Table of Contents

International Journal of Information Systems in the Service Sector

Volume 9 • Issue 2 • April-June-2017 • ISSN: 1935-5688 • eISSN: 1935-5696
An official publication of the Information Resources Management Association

Research Articles

The *International Journal of Information Systems in the Service Sector* is indexed or listed in the following: ACM Digital Library; Bacon's Media Directory; Cabell's Directories; CSA Illumina; DBLP; GetCited; Google Scholar; INSPEC; JournalTOCs; Library & Information Science Abstracts (LISA); MediaFinder; Norwegian Social Science Data Services (NSD); SCOPUS; The Index of Information Systems Journals; The Standard Periodical Directory; Ulrich's Periodicals Directory; Web of Science Emerging Sources Citation Index (ESCI)

International Journal of Information Systems in the Service Sector
Volume 9 • Issue 2 • April-June 2017

Decision Support based on Bio-PEPA Modeling and Decision Tree Induction:
A New Approach, Applied to a Tuberculosis Case Study

Dalila Hamami, Laboratoire d'informatique d'Oran (LIO), University of Oran 1 Ahmed Benbella, Oran, Algeria
Atmani Baghdad, Laboratoire d'informatique d'Oran (LIO), University of Oran 1 Ahmed Benbella, Oran, Algeria
Carron Shankland, Department of Computing Science and Mathematics, University of Stirling, Stirling, UK

ABSTRACT

The problem of selecting determinant features generating appropriate model structure is a challenge in epidemiological modelling. Disease spread is highly complex, and experts develop their understanding of its dynamic over years. There is an increasing variety and volume of epidemiological data which adds to the potential confusion. The authors propose here to make use of that data to better understand disease systems. Decision tree techniques have been extensively used to extract pertinent information and improve decision making. In this paper, the authors propose an innovative structured approach combining decision tree induction with Bio-PEPA computational modelling, and illustrate the approach through application to tuberculosis. By using decision tree induction, the enhanced Bio-PEPA model shows considerable improvement over the initial model with regard to the simulated results matching observed data. The key finding is that the developer expresses a realistic predictive model using relevant features, thus considering this approach as decision support, empowers the epidemiologist in his policy decision making.

KEYWORDS

Bio-PEPA Modelling, Data Mining, Decision Support, Decision Tree Induction, Epidemiology, Modelling and Simulation, Optimisation, Refinement, Tuberculosis

1. INTRODUCTION

The epidemiological field has been greatly enhanced by the use of computational and mathematical models, e.g. the studies of Anderson and May (1991), Weber et al, 1997; Keeling and Rohani (2008), Amouroux et al. (2010) and Hamami and Atmani (2013). Such models are considered indispensable both to understand the pathophysiology of human disease and to follow the spread of disease. The latter in particular allows public health policies to be developed by using predictive models to explore suitable disease control strategies.

For any modelling, the main goal is to provide accurate disease representation and realistic long term prediction; at least, as far as possible given that "the real world is undeniably replete with many complications; economic and social as well as biological" (Anderson and May, 1991). Capturing the complex, dynamic and variable nature of disease spread depends on strong partnership working between epidemiologists and modellers, to achieve careful refinement, elaboration and optimisation of models. Even so, the developed models (Anderson and May, 1991; Frost, 1995; Oaken et al., 2014)

DOI: 10.4018/IJISSS.2017040104

rely heavily on the experience of the experts and developers, and a degree of speculation and inspiration regarding identification of pertinent model features or accurate parameter estimation. Keeling and Rohani (2008) confirm this point of view: "The feasibility of model complexity is compromised by computational power, the mechanistic understanding of disease natural history, and the availability of necessary parameters. Consequently, the accuracy of any model is always limited". However, relying on expert knowledge and assumptions is not enough to ensure model accuracy when this depends on knowledge or features unknown to the expert/developer team.

In this context, many works (vynnycky and Fine, 1997; Debanne, 2000; Geisweiller, 2006; Prandi, 2010; DeEspíndola et al. 2011; Oaken REF, Goeyvaerts, 2015) focus on optimisation, as it becomes as a natural step in the modelling process. Optimisation has grown in recent years from considering simply parameter values, to refining model structure. Of great help in this process is the availability of massively complex datasets on epidemics, containing quantitative, qualitative, textual, Boolean, etc., information (Maumus et al., 2005). Our conclusion is that to decrease uncertainty in epidemic modelling, providing rigorous model descriptions containing the most important system features so parameters can then be correctly estimated, it is urgent to devise a solution to assisting experts/developers in acquiring only the most pertinent information from a dataset, and allow them to review their reasoning about the underlying epidemic system (Moundalexis and Nag, 2013).

To resolve this enigma and overcome the problems of selecting the determinant model features, in particular for tuberculosis (TB), we propose here, that a good epidemiological understanding and control requires a knowledge extraction process from data derived from cohort studies (Mancini, 2014; Poulymenopoulou et al., 2013). This process can involve symbolic methods of data mining (Maumus et al., 2005; Azar et al., 2013).

In epidemiology and public health, the use of data mining methods in general and decision tree induction in particular is growing briskly (Azar et al., 2013; Kotu and Deshpande, 2015; Breiman et al., 1984; Krizmaric et al., 2009; Smitha and Sundaram, 2012). Often these works mention the discovery of unexpected but effective information. As in other areas, it is the availability of wide-ranging historical databases that encourages such developments. By using data mining, patterns are discovered which can lead to better performance in computational modelling, long term prediction and decision-making (Lavanya and Rani, 2013). In our work, this process is automated by using WEKA tool (Hall et al., 2009), this offers a range of algorithms to build decision tree models.

The purpose of this article is:

- To show how the results from data mining can be complementary to the expert knowledge and help to achieve, update or validate an epidemic Bio-PEPA model,
- To present a framework in which data mining and Bio-PEPA modelling can be used together to better understand the mechanisms of detection and spread of epidemics, and
- To demonstrate the application of the framework to TB disease to identify influencing factors and their force.

This paper is structured as follows: section 2 provides background on Bio-PEPA modelling and data mining concepts more extensively on decision tree induction. Section 3 is dedicated to the proposed approach, which describes the different steps undertaken to combine Bio-PEPA with data mining. Details of the case study (tuberculosis), experimental approach and results of applying the Bio-PEPA framework using decision tree induction results are described in Section 4. Finally, in section 5, we conclude by summarizing and highlighting our key findings and contribution, together with perspectives on future work.

2. BACKGROUND

This section reviews the two main areas in our work, Bio-PEPA modelling and simulation and data mining.

2.1. Bio-PEPA Modelling and Simulation

Bio-PEPA (Bio-Performance Evaluation Process Algebra) is a formal language belonging to the Process algebra (PA) family. Developed in the 1970s, PA was mainly based on algebraic concepts (operators and axioms) to study the behaviour of parallel and distributed systems. It has since been used in biology: e.g. in 1993 Tofts (1993) used it to describe the behaviour of social insects, and in 2003 Norman and Shankland (2003) used it for epidemiology. Ciocchetta and Hillston (2009a, 2009b) developed a new, less-complex formalism, Bio-PEPA, to describe biological systems more succinctly. A general view of Bio-PEPA model components is given in Figure 1 (Appendix, all figures and tables are shown in Appendix).

Bio-PEPA is a formalism based on a set of rules and events (Ciocchetta and Hillston, 2009a) describing an interaction between a set of species (agents) belonging to one or a set of compartments and performing different reactions evolving under specific parameters. More formally and conveniently those concepts are described by the syntax below:

$S :: = (\alpha, \kappa)\ op\ S \mid S + S \mid C$
$op = << \mid >> \mid (+) \mid (-) \mid (.)$
$P :: = P <L> P \mid S(x)$

Where 'S': species or well known as individual entities. The dynamic of S is described by the reaction defined by 'α': action to undertake and 'κ': stoichiometry coefficient of the entity in that reaction. During the process 'P', S evolves under a specific operation 'op' as indicated above, where '<<': reactant, '>>': product, '(+)': activator, '(-)': inhibitor, '(.)': generic modifier. Bio-PEPA syntax offers the choice between different behaviours by using '+' (the full syntax details are presented by Ciocchetta and Hillston, 2009a, 2009b).

By applying Bio-PEPA to avian influenza Ciocchetta and Hilston (2009a) draw out the advantages of using Bio-PEPA for epidemiology modelling such as, its ability to deal with population level dynamics, the heterogeneity of individual attributes, stochasticity, spatial structure and discrete/external event. Further, Bio-PEPA offers a series of analyses not previously available to epidemiology through a single description such as stochastic simulation, model checking, ODE derivation and for those who are less familiar, Bio-PEPA allows translating an existing model to SBML (The Systems Biology Markup Language based on XML) (Hamami and Atmani, 2014).

The Ciocchetta and Hillston (2009a, 2009b) epidemiological studies led many authors to extend the use of Bio-PEPA to different infectious diseases. Benkirane et al. (2012) pinpointed the key features of Bio-PEPA by developing a measles model. They put forward seasonal effects and immigration on spreading disease. Hamami and Atmani (2012, 2013) have reviewed a Bonmarin mathematical model of chickenpox (Bonmarin et al., 2008) as well as De-Espindola tuberculosis model (DeEspíndola et al. 2011). Ramanathan et al. (2012) and Oaken et al. (2014) worked on SIR/SEIR models using the Bio-PEPA framework for deeper analysis. Despite the success of Bio-PEPA in epidemiological modelling, developers and experts still must avoid including irrelevant details and features and excluding pertinent ones in the model description.

2.2. Data Mining and Features Filtering

Data mining techniques are powerful tools to identify pertinent patterns and events within a large database. Data mining involves different techniques depending on the objective of the task and data to explore (Wang et al., 2012). They are summarised as predictive or descriptive methods (Kotu and

Deshpande, 2015). That is, predictive methods, such as classification and regression, use known outputs and the relationship between existing features to predict the future. Regression defines models using continuous output, as applied by Piarroux et al. (2011) to detect different levels of Cholera infection by region. Classification uses categorical output as done by Azar et al. (2013) to classify patients infected with Lymph disease. Descriptive methods, such as clustering and association rules, disclose concealed patterns that sum up the relationship between variables without predicting target values. Clustering regroups a set of objects with a similar specificity, as used by Almeida et al. (2014) in cardiovascular risk assessment where the resulting five clusters showed the intrinsic relation between features. Association rules identify a degree of association between features and their frequency, as achieved by Ou-yang et al. (2013) where the impact of prescribed drugs on Stevens–Johnson syndrome was detected. Thus, before applying data mining techniques, it is important to know which kind of method is more appropriate for our dataset study. Recall that the aim of this work is to use data mining techniques to enrich computational modelling by finding the relevant variables that explain the data. This means that the output of the data mining model is known. In addition, according to the categorical nature of our data, this description led us to focus on classification.

Classification is a data mining technique based on supervised learning (Kotu and Deshpande, 2015; López-Vallverdú et al., 2012): the learning is based on using known output values to build a model, useful to predict the class of objects whose class label is unknown. Various techniques such as: Decision Trees, Bayesian networks, Neural Networks, Rule induction, K-nearest neighbour, are used in classification. Many works highlight the decision tree as the classification method popularly used for classifying medical data (Lavanya and Rani, 2013; Mitchell, 1997; Phyu, 2009; Gorunescu, 2011; Carr et al., 2013). The decision tree is described by a tree structure where each non-leaf node denotes an attribute, each branch represents an attribute value and leaf nodes represent classes or class distributions. This structure makes models easy to interpret into rules: If Condition Then Conclusion, where Condition denotes a disjunction/conjunction of attributes, and Conclusion is the class reached by the condition (Atmani and Beldjilali, 2007). More advantages are reported in literature: Lavanya and Rani (2013) argued that decision tree algorithms are most commonly used because the parameter setting of domain knowledge is not required to construct the tree. Phyu (2009), by undertaking a survey of classification techniques, concluded that decision tree algorithms tend to perform better when dealing with discrete/categorical features. Gorunescu (2011) noted that the greatest benefit of decision tree approaches is flexibility, understandability and usefulness in prediction. Delen et al. (2005) used a series of decision tree algorithms (ID3, C4.5, C5 (Quinlan, 1993), and CART (Aguiar et al, 2012)) to identify variables and corresponding thresholds which separate observations in branches containing a set of leaves. Delen et al. (2005) outlined that the objective of decision tree algorithms is to minimise the number of homogeneous groups, and went on to apply the C5 algorithm to breast cancer data, extracting the most important features for the breast cancer prognosis. Azar et al. (2013) applied decision tree algorithm following classification to prove an increase in diagnostic confidence, by selecting six relevant features rather than the defined eighteen as data input. Krizmaric et al. (2009) focused on survival prediction of patient subject to cardiac arrest where features such as arrival time and cardiopulmonary resuscitation were detected as more pertinent for this study. Smitha and Sandaram (2012) applied a decision tree algorithm to predict the inhabitants infected by disease in a slum area. The resulting tree explains clearly that the infection is related first to climatic parameters and second to other parameters such as spread of deadly diseases, population immunity and control activities, vector abundance and family history.

These wide-ranging examples reveal that the decision tree is by far the most adequate classifier for our study, because:

- The resulting model is expressed by a tree (set of rules), easily interpreted by non-expert and well matched to Bio-PEPA model components.

- Decision tree algorithms perform better when dealing with the nature of features defined in our dataset (discrete, categorical),
- Decision tree algorithms identify variables relevant to the example, which is our principal aim in developing more realistic formal models. In the epidemiological setting, this draws out the causal relationships between predictors of the disease using a decision tree classifier, allowing relevant information to be extracted to understand and monitor epidemics.

To summarize, to improve and simplify the interaction between expert and developer, this section has identified two complementary techniques: decision tree induction and Bio-PEPA modelling. The former identifies the relevant features. The latter is used by the modeler to explore the usefulness of those features in optimising and refining a realistic and accurate model aiming to predict and improve the decision making of the epidemiologist. The next section explains how these techniques can be combined.

3. METHODOLOGY

Our aim is to use decision tree induction to extract useful information from the database to inform, refine and optimise our formal modelling.

To prove the usefulness of our approach, we begin by considering the typical manual modelling process and show how this can be enhanced with data mining. Figure 2 shows the structure of the methodology. Typical steps of our approach include:

1. **Interaction expert/developer:** Consecutive exchange knowledge between epidemiologist and developer is performed as follows:
   a. **Problem Definition:** identify critical areas in the process to be modelled.
   b. **Design the Study:** collect data (and possibly expert knowledge of the problem).
   c. **Design the Conceptual Model:** describe all dependencies between system components.
   d. **Process Definition:** determine the predictability and accuracy of the model, where inputs, outputs, assumptions and rules are specified separately.
2. **Bio-PEPA modelling:** Based on Bio-PEPA structure, the formal model is constructed using all the information gained in the previous step, inputs, outputs and rules.
3. **Simulation and analysis:** Once the formal model is constructed, it can be analysed. For this work we use stochastic simulation of the Bio-PEPA model. The resulting outputs are used to validate model accuracy by comparing with observed data.
4. **Optimisation:** The implementation of the model can be an accurate/inaccurate representation of the real system depending on the assumptions made by either the developer or the epidemiologist/domain expert. In either case, more information is required to refine/optimise the model.
   a. **Manual Optimisation (dashed line in Figure 2):** By returning to the process definition step defined earlier (Interaction expert/developer step), the expert enhances this step with new information. The optimisation based on expert/developer interaction is repeated until the results match well with observed data. This process is extremely reliant on expert/developer capability and knowledge, when basing only on inspiration and assumptions could derive to time consuming and increasing in complexity. To overcome these problems, we propose to use data mining at optimisation step.
   b. **Optimisation Using Data Mining:** Rather exploring the expert/developer interaction at optimization step, decision tree induction is used as factor retrieval on the disease dataset. This optimization begins with the data mining process (data cleaning, data transformation, feature selection, classification and validation) and ends in Bio-PEPA model refinement. The steps lie as follow:

- **Data Cleaning and Data transformation:** As disease dataset is collected from different sources, noise and errors can be expressed. Data cleaning attempts to correct inconsistencies, remove errors, noise and missing values in the data (Han and Kamber, 2006; Gibert et al, 2008), when data transformation (Inbarani et al, 2013) converts the data into appropriate forms for mining that makes data operationally efficient and understandable. To achieve this goal, a series of algorithms are available in literature (Inbarani et al, 2013; Witten, 2011), such as discretisation and removing missing values. The choice of those algorithms remains strongly dependent on dataset used.
- **Feature Selection Algorithms:** Not all features recorded in the dataset are useful in decision making. Feature selection is a preliminary step to classification, it reduces the attribute space with the aim of finding a minimal attribute set to describe the data (Guyon and Elisseeff, 2003). Those attributes are the classifier input deriving the optimal tree (optimal tree size and number of leaves) with highest accuracy. In data mining a range of feature selection algorithms are defined. According to Witten (2011), Saeys et al. (2007) and Karegowda et al. (2010), methods used for feature selection are classified into two types: Attribute subset evaluator and Single attribute evaluator. The choice of algorithm depends on the aim of feature selection. As argued by Saeys et al. (2007), attribute subset evaluators are used to improve prediction performance by considering feature dependencies whereas single attribute evaluators consider each feature separately to improve cluster detection (Inbarani et al., 2013). The aim of this study is to detect pertinent information expressed by the relation between different attributes to improve modelling prediction, therefore the Attribute Subset Evaluator algorithms are more suitable for this field. Many of them are defined in the literature. For example, Karegowda et al. (2010) applied a Correlation-based Feature Selection algorithm (CFS algorithm) combined with a neuronal network classifier to diabetic data to identify highest classifier accuracy through a highest correlation between features, while Macaš et al (2012) used Wrapper Subset Evaluator and Filtered Subset Evaluator combined with a series of classifiers. The choice of algorithm remains dependent on the nature of data to be mined.
- **Classification:** Once the feature selection step is achieved, the selected attributes can be used as an input to the classifier.As argued in section 2.2, decision tree algorithms are used in this study. A range of algorithms can be used to create the classifier, the most commonly reported in literature (Ou-yang et al., 2013; Shi, 2008; Zhao and Zhang, 2008; Gibert et al, 2010) are: Best First Decision Tree (BFTree), J48, J48Graft, Naive Bayesian Tree (NBTree), Alternating decision Tree using the LogitBoost strategy (LadTree), REPTree, RandomTree and Cart /Simple Cart.
- **Validation:** Once the models resulting from the classifiers listed above are achieved and trained, their performance is evaluated and significance is interpreted. To this end, a series of measures are undertaken such as: accuracy rate, confusion matrix, positive rate and negative rate. According to Witten (2011), confusion matrix is very useful measure for better understandability. The matrix is defined by predicted classes (matrix columns) and actual classes (matrix rows), where all correct predictions are expressed by its diagonal, see for example Table 5. Once the performance evaluated using the above measures, a comparison is done between all classifiers resulting in a ranked set.
- **Optimisation of Bio-PEPA Model:** The best ranked model resulting from the validation step is analysed to distinguish which parameters influence the classification results. To simplify this step, the selected model structured as a tree, where the first node is a root and terminal nodes reflect decision outcomes is converted into sets of rules described by a relation (arc) between a set of attributes (nodes) and then defined as: X and Y then Z, where X, Y are called antecedent (condition) and Z the consequent (conclusion) of the

rule. At this step, the selected attributes could be further validated by the epidemiologist as being primary reasons disrupting the analysis of disease spread which were unknown/missed by epidemiologist at the start of the study.

Having extended these rules, this information from data mining is incorporated into the Bio-PEPA model as follows:

1. Extract from mined rules, pertinent attributes not currently included in the Bio-PEPA model.
2. Refine the existing Bio-PEPA model by integrating relevant features.
3. Recalculate parameters useful to developing the Bio-PEPA model by restructuring the initial database.
4. Analyse the new simulated results.
5. Come back to step 1 here or to the data mining/dataset interaction phase to regenerate new rules if the aim is not achieved (i.e. the model is improved, but there is still a significant gap between observed data and model simulation).

To illustrate this methodology, the next section describes its application to a tuberculosis data set.

4. RESULTS

The Tuberculosis has been a major killer disease for several years which makes it a disease of interest for number of studies either in modelling and simulation field such as: Blower et al. (1998), Aparicio and Catillo-chavez (2009), DeEspíndola et al. (2011), Ozcaglar et al. (2012) and Hamami and Atmani (2013), or in data mining field such as: Sebban et al. (2002), Aguiar et al. (2012) and Venkatesan and Yamuna (2013).

According to the last report of the World Health Organization (WHO), the international standard for tuberculosis control, TB remains the leading infectious deadly disease today (WHO, 2012). WHO applies a strategy to reduce the transmission of the infection through prompt diagnosis and effective treatment of symptomatic TB patients who present at health care facilities, where strict supervision is based on recording individual patient data and their medicines taken during treatment period.

In 1985, the medical authority of Algeria, created the Service of Epidemiology and Preventive Medicine (SEMEP: Service d'Epidémiologie et MEdecine Préventive). The role of SEMEP is to co-ordinate and monitor health and prevention activities. SEMEP services work closely with the Department of Health and Population (DSP: Direction de la Santé et de la Population) for the collection of health information and its analysis. This is useful for statistical analysis of data, epidemiological interpretation, dissemination and exploitation of results. Although the SEMEP provides a great support to epidemiological monitoring, the large number and complexity of recorded data increase the difficulty to follow the spread of TB.

To demonstrate the value of our approach, we used data set obtained from the SEMEP of Mostaganem (Algeria). This data set consists of a set of locations situated in Mostaganem (Algeria). It records the details of individuals infected by tuberculosis from January, 2008 up to December, 2012: a total number of 998 cases. This data is an Excel spreadsheet with 23 attributes to describe each record described in Table 1, where nine attributes were ignored following data mining steps (more details are given in the section 4.4).

The process as described in section 3 is divided into three steps:

1. Realize TB Bio-PEPA model based on expert knowledge.
2. Analyse TB data using data mining techniques if the simulated output does not match observed data.

3. Rebuild existing model taking into consideration the extracted pertinent information from the second step.

4.1. Interaction Expert/Developer

Figure 3 formulates a global schema of the TB model that incorporates treatment and reinfection based on expert knowledge. The host population is divided into the following epidemiological classes or subgroups: susceptible moves through to infected by pulmonary TB (TP) when he is diagnosed. The TP moves to one of the different states (recovered, died, Trt_comp, lost, failed and transferred). It is noted in Figure 3 that:

- Because the TB treatment just allows recovery and does not give immunity, the recovered individual comes back to the susceptible class.
- Because of treatment failure, the individual in the Failure state comes back to the infected TP state.
- Because lost individuals are no longer part of the treated population, they will return into the infected class.

The main parameters that drive these transitions are shown in Table 2 with their values, and formula used to calculate the values from TB data.

4.2. Bio-PEPA Modelling

The aim here is to express the TB model, illustrated in Figure 3, in Bio-PEPA and to analyse the results.

As shown in Figure 4, the Bio-PEPA model is composed in a modular way through the interactions between the processes by defining:

Parameters/rates (P, θ1, …, θ6): numeric rates (Figure 4 from line 1 to 8), calculated using the observed data or collected from the literature (Aparicio and Catillo-chavez, 2009; Keeling and Rohani, 2008), see Table 2.

Location (space): Bio-PEPA defines a "Location" parameter which describes the place where the population is situated. For our initial model, we consider our population as homogeneous within a unique space (location) "City" (see Figure 4 line 9).

Species and Functional rates (KineticLawOf). The species correspond to the compartments defined in Figure 3 (Susceptible, Infected, transferred, Failed, Lost, Trt_comp and recovered). Each species carries out activities to change their own levels or those of others they may interact with (see Figure 4 from line 17 to 24). The rate of change is defined by the functional rate (see Figure 4 from line 10 to 16). For example, the action Recovery (line 12) leads to an increase in Recovered species (line 20) using the ">>" operator, while it leads to a decrease in Infected species (line 18) using the "<<" operator. By using the operator '+', the Infected species (line 18) has a choice between different actions at each time step.

The last line of the model (line 25) is the model component, defining the initial sizes and the interaction between species.

4.3. Simulation and Analysis

Once the model is achieved, a series of simulations are carried out in the Bio-PEPA plug-in (Duguid et al., 2009) (100 simulations are performed: Two Way ANOVA followed by Tukey Multiple Comparisons showed that the mean responses were not statistically different when more simulations were performed). The simulation is of one year, starting at t=0, where only the susceptible individuals and infected by pulmonary TB individuals are present, and ending at t=364.

Complete data series of five years are available, from 2008 to 2012. As some of Bio-PEPA parameters are calculated from observed data, Table 3 illustrates the period used according to the year of prediction. For example, to predict 2011, the average value of the set (2008-2009-2010) is considered to calculate parameters reported in Table 2.

To validate the model, the first simulation for observed data of 2008 is carried out and predicted data for 2009. A comparison is done between simulated (rounded mean) and observed data illustrated in histograms of Figure 5, for each class (died, failed, recovered, lost, transferred and Trt_comp).

As shown in Figure 5, the simulated model corresponds well to observed data. The histograms illustrate the state of individuals after 180 days of treatment. In order to strengthen the validity of these results, a χ^2 goodness of fit test was performed at 5% significance level. The null hypothesis (H0) is that the observed data follows the same distribution as the simulated data while the alternative hypothesis (H1) is that the observed data follows some other unspecified distribution. The results of this analysis was $\chi^2 = 0.381$, degree of freedom = 3, p-value = 0.944. Thus, there is insufficient evidence at the 5% level to reject H0 in favour of H1, which confirms that observed data is not different to the predicted. In the rest of the paper we summarize this argument by writing that the simulated data is not statistically different from the observed data (χ^2=x, degree of freedom=y, p-value=z). As there is no large variability between simulated and observed data, the optimization step is not required and the model is considered as an accurate one.

Moving on to 2010, further simulations (100 simulations) are carried out in the Bio-PEPA plug-in. Figure 7 shows histograms of the state of individuals, who were detected as infected in 2010, after 180 days of treatment. The same Bio-PEPA model was used to carry out this simulation as for 2009, keeping the same rates. It is clearly shown that a large difference separates simulated data and observed data, particularly the Lost class (resp. Trt_comp class) where the gap is estimated at 9 individual (resp. 9 individual). Nevertheless, the Lost state draws more attention than the Trt_comp state as it was under-predicted. It is noted that the Lost state is significant in both 2009 and 2010. As shown in Figure 5 and 7, the lost state is the largest group after recovered. Additionally, in 2010 our Bio-PEPA model predicts that the lost state characterizes 12% of infected when in the observed data it characterizes 23% of infected. This difference may mislead decisions by the epidemiologist. This state means that the patients are still infected and could cause potential infection in the population in the following years.

Thus, the Bio-PEPA model is inaccurate for 2010. At this step, as discussed in section 3, the principle of modelling and simulation is to apply the optimisation step.

4.4. Optimization

Deliberately, we apply manual optimisation first to better state its limitations and enhance the model with available information from either expert or literature. In this case, our expert observed the larger number of Lost in the observed data and proposed that the parameters may be adjusted. A series of experiments ranging over the flexible parameters: contact_ number and infection_period (see Table 2) show us that the results are insensitive to the balanced values of rates and converge to the same histograms in Figure 6. Therefore, what happened in 2010 that expert does not know? Which information is omitted from the simulated model? Which specific features could explain this large difference?

To refine the model and enrich the information given by the expert, data mining techniques are used. This process will not itself give a closer match to data, but it will explain what part of the population tends towards this lost state and helps us understand the underlying system. The model can then be revised accordingly.

WEKA (Waikato Environment for Knowledge Analysis, Hall et al., 2009) is a collection of machine learning algorithms for data mining tasks. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes (Ou-yang et al., 2013).

In this study WEKA is used to carry out experiments. Table 4 summarises the range of data set used for each data mining experiment. For example, if we are predicting the year 2010 in Bio-PEPA then the range of data set used in WEKA is from 2008 to 2010. According to the steps depicted in the section 3, the corresponding results are discussed below:

4.4.1. Data Cleaning and Data Transformation

By using either manual process or WEKA algorithms, some of irrelevant variables were removed and some others were transformed. In sum nine of them were removed and one transformed:

- "ID, First name, Last name, RecStatus" are not relevant to our analysis, and then removed manually.
- "DiagnoTEP and Preuve" do not help in diagnosis, as they concern extra-pulmonary tuberculosis, where in our study we focus on disease which spreads. Therefore, only pulmonary tuberculosis is considered.
- As the last action leads to remove all extra-pulmonary tuberculosis records, the "Loc" attribute as well "DiagnoTP" are referring to the only pulmonary tuberculosis, where their values do not vary at all, hence there were considered useless by WEKA filter and then removed.
- "MalAsso" for which 99% of records are not reported, and then considered as useless by WEKA filter.
- As "Age" is defined by a large number of possible values ranging from 3 to 94, discretisation was applied to reduce this number, where ages were transformed to nine groups using WEKA filter.

That leaves 14 attributes which could be significant. These were input to the feature selection and classification algorithms as described in Section 3.

4.4.2. Feature Selection and Classification

As argued in section 3.2, the Attribute Subset Evaluator is more suitable for this field. To analyse the performance of our approach, we compared five attribute Subset Evaluators combined with eight classifier algorithms, where the feature set resulting from each feature selection algorithm is assigned as an input to each classifier. In addition, the classification process is based on separated training and test data, as our data are limited due to the cleaning step undertaken above, a k-fold cross-validation (k =5, 10, 20, 25) algorithm is pre-applied (Saeys et al., 2007; Witten, 2011). This splits training and test data in different ways, to ensure we are not overfitting to training data.

In total, WEKA (Hall et al., 2009) runs 160 (5*8*4) experiments. Although this number of experiments seems staggering, WEKA automates the process and much more, WEKA repeats the process N times to give mean accuracy and standard deviation value. Data mining relies on additional parameters of the algorithms. We carried out a series of preliminary experiments: our recommendation is that the default values for WEKA parameters are used. These depend on our data (e.g. minimum number of objects: 2, confidence factor: 0.25, pruning: true). Results showed that the best accuracy was performed by running 10 times k-fold cross-validation, with k = 25. Indeed, splitting our dataset on 25 folds enables the fine grained heterogeneity of our data to be explored.

4.4.3. Validation

As identified in section 4.3, the lost state is the inaccurate part of our modelling; therefore, looking across our 160 experiments, we select classifiers reaching the highest class-wise accuracies particularly for the Lost state. In terms of feature selection algorithms, Filtered subset evaluator always provides the highest classification (for our data). In terms of classification algorithms, J48, J48Graft and LadTree algorithms reached the highest accuracy.

The rules resulting from J48 and J48Graft are described below, where the condition expresses the pertinent attributes and the conclusion expresses the state of individual.

As can be observed, the features Bacil 1-3 and Daira are those pertinent features inducing the state of individual during treatment. Recall that the Lost state is the inaccurate part of our modelling, the conditions leading to the Lost state are the point of interest (see dashed rectangle in the rules above). Although this indicates Bacil 1-3 as a classifier, in fact Bacil 1-3 are used as the definition of Lost: if we fail to have all of these tests, then the subject is defined as lost. Therefore, these add no additional information either to the expert knowledge or how the higher number of Lost cases arise.

We conclude the model arising from J48 and J48Graft does not give useful information, and we consider the LadTree model, which has the next highest accuracy after J48 and J48Graft. The rules resulting from Ladtree algorithm, as seen below, produce a multi class in the conclusion of the rule with their predictive values rather one class as in J48 tree.

If (Bacil3= MM) Then
(-1.16,4.217,-1.159,-1.158,0.415,-1.155)
If (Bacil3 ≠ MM) Then
(0.643,0.376,-0.197,-0.632,0.49,-0.68)
| If (Bacil2 = MM) Then
(-0.769,2.585,-0.696,-0.66,0.192,-0.652)
| If (Bacil2 ≠ MM) Then
(0.343,-0.628,0.297,-0.008,0.141,-0.144)
| | If (Daira = Ain Tedles) Then
(-0.59,0.64,-0.008,0.777,-0.237,-0.581)
| | If (Daira = Kheireddine) Then
(0.526,-0.647,0.199,-0.794,0.116,0.6)

The Ladtree is based on decision nodes and prediction nodes, where a decision node refers to conditions in the rules above, and a prediction node refer to conclusions. As the LadTree algorithm is well known as a multiclass decision tree, the conclusion is expressed by a vector of predictive values corresponding to each class. In our example the predictive values refer respectively to: Lost, Recovered, Failed, Died, Trt_comp, Transferred.

Recall that our aim is to filter the branch reaching to the Lost state. In LadTree, we follow all paths leading to the Lost state for which all decision nodes are true (the "true" refers to the positive values expressed between brackets in the above rules). By maximising the sums of all predictive values corresponding to each branch, the best classifier is then selected. In our example the strongest classifier leading to the Lost state was from maximising the values (0.643, 0.643+0.343, 0.643+0.343+0.526). This result leads us to conclude that the attribute "Daira" is the main factor arising to this classifier.

The aim of analysing these conditions is not to predict TB, but to detect, extract and understand what is common in general to all TB individuals described in the database and in particular those that are lost.

Table 5 shows the Ladtree algorithm results depicting the class-wise accuracy and confusion matrix for six classes, where columns denote the instances in a predicted class and rows denote the instances in an actual class. The Recovered class yields highest accuracy (0.978) followed by the Lost class (0.811). It is clear that LadTree algorithm successfully classified and identified patients who are lost after the end of treatment.

As the aim of this research is to find out the determining factors for being lost, Table 5 and described rules strengthen the usefulness of "Daira" attribute.

In fact, the rules described above mean that the lost individual, infected by pulmonary TB, for whom the smear test 2 and 3 are either positive or unavailable, has more chance to be located in Daira of Kheireddine than in Daira of AinTedles. This suggests that a more refined model structured

on Daira could be more consistent with observed data, by integrating the selecting rules to the initial Bio-PEPA model. The next section describes this step in detail.

4.4.4. Optimisation of Bio-PEPA Model

Two stages are required prior to further simulation: restructuring the TB database according to the condition described in the last section, and updating Bio-PEPA model.

- Restructuring tuberculosis database: To make the TB database heterogeneous, it should be divided into two parts, those situated in Kheireddine and those situated in Ain Tedles.
- Updating Bio-PEPA model: As the main concepts of Bio-PEPA are: parameters, compartments, functional rate and species, updating the initial model requires us to update each one of these concepts.

Conveniently, Bio-PEPA allows species to be grouped in compartments. In the first model the compartment was based on one location "City". Here, we split the "City" compartment into two sub-compartments corresponding to the Daira of interest which contains only two sub-locations: Kheireddine and Ain Tedles. The set of rates is essentially as before, but specialized to use only individuals and rules in the specified location from which the new values were calculated. These two distinguished compartments help us to follow each group separately in simulation. The full Bio-PEPA model is available online (Hamami, 2015).

The revised model can now be analysed using simulations (100 simulations, time period as before) and comparing to 2010 data to answer the questions: which part of the population makes the simulated model illustrated in Figure 6 different than the observed data? Further, which attribute is pertinent to conduct this analysis and detect the missing information?

Histograms in Figure 7 (resp. 8), illustrate comparison between simulated and observed data of individuals located at Ain Tedles (resp. Kheireddine), in 2010. As can be seen from Figure 7 (resp. Figure 8), the gap between simulated and observed data is more important for Lost individuals located in Kheireddine, than those located in Ain Tedles. By comparing them to the observed number of infected in each location, the gap for those located in Ain Tedles is 2% (with number of lost in simulated data 5 compared to 3 in observed data), and the gap for those located in Kheireddine is 30% (with number of lost in simulated data 5 compared to 15 in observed data). Figure 7 and 8 show clearly that the rest of classes matched well between observed and simulated data with insignificant differences. The observed data is statistically analysed at 5% significance level (with $\chi^2 = 0.862$, degree of freedom=2 and p-value = 0.650 for Ain Tedles and χ^2=5.742, degree of freedom = 2, p-value = 0.057 for Kheireddine).

Further, this simulation explains that group located in Kheireddine is the cause of the discrepancy between simulated and observed data which involves that more information is required to correctly predict an epidemic state. In general, in our approach, data mining can be repeated to extract further information from the restructured dataset. For the TB example the data is limited - just 40 instances for Kheireddine location. No new information was issued except Bacil 1-3; and these are not useful. By using symbolic decision tree induction, we have refined the initial model and more tightly identified the problem area which helps the expert to undertake the next step, to further investigate this particular portion of population and collect additional useful knowledge. Revealing this direct relationship between location and the lost state will lead the expert to investigate the district of Kheireddine more closely, and make a better decision.

By identifying the specific problem area, it is clearer now why the manual optimisation undertaken in section 4.4 did not lead to a more accurate result. The population in the global model was homogeneous and well mixed, with only one global rate of infection. By re-estimating this rate, it is impossible for the initial model to estimate accurately those lost in Kheirddine without leading to an imperfection for those lost located in Ain Tedles, and vice-versa. By defining the rate of infection

for each location, the revaluation using the range values defined in Table 2 is more accurate for that location. Indeed, the optimisation considers rates (rate of infection and rate of lost) related to the lost state in Kheireddine Location without changing those at Ain Tedles location. It is worth noting that the major concern in those rates is actually increasing the contact rate within population of Kheireddine. The choice was argued both by the formula defining the rate of infection illustrated in Table 2 and the capability to re-evaluate formula parameters. As infection probability is estimated from our data, this leads to re-evaluate the contact rate, which was increased from 27 to 29 for 2010. Results for updated model of the year 2010 are illustrated in Figure 9. Our histograms show better results when comparing simulated data to observed data, as a consequence of increasing the related rates which were under-estimated. Furthermore, to assess the global perspective of the last results corresponding to Kheireddine location (Figure 9), they were merged to those corresponding to Ain Tedles location (Figure 7) and compared to the global observed data (histograms in the right side of Figure 6). The final histograms in Figure 10 show clearly the positive impact of optimisation on our Bio-PEPA results. Recall that simply changing the parameter values (without changing the structure of the model) is not sufficient. Our study highlights the utility of decision tree induction in uncovering relevant features in the data, but also the requirement to couple this with constant reassessment of parameter values to achieve robust modelling results. The key element is that both of these are strongly tied to the nature of the disease, and the data collected.

In order to emphasize the generalized capability of our approach, the same process and simulations, as done for 2010, are carried out in the Bio-PEPA plug-in, for both years 2011 and 2012 by considering them as blind data, to show the refined model fits other years.

To predict 2011 (resp. 2012), the same initial Bio-PEPA model was used to carry out this simulation as for 2009 and 2010 keeping the same species and functions and varying rates depending on information extracted from 2008, 2009 and 2010 (resp. From 2008 to 2011). As shown in Figure 11 the simulated model corresponds well to observed data. The histograms illustrate the state of individuals, who were detected as infected in 2011, after 180 days of treatment. The statistical analysis (χ^2=1.550, degree of freedom = 3, p-value = 0.671) shows that the observed data is statistically similar to the simulated data at 5% significance level.

If (BACIL3 = NF)
| If (BACIL2 = NF)
| | If (BACIL1 = NF) Then Lost
| | If (BACIL1 = MM) Then Recovered
| | If (BACIL1 = MP)
| | | If (DAIRA = AIN TEDLES) Then Failed
| | | If (DAIRA = KHEIR EDDINE) Then Lost
| | If (BACIL1 = MP+) Then Lost
| If (BACIL2 = MM) Then Recovered
| If (BACIL2 = MP+)Then Transferred
If (BACIL3 = MM) Then Recovered
If (BACIL3 = MP) Then Failed

As for 2009, this simulation predicts well what happened in 2011, which leads us to strengthen our opinion that the Bio-PEPA model works well when epidemic knowledge is correctly stated. Contrariwise, for 2012 it is clearly shown in Figure 12 that only for the Lost state a large difference separates simulated data and observed data, as it is under-predicted. The same steps were undertaken, as it was done for 2010, to extract pertinent information from decision tree induction, thus the initial model is refined by integrating Daira attribute extracted from the resulting rules as shown above.

Results for the updated model are illustrated in Figure 13 (resp. 14). Histograms show comparison between simulated and observed data of individuals located at Ain Tedles (resp. Kheireddine), in 2012.

As can be seen from these Figures, the gap between simulated and observed data is more important for Lost individuals located in Kheireddine than those located in Ain Tedles. By comparing them to the observed number of infected in each location, the gap for those located in Ain Tedles is 6% (with 5 lost in simulated data rather than 10 in observed data), and for those located in Kheireddine is 14% (with 4 lost in simulated data rather than 17 in observed data) (see Figure 14).

The refined model, enriched by Daira attribute, identified more specifically the area of difference with the data of 2012. That is, we have used the information of 2010 to create a model which corresponds for other years not considered in our data mining step. Independently, we applied decision tree induction for 2012 to confirm the use of the Daira attribute.

Results achieved by using the same set of feature selection algorithms combined with classification methods, define J48 algorithm as the most accurate by using 20-fold cross-validation. According to Table 6 depicting the class-wise accuracy and confusion matrix for six classes, the Recovered class yields higher (0.996) followed by Lost class (0.904). It is clear that J48 algorithm successfully classified and identified patients who are lost after the end of treatment. It is clear that J48 algorithm outperformed for the 2012 data comparing to the 2010 data.

As the aim of this research is to find out the determining factors for being lost, the rules defined above reveal the pertinent attributes resulting from use of the Filtered Subset Evaluator and J48 classifier.

It is clear that Daira attribute remains the most pertinent information extracted from tuberculosis dataset.

Further, based on assumptions made by the expert for 2010 concerning strong influence that the Lost state and Kheireddine location have on contact rate, the latter was increased from 27 to 30. The corresponding results are illustrated in Figure 15.

The analysed histograms validate the usefulness of increasing the related rates which were under-estimated.

We also performed the merging process between the last histograms depicting simulated data for Kheireddine location (Figure 15) and those for Ain tedles Location (Figure 13), with the aim of comparing the merged histograms to the global observed data (histograms in the right side of Figure 12).

The final histograms in Figure 16 show better fitting between simulated and observed data compared to the first model results.

These results strengthen our assumptions that the expert missed important information that could enrich our Bio-PEPA prediction for both years 2010 and 2012. It is clear that something happened in Kheireddine location during 2010 and 2012, leading to perform a specific optimisation for a specific part of population rather than refining parameters of the whole population. Even if, we succeed to achieve an accurate model comparing to the observed data, the expert should investigate more research to understand really what happened at Kheireddine location which leads to this group of lost. At that time, our model can be subject for further future predictions.

5. CONCLUSION

In this paper we have presented results demonstrating the usefulness of combining data mining with Bio-PEPA modelling in the epidemiological field. We have done this by creating a framework in which data mining and Bio-PEPA modelling can be used together to better understand the mechanisms of detection and spread of epidemics, and by demonstrating its application to TB disease to identify influencing factors and their force. Thus we have met the objective set out at the beginning.

More specifically, we carried out a series of simulations to predict outbreaks in 2009, 2010, 2011 and 2012. The results showed that there is clearly variation between those different years. For 2009 and 2011, the initial prediction corresponded well to observed data, which means that all information used was sufficient to reproduce an accurate model. Conversely, for 2010 and 2012 the results showed

that the Bio-PEPA model ought to be enriched by new information (unknown by the expert). This is to be expected: variation within the system and unexpected future circumstances mean that the past is not always a good predictor of the future. However, using decision tree induction at this point helped to uncover which portion of the population should be subject to more investigation. This process was achieved by experimenting with eight decision tree classifiers combined with five feature selection algorithms, where the accuracy of classification reached to 76.41%. This rate is relatively low in data mining terms: this is due to our rather small, highly variable dataset. We therefore used accuracy enhanced by the true positive rate as a way of qualitatively identifying pertinent features to incorporate in our Bio-PEPA model. It is important to state that by analysing all dataset from 2008 to 2012 the accuracy was increased by 3%. In terms of feature selection algorithms and classifiers, the filtered subset evaluator yielded the highest accuracy for all classifiers where the best classifiers were Ladtree classifier for 2010 and J48 classifier for 2012. The results show that the most appropriate feature extracted is "Location". This pertinent attribute leads to divide the Bio-PEPA model into two parts: "Kheireddine" location and "Ain tedles" location. It is clear that the Kheireddine location is the principal part of the model where the developer should parameterize parameters differently to the rest of the model. In addition, it suggests to the expert subareas and subsets he should explore to make the right decision.

The last step in this experiment, based on expert hypothesis, was to prove the influence of the Location attribute on the infection rate by inferring the number of contacts through experiments.

By comparing our analysis to other modelling and simulation works, as done by Aparicio and Castello-chavez (2009), when the simulated model does not fit with observed data, it is better to use pertinent parameters extracted from data mining than to select by inspiration. Aparicio and Castello-chavez (2009) argued in their last work, the importance of modelling age and its influence on the number of contacts. The parameters used for these attributes are drawn from a literature review. The question is: are those parameter values ranges the right ones? And are there other features more important than this one? For example, if the ranges of age groups resulting from expert analysis are not clustered correctly then significant and pertinent information will be hidden from the expert. As argued by Anderson and May (1991): "even if using a roughly flat age distribution in the host population had large impact on the force of infection for a specific period, this could be an unreasonable assumption for another period". Further, in our study the age was among features defining TB data, but at no point was it depicted as the pertinent one by data mining. Through the use of decision tree induction, medical experts can detect relevant paths and even anomalies better than just human observation of datasets. By using Bio-PEPA modelling and simulation tools, we were able not only to validate the usefulness of extracting rules for the epidemiological study, but also to design the patterns which help to identify which, among a series of parameters, is the cause of an epidemic. By doing this, Bio-PEPA with symbolic induction decision tree aids the decision making of the epidemiologist.

In this study we proved the performance of using data mining at optimisation step in the existing computational model gaining on time and complexity.

This work is the first step in showing that data mining techniques generally can be used to support formal modelling. In future work we plan to optimise a selection of parameters affecting classifier performance and to carry out a large comparative study of all the data mining techniques, including association rules and clustering algorithms, as well as combining a set of classifiers, and their combination with our modelling approach. This will expand the range of measures used to select new content for our formal models. That is, rather than simply using accuracy rate (as here) as the principal measure to choose the best classifier, we can use information about clusters and associations to enhance the model.

REFERENCES

Aguiar, F. S., Almeida, L. L., Ruffino-Netto, A., Kritski, A. L., Mello, F. C., & Werneck, G. L. (2012). Classification and regression tree (CART) model to predict pulmonary tuberculosis in hospitalized patients. *BMC Pulmonary Medicine*, *12*(1), 40. doi:10.1186/1471-2466-12-40 PMID:22871182

Almeida, V. G., Borba, J., Pereira, H. C., Pereira, T., Correia, C., Pêgo, M., & Cardoso, J. (2014). Cardiovascular risk analysis by means of pulse morphology and clustering methodologies. *Computer Methods and Programs in Biomedicine*, *117*(2), 257–266. doi:10.1016/j.cmpb.2014.06.010 PMID:25023535

Amouroux, e., taillandier, p., & drogoul, a. (2012). Complex environment representation out epidemiology abm: application on h5n1 propagatio. *Tạp chí Khoa học và Công nghệ, 48*(4).

Anderson, R. M., May, R. M., & Anderson, B. (1992). *Infectious diseases of humans: dynamics and control* (Vol. 28). Oxford: Oxford university press.

Aparicio, J. P., & Castillo-Chavez, C. (2009). Mathematical modelling of tuberculosis epidemics. *Mathematical Biosciences and Engineering*, *6*(2), 209–237. doi:10.3934/mbe.2009.6.209 PMID:19364150

Atmani, B., & Beldjilali, B. (2012). Knowledge discovery in database: Induction graph and cellular automaton. *Computing and Informatics*, *26*(2), 171–197.

Azar, A. T., Elshazly, H. I., Hassanien, A. E., & Elkorany, A. M. (2014). A random forest classifier for lymph diseases. *Computer Methods and Programs in Biomedicine*, *113*(2), 465–473. doi:10.1016/j.cmpb.2013.11.004 PMID:24290902

Benkirane, S., Norman, R., Scott, E., & Shankland, C. (2012). Measles epidemics and PEPA: an exploration of historic disease dynamics using process algebra. In FM 2012: Formal Methods (pp. 101-115). Springer Berlin Heidelberg. doi:10.1007/978-3-642-32759-9_11

Blower, S. M., & Gerberding, J. L. (1998). Understanding, predicting and controlling the emergence of drug-resistant tuberculosis: A theoretical framework. *Journal of Molecular Medicine*, *76*(9), 624–636. doi:10.1007/s001090050260 PMID:9725765

Bonmarin, I., Santa-Olalla, P., & Lévy-Bruhl, D. (2008). Modélisation de limpact de la vaccination sur lépidémiologie de la varicelle et du zona. *Revue dEpidemiologie et de Sante Publique*, *56*(5), 323–331. doi:10.1016/j.respe.2008.07.087

Brieman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterrey, CA: Wadsworth & Brooks.

Carr, M., Ravi, V., Reddy, G. S., & Veranna, D. (2013). Machine Learning Techniques Applied to Profile Mobile Banking Users in India. *International Journal of Information Systems in the Service Sector*, *5*(1), 82–92. doi:10.4018/jisss.2013010105

Ciocchetta, F., & Hillston, J. (2009). Bio-PEPA: A framework for the modelling and analysis of biological systems. *Theoretical Computer Science*, *410*(33), 3065–3084. doi:10.1016/j.tcs.2009.02.037

Ciocchetta, F., & Hillston, J. (2009a). Bio-PEPA for epidemiological models. *ENTCS*, *261*, 43–69.

de Espíndola, A. L., Bauch, C. T., Cabella, B. C. T., & Martinez, A. S. (2011). An agent-based computational model of the spread of tuberculosis. *Journal of Statistical Mechanics*, (05): P05003.

Debanne, S. M., Bielefeld, R. A., Cauthen, G. M., Daniel, T. M., & Rowland, D. Y. (2000). Multivariate Markovian modeling of tuberculosis: Forecast for the United States. *Emerging Infectious Diseases*, *6*(2), 148–157. doi:10.3201/eid0602.000207 PMID:10756148

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. doi:10.1016/j.artmed.2004.07.002 PMID:15894176

Duguid, A., Gilmore, S., Guerriero, M. L., Hillston, J., & Loewe, L. (2009, December). Design and development of software tools for Bio-PEPA. *Proceedings of theWinter Simulation Conference* (pp. 956-967). Winter Simulation Conference. doi:10.1109/WSC.2009.5429725

Frost, W. H. (1995). The age selection of mortality from tuberculosis in successive decades. *American Journal of Epidemiology*, *141*(1), 4–9. PMID:7801964

Geisweiller, N. (2006). EM-PEPA, A Software to Find the Most Likely Rates Inside a PEPA Model. Retrieved from http://empepa.sourceforge.net/

Gibert, K., Sanchez-Marre, M., & Codina, V. (2010). *Choosing the right data mining technique: classification of methods and intelligent recommendation* (Doctoral dissertation). International Environmental Modelling and Software Society.

Gibert, K., Spate, J., Sànchez-Marrè, M., Athanasiadis, I. N., & Comas, J. (2008). Chapter twelve data mining for environmental systems. *Developments in Integrated Environmental Assessment*, *3*, 205–228. doi:10.1016/S1574-101X(08)00612-1

Goeyvaerts, N., Willem, L., Van Kerckhove, K., Vandendijck, Y., Hanquet, G., Beutels, P., & Hens, N. (2015). Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-specific influenza-like illness incidence. *Epidemics*, *13*, 1–9. doi:10.1016/j.epidem.2015.04.002 PMID:26616037

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques* (Vol. 12). Springer Science & Business Media. doi:10.1007/978-3-642-19721-5

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11*(1), 10-18.

Hall, M., Witten, I., & Frank, E. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington: Kaufmann.

Hamami, D. (2015). URL Bio-PEPA code. Retrieved from http://www.cs.stir.ac.uk/~dha/

Hamami, D., & Atmani, B. (2012). Modeling the effect of vaccination on varicella using Bio-PEPA.*Proc. of IASTED* (pp. 783-077). doi:10.2316/P.2012.783-077

Hamami, D., & Atmani, B. (2013, April). Tuberculosis Modelling Using Bio-PEPA Approach. In Proceedings of World Academy of Science, Engineering and Technology (No. 76, p. 871). World Academy of Science, Engineering and Technology (WASET).

Hamami, D., & Atmani, B. (2014). From Simulated Model By Bio-PEPA to Narrative Language Through SBML. *International Journal of Control Theory and Computer Modeling*, *4*(1/2), 27–43. doi:10.5121/ijctcm.2014.4203

Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann Publishers.

Holmes, G., Pfahringer, B., Kirkby, R., Frank, E., & Hall, M. (2002). Multiclass alternating decision trees. Proceedings of the Machine learning ECML '02 (pp. 161-172). Springer Berlin Heidelberg. doi:10.1007/3-540-36755-1_14

Inbarani, H. H., Azar, A. T., & Jothi, G. (2014). Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis. *Computer Methods and Programs in Biomedicine*, *113*(1), 175–185. doi:10.1016/j.cmpb.2013.10.007 PMID:24210167

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271–277.

Keeling, M. J., & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kotu, V., & Deshpande, B. (2015). Data Mining Process Predictive Analytics and Data Mining. Morgan Kaufmann.

Krizmaric, M., Verlic, M., Stiglic, G., Grmec, S., & Kokol, P. (2009). Intelligent analysis in predicting outcome of out-of-hospital cardiac arrest. *Computer Methods and Programs in Biomedicine*, *95*(2), S22–S32. doi:10.1016/j.cmpb.2009.02.013 PMID:19342117

Lanzas, C., & Chen, S. (2015). Complex system modelling for veterinary epidemiology. *Preventive Veterinary Medicine*, *118*(2), 207–214. doi:10.1016/j.prevetmed.2014.09.012 PMID:25449734

Lavanya, D., & Rani, K. U. (2013). A Hybrid Approach to Improve Classification with Cascading of Data Mining Tasks. *International Journal of Application or Innovation in Engineering Management*, *2*, 345–350.

López-Vallverdú, J. A., Riañ, O. D., & Bohada, J. A. (2012). Improving medical decision trees by combining relevant health-care criteria. *Expert Systems with Applications*, *39*(14), 11782–11791. doi:10.1016/j.eswa.2012.04.073

Macaš, M., Lhotská, L., Bakstein, E., Novák, D., Wild, J., Sieger, T., & Jech, R. et al. (2012). Wrapper feature selection for small sample size data driven by complete error estimates. *Computer Methods and Programs in Biomedicine*, *108*(1), 138–150. doi:10.1016/j.cmpb.2012.02.006 PMID:22472029

Mancini, M. (2014). Exploiting big data for improving healthcare services. *Journal of e-Learning and Knowledge Society, 10*(2).

Mantas, J. (2014). Machine learning for knowledge extraction from phr big data. *Integrating Information Technology and Management for Quality of Care*, *202*, 36. PMID:25000009

Marco, D., Shankland, C., & Cairns, D. (2012, July). Evolving Bio-PEPA process algebra models using genetic programming.*Proceedings of the 14th annual conference on Genetic and evolutionary computation* (pp. 177-184). ACM. doi:10.1145/2330163.2330189

Maumus, S., Napoli, A., Szathmary, L., & Visvikis-Siest, S. (2005). Fouille de données biomédicales complexes: extraction de règles et de profils génétiques dans le cadre de l'étude du syndrome métabolique. In Journées Ouvertes Biologie Informatique Mathématiques JOBIM '05 (pp. 169-173).

Mitchell, T. (1997). Decision tree learning. In Machine Learning (Vol. 414, Ch. 3, pp. 52–78). The McGraw-Hill Companies, Inc.

Moualeu-Ngangue, D. P., Röblitz, S., Ehrig, R., & Deuflhard, P. (2015). Parameter Identification in a Tuberculosis Model for Cameroon. *PLoS ONE*, *10*(4), e0120607. doi:10.1371/journal.pone.0120607 PMID:25874885

Moundalexis, M. L., & Nag, B. N. (2013). Decision making, dashboard displays, and human performance in service systems. *International Journal of Information Systems in the Service Sector*, *5*(4), 32–46. doi:10.4018/ijisss.2013100103

Norman, R., & Shankland, C. (2003). Developing the use of process algebra in the derivation and analysis of mathematical models of infectious disease. In *Computer Aided Systems Theory-EUROCAST 2003* (pp. 404–414). Springer Berlin Heidelberg. doi:10.1007/978-3-540-45210-2_37

Oaken, D. R. (2014). Optimisation of Definition Structures & Parameter Values in Process Algebra Models Using Evolutionary Computation.

Ou-Yang, C., Agustianty, S., & Wang, H. C. (2013). Developing a data mining approach to investigate association between physician prescription and patient outcome–A study on re-hospitalization in Stevens–Johnson Syndrome. *Computer Methods and Programs in Biomedicine*, *112*(1), 84–91. doi:10.1016/j.cmpb.2013.07.004 PMID:23910224

Ozcaglar, C., Shabbeer, A., Vandenberg, S. L., Yener, B., & Bennett, K. P. (2012). Epidemiological models of Mycobacterium tuberculosis complex infections. *Mathematical Biosciences*, *236*(2), 77–96. doi:10.1016/j.mbs.2012.02.003 PMID:22387570

Phyu, T. N. (2009, March). Survey of classification techniques in data mining. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *1*, 18–20.

Piarroux, R., Barrais, R., Faucher, B., Haus, R., Piarroux, M., Gaudart, J., & Raoult, D. et al. (2011). Understanding the cholera epidemic, Haiti. *Emerging Infectious Diseases*, *17*(7), 1161–1168. doi:10.3201/eid1707.110059 PMID:21762567

Prandi, D. (2010). Particle swarm optimization for stochastic process calculi.*Proceedings of the 9th Workshop on Process Algebra and Stochastically Timed Activities* (pp. 77-82).

Quinlan, J. (1993). *C4.5: programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics, 23*(19), 2507-2517.

Sebban, M., Mokrousov, I., Rastogi, N., & Sola, C. (2002). A data-mining approach to spacer oligonucleotide typing of Mycobacterium tuberculosis. *Bioinformatics (Oxford, England)*, *18*(2), 235–243. doi:10.1093/bioinformatics/18.2.235 PMID:11847071

Shi, H. (2008). Best-first decision tree learning (Thesis). Citeseer, Hamilton.

Smitha, T., & Sundaram, V. (2012). Classification rules by decision tree for disease prediction. *International Journal of Computers and Applications*, *43*, 35–37.

Tofts, C. (1993). Using process algebra to describe social insect behaviour. *Transactions of the Society for Computer Simulation*, *9*(4), 227–283.

Venkatesan, P., & Yamuna, N. R. (2013). Treatment response classification in randomized clinical trials: A decision tree approach. *Indian Journal of Science and Technology*, *6*(1), 3912–3917.

Vynnycky, E., & Fine, P. E. M. (1997). The natural history of tuberculosis: The implications of age-dependent risks of disease and the role of reinfection. *Epidemiology and Infection*, *119*(02), 183–201. doi:10.1017/S0950268897007917 PMID:9363017

Wang, Z., Yan, R., Chen, Q., & Xing, R. (2012). Data mining in nonprofit organizations, government agencies, and other institutions. *Advancing the Service Sector with Evolving Technologies: Techniques and Principles: Techniques and Principles*, 208.

Weber, A., Weber, M., & Milligan, P. (2001). Modeling epidemics caused by respiratory syncytial virus (RSV). *Mathematical Biosciences*, *172*(2), 95–113. doi:10.1016/S0025-5564(01)00066-9 PMID:11520501

Wolkewitz, M., & Schumacher, M. (2011). Simulating and analysing infectious disease data in a heterogeneous population with migration. *Computer Methods and Programs in Biomedicine*, *104*(2), 29–36. doi:10.1016/j.cmpb.2010.05.007 PMID:20633950

World Health Organization (WHO). (2009). Tuberculosis. Retrieved from http://www.who.int/topics/tuberculosis/en/

Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research*, *41*(12), 1955–1959. doi:10.1016/j.asr.2007.07.020

APPENDIX

Figure 1. Bio-PEPA model component

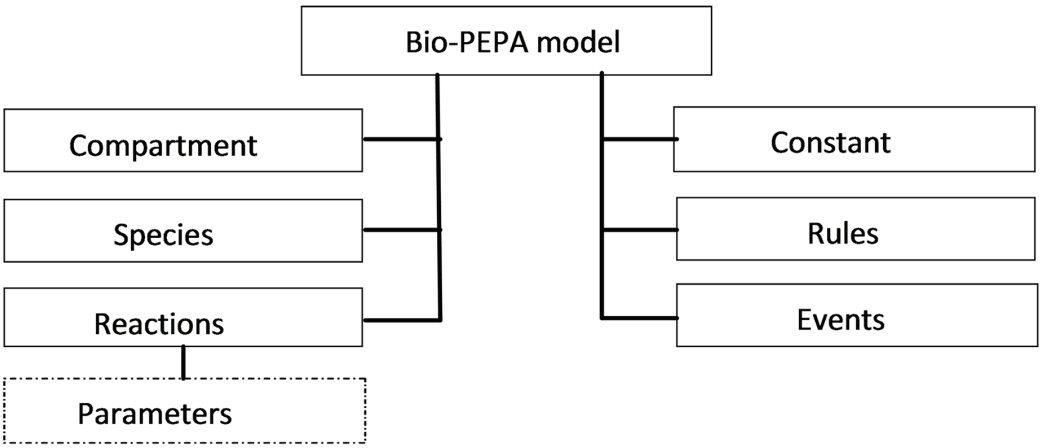

Figure 2. Modelling and simulation process

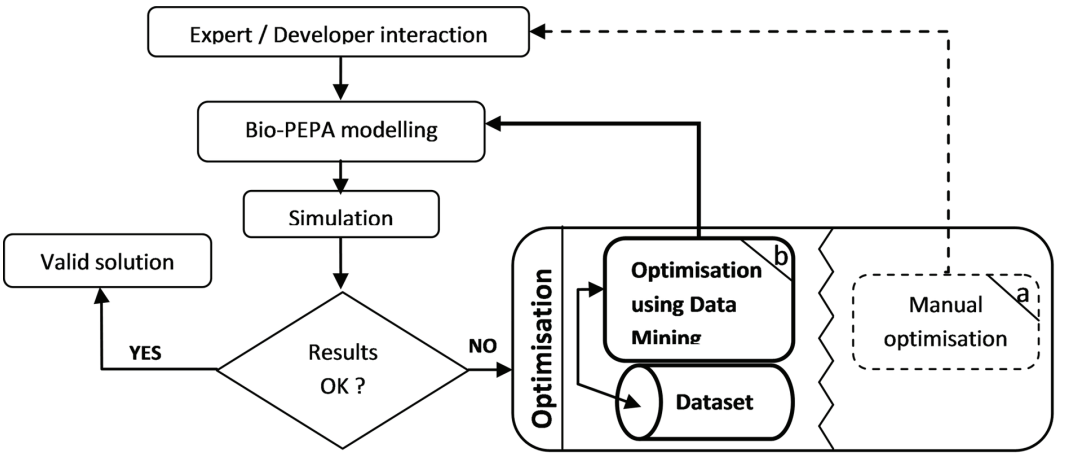

Figure 3. Simplified tuberculosis model

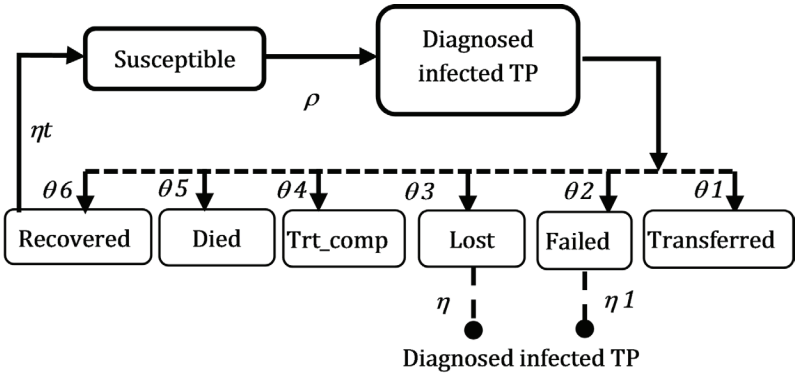

Figure 4. Tuberculosis Bio-PEPA Model

```
1   P= 0.000000075;
2   θ1 =0.00034;
3   θ2 =0.0032;
4   θ3 =0.00014;
5   θ4 =0.0011;
6   θ5 =0.00007;
7   θ6 =0.00069;
8   size-Population = 137990;
9   Location City : size = size-Population, type = compartment;
10  kineticLawOf susceptible_infected :  P* Susceptible@City* Infected@City;
11  kineticLawOf Failure :  θ1 * Infected@City;
12  kineticLawOf Recovery : θ2 * Infected@City;
13  kineticLawOf Transfer : θ3 * Infected@City;
14  kineticLawOf End_Treatment : θ4 * Infected@City;
15  kineticLawOf Death : θ5 * Infected@City;
16  kineticLawOf Loss : θ6 * Infected@City;
17  Susceptible = (susceptible_infected,1) << Susceptible ;
18  Infected =  (susceptible_infected,1) >> Infected  + (Recovery,1) << Infected +
    (Failure,1) << Infected + (Transfer,1) << Infected  + (End_Treatment,1) <<
    Infected  + (Death,1) << Infected + (Loss,1) << Infected;
19  Failed = (Failure,1) >> Failed ;
20  Recovered= (Recovery,1) >> Recovered ;
21  Transfere=(Transfer,1) >> Transfere ;
22  Trt_comp=(End_Treatment,1) >> Trt_comp ;
23  Died = ( Death,1) >> Died;
24  Lost = (Loss,1) >> Lost ;
25  Susceptible@City[137990]<*> Infected@City[15]<*> Failed@City[0]<*>
    Recovered@City[0]<*> Transfere@City[0]<*> Trt_comp@City[0]<*>  Died@City[0]<*>
    Lost@City[0]
```

Figure 5. Histograms for tuberculosis model for 2009

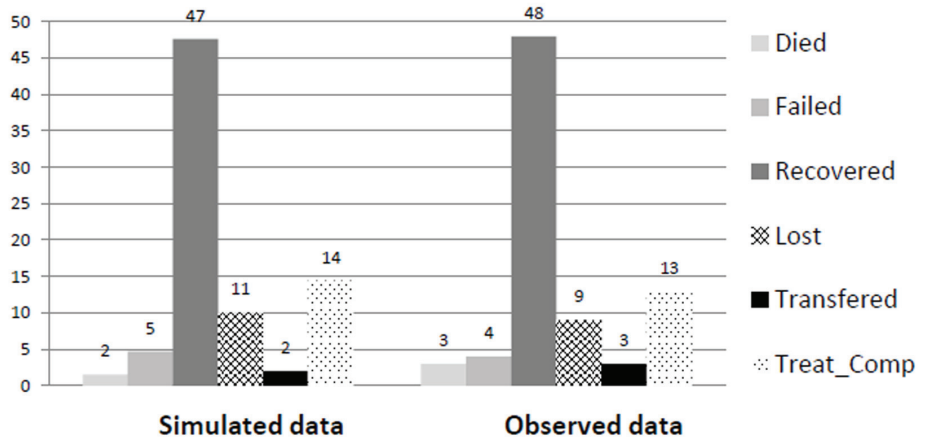

Figure 6. Histograms for tuberculosis model for 2010

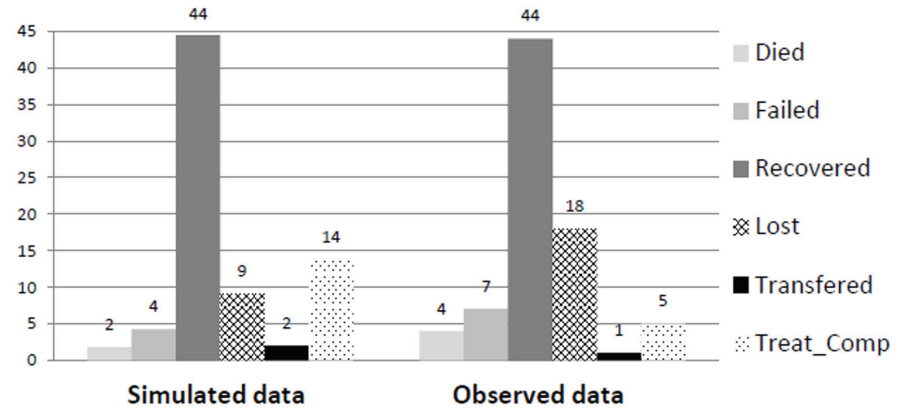

Figure 7. Histograms for tuberculosis model for Ain Tedles 2010

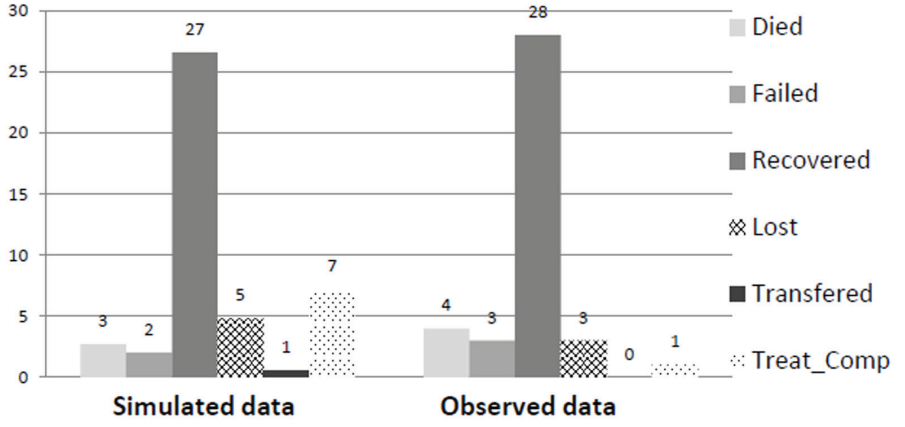

Figure 8. Histograms for tuberculosis model for Kheireddine 2010

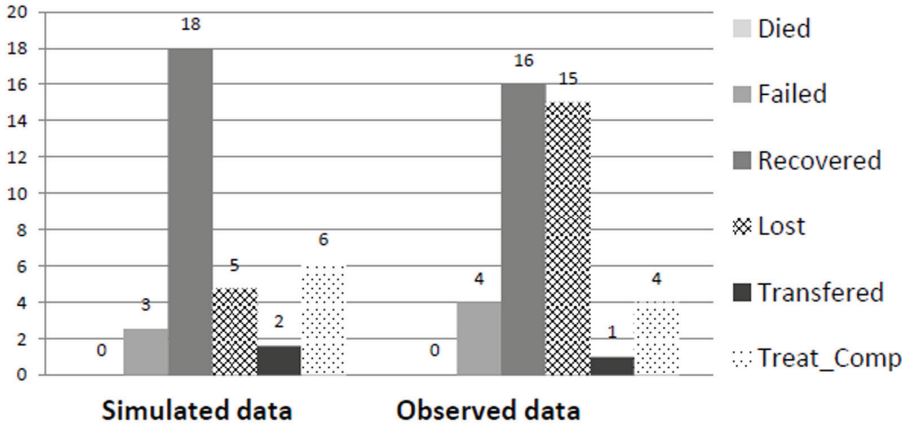

Figure 9. Updated tuberculosis model for Kheireddine Location 2010

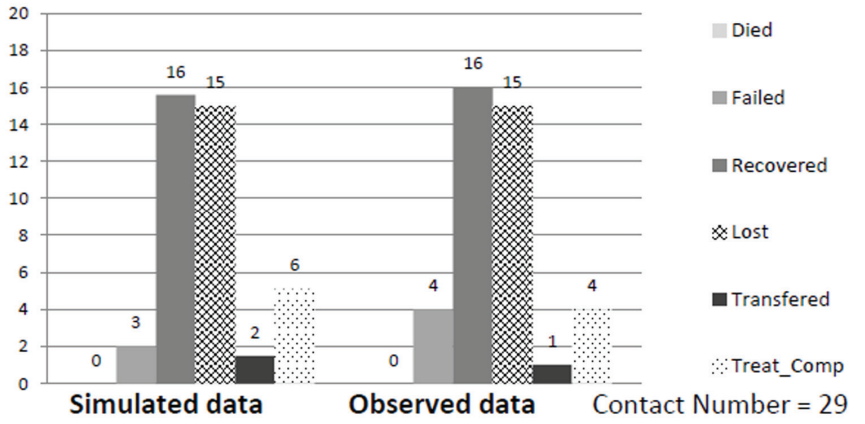

Figure 10. Final Histograms for tuberculosis model for 2010

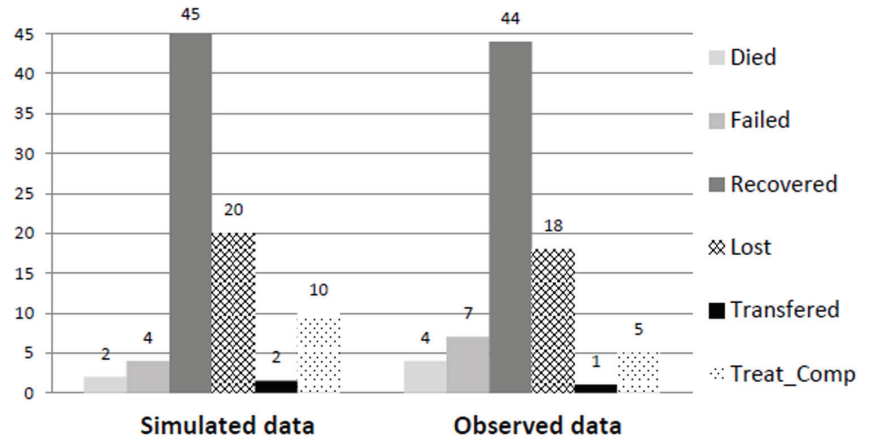

Figure 11. Histograms for tuberculosis model for 2011

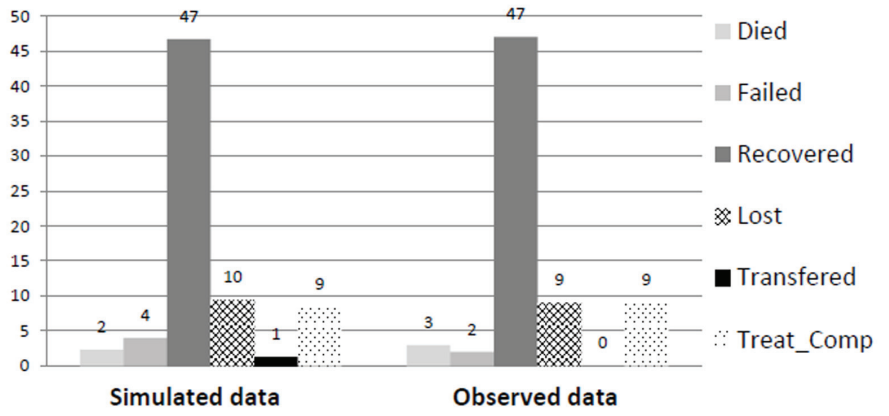

Figure 12. Histograms for tuberculosis model for 2012

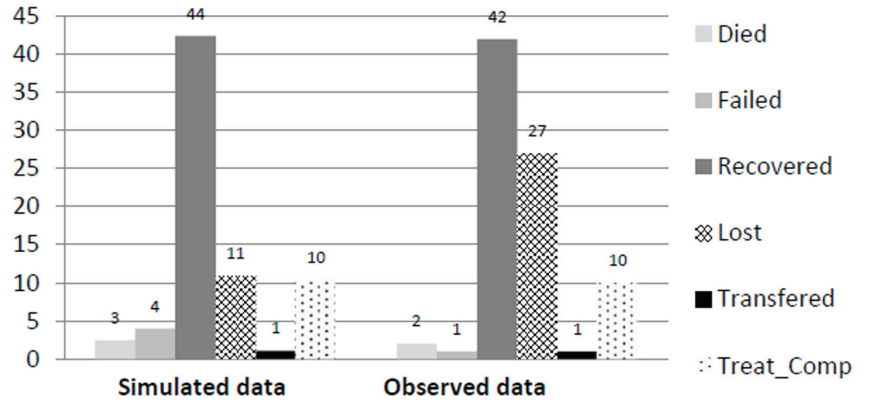

Figure 13. Histograms for tuberculosis model for Ain Tedles Location 2012

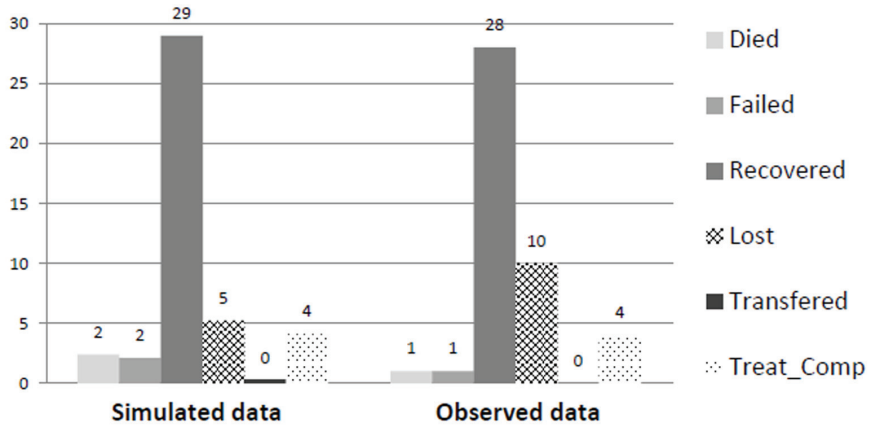

Figure 14. Histograms for tuberculosis model for Kheireddine Location 2012

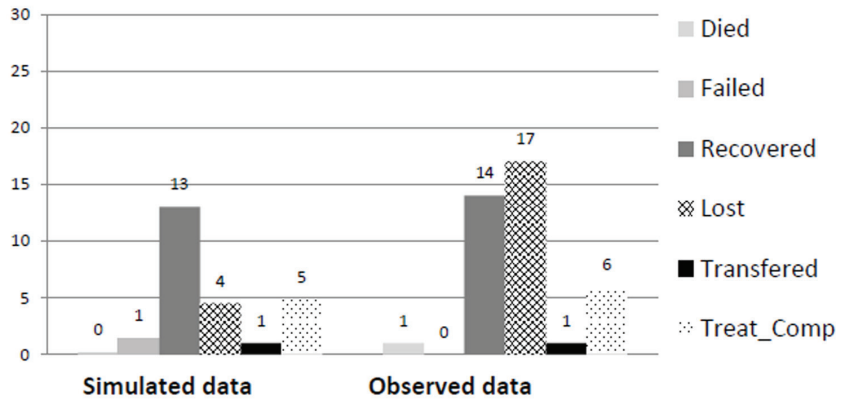

Figure 15. Updated tuberculosis model for Kheireddine Location 2012

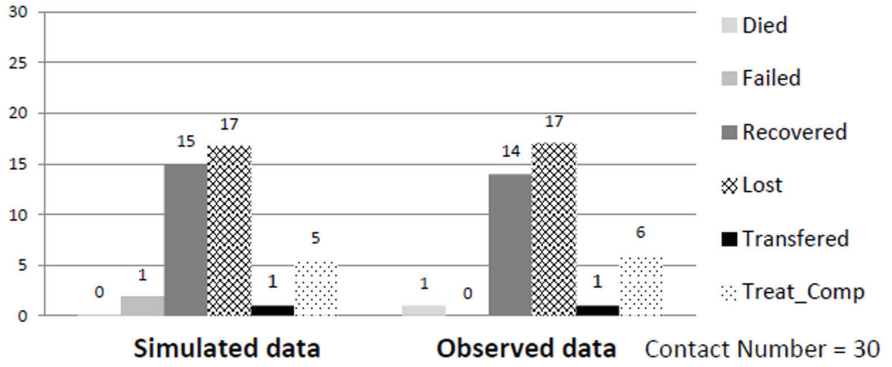

Figure 16. Final Histograms for tuberculosis model for 2012

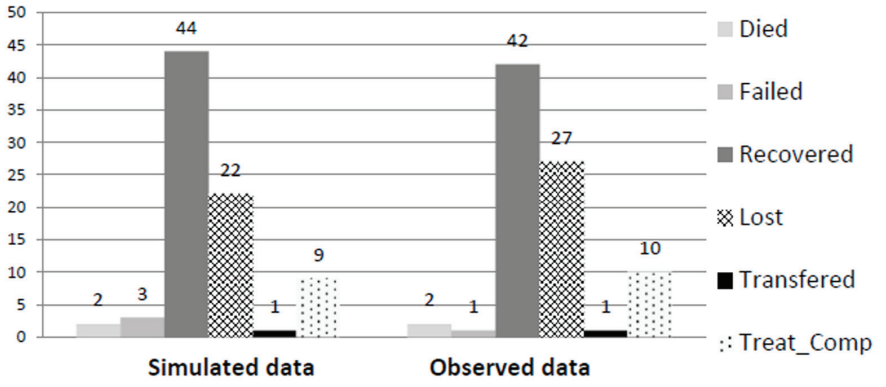

Table 1. Attributes and description

Attributes	Description
Attributes used in the study	
Age	Age of individual
Sexe	Male / Female
Mois	month of detection
Trim	season of detection
Annee	year of detecting disease
Com	City
Daira	municipality
adress	Flat nimber, Zip code, etc.
Date_Debut_TRT	Date of starting treatment
Typmal	New/Relapse/Failure/Other
Bacil1, Bacil2, Bacil3	Baciloscopy1, Baciloscopy 2, Baciloscopy 3. MM: negative, MP: positive, known as microscopy for Bacilli, test performed during six months of treatment by using a microscope to detect bacteria of tuberculosis in sputum samples. This test is used to manage mycrobacterial infections of tuberculosis.
AretTRT	State of patient at the end of treatment period: Lost: individual diagnosed but not treated, failed: treated but not recovered, Trt_comp: completed treatment without proving recovery, died, transferred: resistant TB, recovered.
Attributes not used in the study	
ID, First name, Last name RecStatus Loc DiagnoTP DiagnoTEP Preuve MalAsso	Identifier, first name of patient, last name of patient national identity number pulmonary or Extra pulmonary tuberculosis Patient diagnosed as a pulmonary tuberculosis Patient diagnosed as extra-pulmonary tuberculosis Examination of Extra pulmonary tuberculosis Other disease related to the record

Table 2. Model Parameters

Parameter	Description	Value	Formula
p	Rate of developing active pulmonary tuberculosis from susceptible state	**5.4 e^{-8}**	(Contact _ Number / Infection_Period)* Infection_Probability [1] (Keeling and Rohani, 2008).
θ 1	Transfer rate	**0.1 e^{-3}**	(1/ Infection_Period) * Transfer _Probability[1] (Keeling and Rohani, 2008).
θ 2	Failure rate	**3.7 e^{-4}**	(1/ Infection_Period) * Failure_Probability
θ 3	Lost rate	**8.5 e^{-3}**	(1/ Infection_Period)* Lost _Probability[1]
θ 4	Complete treatment rate	**7.8 e^{-4}**	(1/ Infection_Period)* Treatment_completed_Probability
θ 5	Death rate	**1.9 e^{-4}**	(1/ Infection_Period)* Death_Probability[1]
θ 6	Recovery rate	**4 e^{-4}**	(1/ Infection_Period)* Recovery_Probability[1]
ηt	Rate of recovered individual moving to susceptible state	**1**	All recovered move to Susceptible state.
η1	Rate of failure state transiting to infected TP state	**1**	All failed move to Infected state
η	Rate of lost transferred to Infected state	**1**	All lost move to Infected state
Contact Number	Contact with one infected case	**27**	Range over the interval [7,30]: possible freedom to vary these to fit observed data (Aparicio and Castillo-chavez, 2009)
Infection Period (month)	The period during which the virus can be transmitted	**6**	Range over the interval [6,24]: possible freedom to vary these to fit observed data (Aparicio and Castillo-chavez, 2009)

[1]the probabilities are calculated from observed data.

Table 3. Description the uses of data by year in Bio-PEPA process

Year of prediction	2009	2010	2011	2012
Set of years used	2008	2008-2009	from 2008 to 2010	from 2008 to 2011

Table 4. Description the uses of data by year in data mining process

Year of prediction	2010	2012
Set of years used	From 2008 to 2010	from 2008 to 2012

Table 5. Confusion matrix and class wise accuracy of Ladtree algorithm

	Predicted classes					
Class label	**Died**	**Failed**	**Recovered**	**Lost**	**Transferred**	**Trt_comp**
Died	0	0	1	7	0	0
Failed	0	0	2	10	0	4
Recovered	1	0	136	0	0	2
Lost	0	1	1	30	0	5
Transferred	0	0	1	3	2	0
Trt_comp	0	1	12	15	1	5
True positive rate	0	0	0.978	0.811	0.333	0.147
False Positive Rate	0.004	0.009	0.168	0.172	0.004	0.053

Table 6. Confusion matrix and class wise accuracy of j48 algorithm

	Predicted classes					
Class label	**Died**	**Failed**	**Recovered**	**Lost**	**Transferred**	**Trt_comp**
Died	**0**	**0**	**1**	**11**	**1**	**0**
Failed	**0**	**5**	**4**	**10**	**0**	**0**
Recovered	**0**	**0**	**227**	**0**	**1**	**0**
Lost	**0**	**1**	**6**	**66**	**0**	**0**
Transferred	**0**	**1**	**1**	**2**	**3**	**0**
Trt_comp	**0**	**0**	**19**	**33**	**1**	**0**
TP rate	**0**	**0.005**	**0.996**	**0.904**	**0.429**	**0**
FP Rate	**0**	**0.263**	**0.188**	**0.175**	**0.008**	**0**

Dalila Hamami is a PhD student at Computing Science Department, Oran University (Algeria) in collaboration with School of Natural Science, Stirling University (Stirling). She completed her Master's in Computing science in 2007. Her research interests include modelling, simulation, data mining and optimization and decision support systems. She is currently assistant lecturer at Computing science and mathematics department, University of Abdelhamid Ibn Badis, Mostaganem, Algeria.

Baghdad Atmani is a professor of Computing Science at the University of Oran. His field of interests are Data Mining and Machine Learning Tools. His research is based on Knowledge Representation, Knowledge-based Systems and CBR, Data and Information Integration and Modelling, Data Mining Algorithms, Expert Systems and Decision Support Systems. His research is guided and evaluated through various applications in the field of control systems, scheduling, production, maintenance, information retrieval, simulation, data integration and spatial data mining.

Carron Shankland is a professor of Computing Science at the University of Stirling, and deputy head of the school of Natural Sciences. Her research lies in the intersection of computer science, mathematics and biology: understanding the behaviour of biological systems through mathematical and computational models. Her models (in process algebra) can describe systems at a high level of abstraction as networks of communicating individuals, scaling up to the emergent population dynamics. Her group has worked across a range of biological systems (disease dynamics, immunological systems, collective dynamics of cells, cell signaling response to cancer therapies) as well as in computer networks and protocols. In addition, her group is developing an exciting technique combining genetic programming with modelling to produce models directly from experimental data. Prof Shankland leads activities in the modelling and abstraction theme in the Scottish Computing community, and nationally co-leads the POEMS network linking modelling to healthcare technology.

Call for Articles

International Journal of Information Systems in the Service Sector

Volume 9 • Issue 2 • April-June 2017 • ISSN: 1935-5688 • eISSN: 1935-5696
An official publication of the Information Resources Management Association

MISSION

The **International Journal of Information Systems in the Service Sector (IJISSS)** provides a significant channel for practitioners and researchers (from both public and private areas of the service sector), software developers, and vendors to contribute and circulate ground-breaking work and shape future directions for research. IJISSS assists industrial professionals in applying various advanced information technologies. It explains the relationship between the advancement of the service sector and the evolution of information systems.

COVERAGE AND MAJOR TOPICS

The topics of interest in this journal include, but are not limited to:

Business services • Creative problem solving • Customer value and customer relationship management • Data warehousing and mining in services • Decision-making under uncertainty • Decision-support systems • E-business in service industries • Economic analysis and organizational behavior • Forecasting, planning, scheduling, and control • Green service and sustainability • Hospitality and tourism information systems • Information technology in services • Knowledge Management • Logistics network configuration • Matching supply with demand • Multiple-objective decision making • Optimization of service systems • Performance measures and quality control • Public service management • Revenue and risk management • Self-service systems • Service business models • Service information systems • Service practice, productivity, and innovation • Service systems simulation • Service-oriented architecture • Service-oriented computing • Solid or soft modeling and analysis • Strategic information systems and strategic alliances • Supplier relationship management • System analysis of service industry • Web Services

ALL INQUIRIES REGARDING IJISSS SHOULD BE DIRECTED TO THE ATTENTION OF:

John Wang, Editor-in-Chief • IJISSS@igi-global.com

ALL MANUSCRIPT SUBMISSIONS TO IJISSS SHOULD BE SENT THROUGH THE ONLINE SUBMISSION SYSTEM:

http://www.igi-global.com/authorseditors/titlesubmission/newproject.aspx

IDEAS FOR SPECIAL THEME ISSUES MAY BE SUBMITTED TO THE EDITOR(S)-IN-CHIEF

PLEASE RECOMMEND THIS PUBLICATION TO YOUR LIBRARIAN

For a convenient easy-to-use library recommendation form, please visit:
http://www.igi-global.com/IJISSS