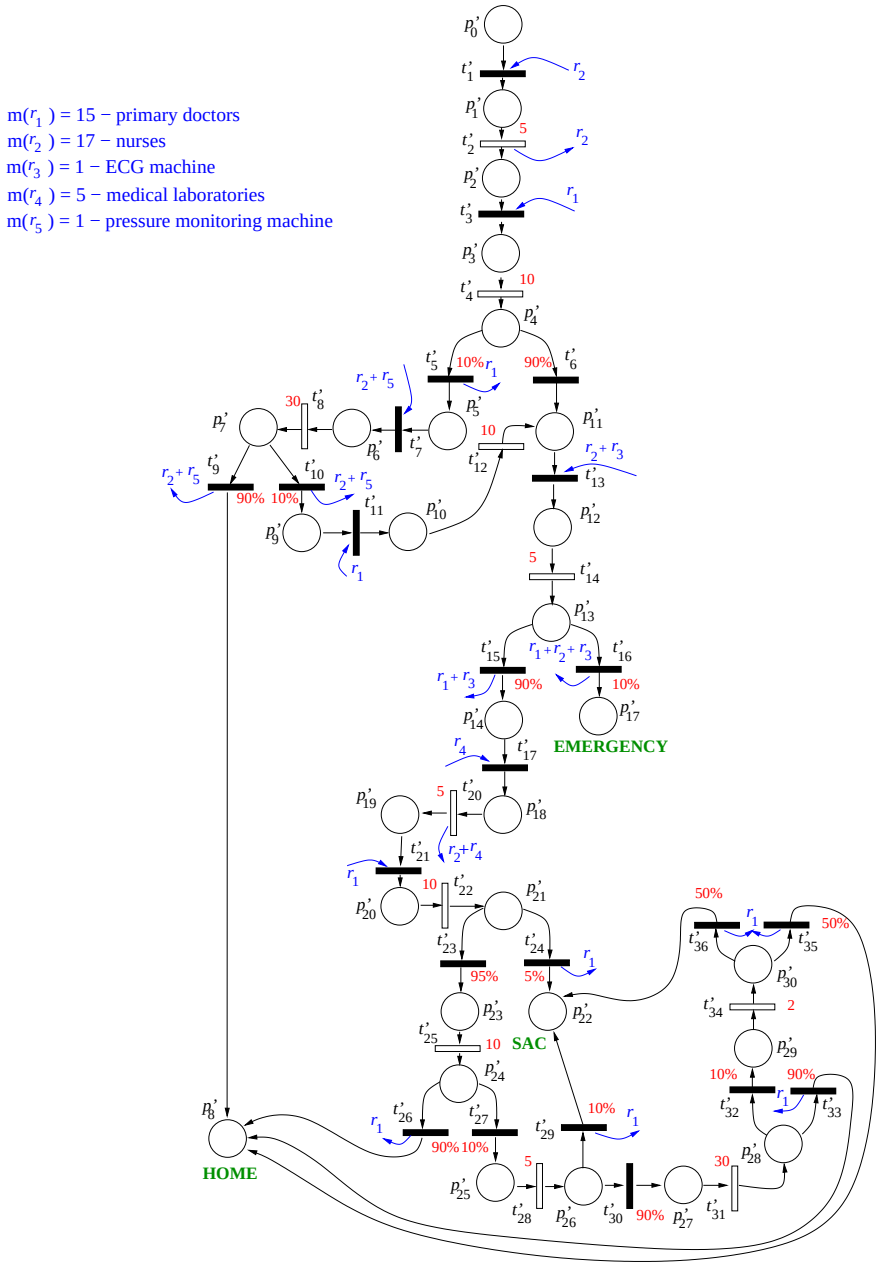

Figure 6: Petri net model of the medical protocol of patients with low blood pressure. The resource places are not represented for sake of clarity while the average time duration in minutes and the probabilities of the choices are shown in red. The initial marking of the resources is given in the left-upper corner.

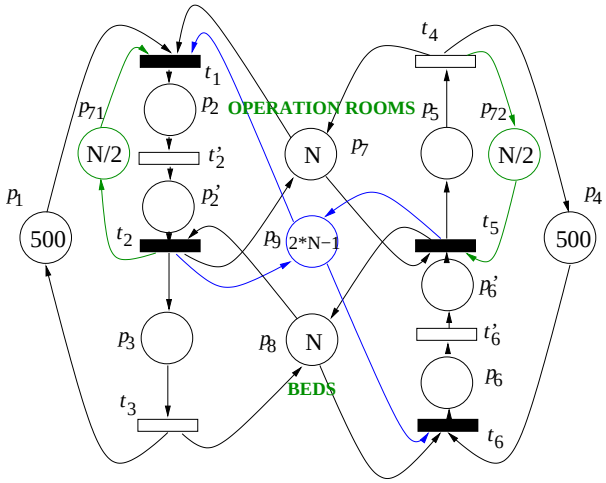

Figure 7: Two medical protocols with shared resources.

introduced into the model when the HPN is obtained by composition) is a T-semiflow since by firing of these transitions a patient will follow the protocol and is returned home.

Risk assessment for the violation of medical protocols. In the HS it is very unusual to obtain a state in which a medical protocol cannot continue due to the absence of a resource (a deadlock state in discrete-event systems). These situations are solved "in situ" by medical personnel detecting the situation, identifying the affected actors and resources, and recovering the situation using ad-hoc solutions. However, in some occasions these ad-hoc solutions may imply violations of the medical protocol and may have legal or ethical consequences.

Let us consider two protocols of a hospital department for surgeries given in Fig. 7. The first protocol corresponds to the emergency surgery consisting in the following sequence of activities: surgery (p_2), waiting for a bed (p_2') and then the hospitalization (p_3). The second protocol corresponds to the normal surgery: first the patient is hospitalized (p_6) then he waits for operation room (p_6') and then the surgery is performed (p_5). In order to apply these activities, the following resources are required: operation rooms (p_7) and hospital beds (p_8). For simplicity we assume the same number N of both resources. Places p_1 and p_4 are used to model the patients.

Let us assume $N = 2$, i.e., there exist two operation rooms and two hospital beds and consider that two patients arrive and emergency surgery is performed (transition t_1 fires twice putting two tokens in p_2) and then they wait for beds to be hospitalized (transition t_2' fires twice and the tokens are removed from p_2 and put in p_2'). Moreover, assume that in parallel two patients are hospitalized for normal surgery (transition t_6 fires twice putting two tokens in p_6) waiting for operation rooms (t_6' fires twice removing both tokens from p_6 and put two tokens in p_6'). Notice that the first two patients *occupy* the operation rooms and the last two *occupy* the hospital beds. Obviously, in

this situation, no medical protocol can be continued as scheduled and ad-hoc solutions should be considered. One of such solution could be of taking a bed from other hospital department. However, this solution creates different problems: (i) the number of beds in a hospital department is decreased and (ii) disturbance for patients since in the same room a new bed is added.

In PN terms, the system has a *deadlock* state, i.e., a state at which no transition can fire. It is well-known that this property is very difficult to be checked, in some cases being necessary to obtain and enumerate all reachable markings. Fortunately, having our model an equivalent behavior with S^4PN there exist necessary and sufficient conditions based on the PN structure. In our example, the deadlock state is obtained by starting both protocols twice in parallel: firing twice t_1 and t_6. Notice that at the obtained marking (two tokens in p_2' and p_6') both t_2 and t_5 (corresponding to the continuation of the protocols) cannot be fired without an external intervention because each protocol waits for a resource that is actually used by the other protocol.

Using the PN model we may use different methods to enforce deadlock-freeness. One of such solutions is to ensure that both medical protocols cannot start in parallel, i.e., places $p_2 + p_2'$ and $p_6 + p_6'$ cannot be marked with four tokens simultaneously. The controller for this solution is designed by adding a *monitor* place p_9 in Fig. 7 with an initial marking equal to $2 \cdot N - 1 = 3$ imposing maximum 3 tokens in places p_2, p_2', p_6 and p_6'.

Another solution of enforcing deadlock-freeness is to divide one resource into two parts and make each part private to the protocols. For example, the two available operation rooms can be split into two and make one room used only for emergency surgery and other one only for normal surgery (p_7 is replaced by p_{71} and p_{72} in Fig. 7). Having two control strategies to avoid the deadlock state, an immediate question is: which one is better? Which control algorithm ensures that the number of patients receiving a surgery is greater?

Performance evaluation, prediction and optimization. These quantitative properties that require time properties to be included in the model help the medical managers to take better decisions, since they can evaluate and analyze different choices. Time delays with exponential pdf are associated to transitions t_2', t_3, t_6' and t_4 of the net in Fig. 7. The time model will belong to the class of Generalized Stochastic PNs (GSPN) [1], for which, its reachability graph is isomorphic to a Markov chain. Hence, steady-state probabilities can be analytically computed.

Let us assume that all timed transitions (described before) of the PN in Fig. 7 have associated time delays given by random variables with exponential probability density functions with mean 1. By simulation we compute the throughput (the number of patients recovered per time unit) using different values of N and for both methods ensuring deadlock-freeness. The results are given in Table 1 illustrating that, for this particular situation, it is better to split (divide) the operation rooms into two and use them privately by protocols than using a controller to *cut* the deadlock state.

Let us now consider the medical protocols given in Figs. 5 and 6 implemented in PCT1. We assume that 400 patients are arriving daily at this medical center. We assume that in average, every 15 minutes is arriving a patient with anginal chest pain that should follow the medical protocol in Fig. 5, every 60 minutes is arriving a patient with low blood pressure that should follow the medical protocol in Fig. 6 while the

N	Monitor place		Private resources	
	Th. t_3	Th. t_4	Th. t_3	Th. t_4
2	0.38	0.38	0.27	0.46
4	0.61	0.61	0.66	0.97
6	0.8	0.8	1.06	1.47
8	0.99	0.99	1.5	1.97
10	1.15	1.15	1.93	2.476

Table 1: Simulations of the PN system in Fig. 7 with both policies ensuring the deadlock-freeness.

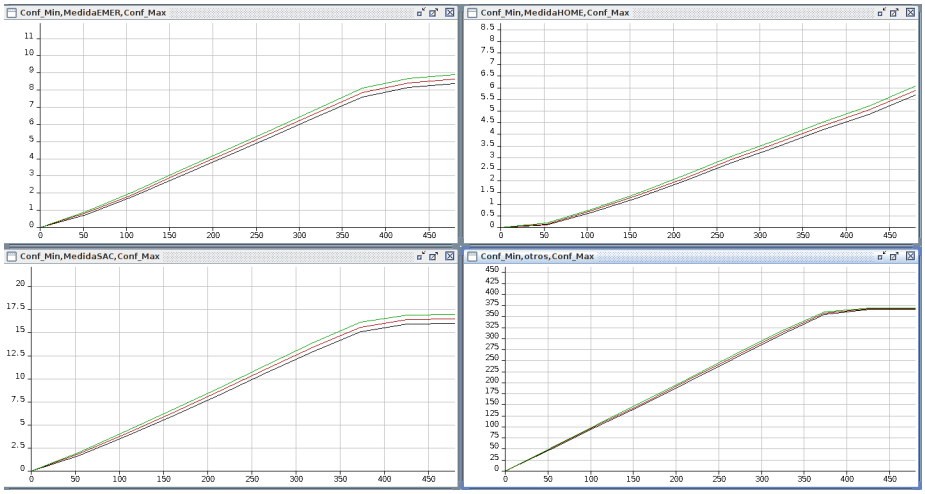

Figure 8: Simulation of two medical protocols inside a PCT without considering the use of resources during 8 hours.

other patients are arriving in average every minute.

We use the transitory simulation in TimeNET [32] to compute the percentage of the doctor' time (of 8 hours daily working time) used to treat patients with anginal chest pain and low blood pressure. In all simulations a confidence level of 98% has been used. First, we simulate the PCT module with these two medical protocols implemented without considering the resources. The results are given in Fig.8 in which there are three curves corresponding to the minimum, maximum and average values. From all 400 patients that arrive, 24.44 patients are consulted for anginal chest pain, 6.92 patients are presenting low blood pressure while $368.3$are coming with other pathologies. Figure in top-left of Fig.8 is showing the evolution of number of patients sent to the Emergency center of the hospital during the 480 minutes. At the end, a number of 8.79 patients have been sent to the emergency. In top-right figure in Fig.8, the number of patients sent home is represented, being at the end of day a number of 5.87 patients. In bottom-left of the figure is sketched the evolution of the patients sent to a specialist doctor at SAC being at the end of the day a number of 16.45 patients. The last

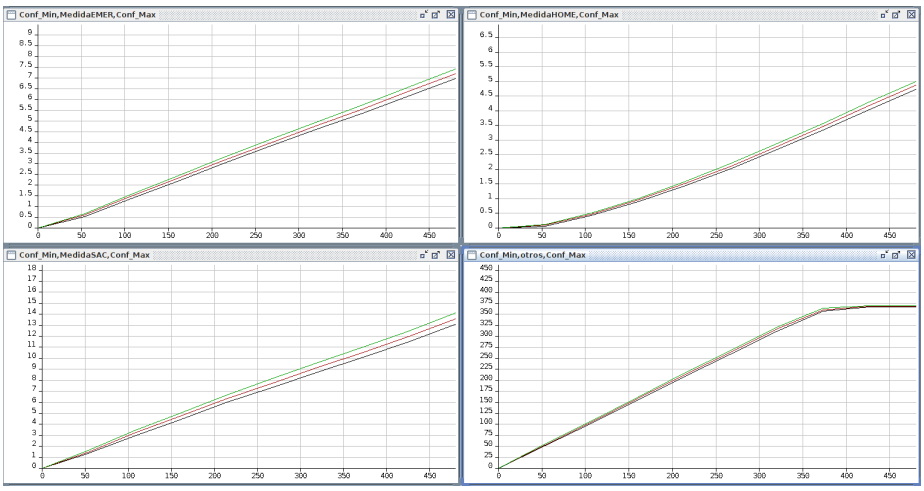

Figure 9: Simulation of two medical protocols inside a PCT considering the use of resources during 8 hours.

graphic represents the number of patients with other pathologies, being its total number equal to 368.3. At the end of the day, the number of patients that leave the PCT is: $8.79 + 5.87 + 16.45 + 368.3 = 399.34$ which is equal to the number of arriving patients.

Let us consider now the full model, i.e., including the resources. The same results are obtained by simulation assuming all available doctors (i.e., 15 doctors, see Fig. 6), meaning that with all available resources it is enough to treat all patients that are arriving in a day. By decreasing the number of medical doctors, the results remain almost the same even with 2 doctors. This does not mean that the number of doctors is overestimated since they should do other activities belonging to other medical protocols (that are not modeled). Simulating the model considering only one medical doctor, can be observed that not all patients are treated in 8 hours. The evolution of the number of patients sent to the emergency (top-left in Fig.9), the number of patients sent home (top-right in Fig.9), the number of patients sent to SAC (bottom-left in Fig.9) and the number of patients with other pathologies (bottom-right in Fig.9) shows that the system did not arrives to steady-state in 8 hours. In fact, the number of patients leaving the PCT in 8 hours is: $7.13 + 4.86 + 13.79 + 368.14 = 393.92$ less than the number of arriving patients that is 400. Since there exist 15 primary doctors and assuming that 1.5 primary doctors are necessary to treat the patients of these two medical protocols, $1.5/15 = 0,1$ (10%) of the time of each doctor is used to treat patients with anginal chest pain and low blood pressure.

Assuming one medical doctor to see both pathologies modeled in Figs. 5 and 6, Fig. 10 shows the waiting queue of the patients with low blood pressure to be initially seen by the primary doctor. This is obtained by looking at the average number of marking of place p_2' in the net of Fig.6.

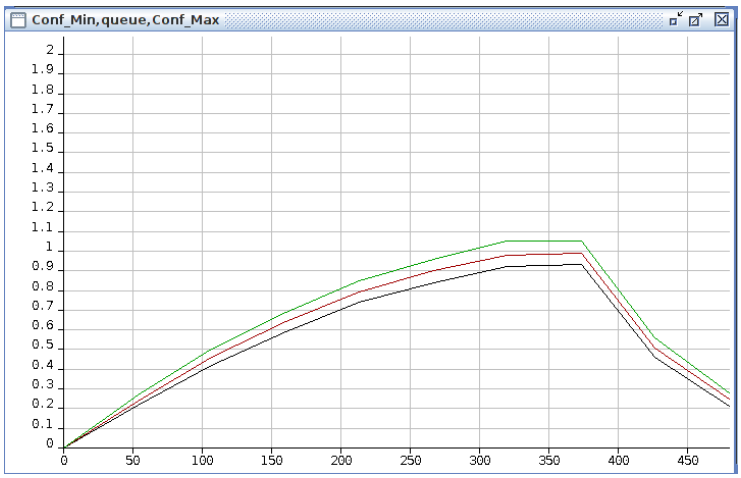

Figure 10: Waiting queue of the patients with low blood pressure to be seen by the primary doctor assuming that there is only one docotor in the PCT for both pathologies (average number of markings in place p'_2 of the net in Fig. 6).

8 Conclusions

This paper proposes a modular modeling approach for the healthcare systems. Three types of modules are identified, each one implementing medical protocols. The number of such protocols depends on the problem to be studied. Each protocol is modeled by a state machine PN and then, the shared resources are added. The global model is obtained by composition of the modules and it is proved to have a behavior equivalent to the well-known class of S^4PR.

The PN model that is obtained can be used for structural analysis to study important problems of the healthcare systems. The interpretation in medical terms of some structural elements is described. Moreover, the resource optimization is the core problem of these systems and the methodology presented here opens interesting ways of study it. Time durations can be associated to transitions and quantitative properties can be computed helping the hospital managers to take better decisions since they can evaluate different possibilities.

The main limitations of our approach are related to the type of resources that can be considered. We may model medical protocols used in healthcare systems with a limited number of available resources and also that are conservative. Moreover, we assume that they are private to one module. In the future work we plan to consider non private resources and we plan to develop a software tool to be used by the medical doctors.

References

[1] Ajmone Marsan, M., Conte, G., Balbo, G.: A class of generalized stochastic Petri nets for the performance evaluation of multiprocessor systems. ACM Trans. Comput. Syst. **2**(2), 93–122 (1984). DOI 10.1145/190.191. URL http://doi.acm.org/10.1145/190.191

[2] Augusto, V., Xie, X.: A modeling and simulation framework for health care systems. IEEE T. Systems, Man, and Cybernetics: Systems **44**(1), 30–46 (2014)

[3] Bahi-Jaber, N., Pontier, D.: Modeling transmission of directly transmitted infectious diseases using colored stochastic Petri nets. Mathematical Biosciences **185**(1), 1–13 (2003)

[4] Bernardi, S., Albareda, J., Colom, J., Mahulea, C.: A model-based approach for the specification and verification of clinical guidelines. In: M2H: Workshop on Models and Methods for Hospital Management and Planning held in conjunction with ETFA'2014 (2004)

[5] Boquet, J. (ed.): Guía de Ayuda al Diagnóstico en Atención Primaria. Sociedad Española de Medicina de Familia y Comunitaria (2005)

[6] Brailsford, S., Harper, P., Patel, B., Pitt, M.: An analysis of the academic literature on simulation and modelling in health care. Journal of Simulation (3), 130–140 (2009)

[7] Carey, R., Lloyd, R.: Measuring Quality Improvement in Health Care: A Guide to Statistical Process Control Applications. ASQ Quality Press, New York (1995)

[8] Clavel, D., Mahulea, C., Albareda, J., Silva, M.: Operation planning of elective patients in an orthopedic surgery department. In: M2H'2016: Second Workshop on Models and Methods for Hospital Management and Planning held in conjunction with ETFA'2016: 21st IEEE International Conference on Emerging Technologies and Factory Automation (2016)

[9] Colom, J.M.: The resource allocation problem in flexible manufacturing systems. In: W. Aalst, E. Best (eds.) Applications and Theory of Petri Nets, *Lecture Notes in Computer Science*, vol. 2679, pp. 23–35. Springer-Verlag, Berlin, Heidelberg (2003)

[10] Davies, R.: An assessment of models of a health system. Journal of the Operational Research Society **36**(8), 679–87 (1985)

[11] Dotoli, M., Fanti, M., Mangini, A., Ukovich, W.: A continuous Petri net model for the management and design of emergency cardiology departments. In: ADHS09: Proc. of the 3rd Analysis and Design of Hybrid Systems. Zaragoza, Spain (2009)

[12] Fanti, M.P., Iacobellis, G., Ukovich, W.: A Metamodelling Approach to Healthcare System management. In: A. Testi, E. Ivaldi, G. Carello, R. Aringhieri, V. Fraghelli (eds.) XXXVI ORHAS conference, Operation Research for Patient-Centered health care delivery, pp. 110–121 (2010)

[13] Fanti, M.P., Mangini, A., Dotoli, M., Ukovich, W.: A three level strategy for the design and performance evaluation of hospital departments. IEEE Trans. on Systems, man, and Cybernetics. Part A: Systems and Humans pp. 1 – 15 (2012)

[14] Fanti, M.P., Mangini, A.M., Ukovich, W., Lesage, J.J., Viard, K.: A petri net model of an integrated system for the health care at home management. In: 2014 IEEE International Conference on Automation Science and Engineering (CASE), pp. 582 – 587 (2014)

[15] Forrester, J.: The Impact of Feedback Control Concepts on the Management Sciences. Reprint. Foundation for Instrumentation Education and Research (1960)

[16] Gunal, M.: A guide for building hospital simulation models. Health Systems **1**(1), 17–25 (2012)

[17] Homer, J., Hirsch, G.: System Dynamics Modeling for Public Health: Background and Opportunities. American Journal of Public Health **96**(3), 452–458 (2006)

[18] Jensen, K., Kristensen, L.M.: Coloured Petri Nets - Modelling and Validation of Concurrent Systems. Springer (2009)

[19] Jun, J.B., Jacobson, S.H., Swisher, J.R.: Application of discrete-event simulation in health care clinics: A survey. Journal of the Operational Research Society pp. 109–123 (1999)

[20] Lamiri, M., Xie, X., Dolgui, A., Grimaud, F.: A stochastic model for operating room planning with elective and emergency demand for surgery. European Journal of Operational Research **185**(3), 1026–1037 (2008)

[21] Lamiri, M., Xie, X., Zhanga, S.: Column generation approach to operating theater planning with elective and emergency patientsy. IIE Transactions **40**(9), 838–852 (2008)

[22] Laskowski, M., Bryan, C., Demianyk, J., Shamir, N., Friesen, M.M., McLeod, R.: Agent-based modeling of the spread of influenza-like illness in an emergency department: A simulation study. IEEE Transactions on Information Technology in Biomedicine **15**(6), 877–889 (2011)

[23] Lehaney, B., Hlupic, V.: Simulation modelling for resource allocation and planning in the health sector. Journal of the Royal Society of Health **115**(6), 382 (1995)

[24] Macal, M., North, M.: Tutorial on agent-based modelling and simulation. Journal of Simulation **4**(3), 151–162 (2010)

[25] Mans, R., Schonenberg, M., Song, M., Aalst, W., Bakker, P.: Application of Process Mining in Healthcare - A Case Study in a Dutch Hospital. Biomedical Engineering Systems and Technologies **25**, 425–438 (2009)

[26] Murata, T.: Petri nets: Properties, analysis and applications. Proceedings of the IEEE **77**(4), 541–580 (1989)

[27] Sibbel, R., Urban, C.: Agent-Based Modeling and Simulation for Hospital Management, pp. 183–202. Springer Netherlands, Dordrecht (2001)

[28] Silva, M.: Introducing Petri nets. In: Practice of Petri Nets in Manufacturing, pp. 1–62. Chapman & Hall (1993)

[29] Sobolev, B.G., Sanchez, V., Vasilakis, C.: Systematic review of the use of computer simulation modeling of patient flow in surgical care. Journal of Public Health Medicine **35**(1), 1–16 (2011)

[30] Tricas, F., García-Valles, F., Colom, J., Ezpeleta, J.: A Petri net Structure-Based Deadlock Prevention Solution for Sequential Resource Allocation Systems. In: Int. Conf. on Robotics and Automation. Barcelona, Spain (2005)

[31] Whittaker, S.J., Rudie, K., McLellan, J.: An Augmented Petri Net Model for Health-Care Protocols. IEEE Transactions on Automatic Control **60**(9) (2015)

[32] Zimmermann, A.: Modeling and evaluation of stochastic Petri nets with TimeNET 4.1. In: VALUETOOLS2012: 6th International Conference on Performance Evaluation Methodologies and Tools, pp. 54–63. IEEE (2012)