The Importance of Health Informatics in Public Health during a Pandemic
J. Mantas et al. (Eds.)

doi:10.3233/SHTI200516

PrositNG – A Machine Learning Supported Disease Model Generation Software

Monika POBIRUCHIN [a,1], Richard ZOWALLA [b],
Maximilian KURSCHEIDT [b] and Wendelin SCHRAMM [a,b]
[a] *GECKO Institute, Heilbronn University, Heilbronn, Germany*
[b] *Center for Machine Learning (ZML), Heilbronn University, Heilbronn, Germany*

Abstract. Decision models (DM), especially Markov Models, play an essential role in the economic evaluation of new medical interventions. The process of DM generation requires expert knowledge of the medical domain and is a time-consuming task. Therefore, the authors propose a new model generation software *PrositNG* that is connectable to database systems of real-world routine care data. The structure of the model is derived from the entries in a database system by the help of Machine Learning algorithms. The software was implemented with the programming language Java. Two data sources were successfully utilized to demonstrate the value of *PrositNG*. However, a good understanding of the local documentation routine and software is paramount to use real-world data for model generation.

Keywords. medical economics, markov processes, real-world data, machine learning, electronic health records

1. Introduction

1.1. Background

Health economic decision models (DM) play an essential role in the evaluation of new treatments, drugs or prevention programs. DMs depict the 'typical' progress of a disease or condition, e.g., cancer or diabetes. Markov Models (MM) are one of the most frequently used methods in the health economics domain [1]. An MM's structure consists of several disease states S={S1,...Sn} and a maximum of n2 transitions between them. Disease states should be defined as mutually exclusive and cumulatively exhaustive, i.e., each state should be a distinctive, well-defined step in the total disease process [2]. Trust, confidence and transparency are critical for DMs [3]. However, from a modeler's perspective the effort to conceptualize, implement and validate a health economic model is of equal importance. Today, the process of DM generation is performed manually, supported by modeling suites such as TreeAge by TreeAge Software, Inc. [4] or general-purpose spreadsheet programs (SPM, e.g. Excel or LibreOffice Calc [5]). However, these software suites and their resulting DMs are not connected to real-world data sets or database systems, e.g., a hospital information system or a registry information system.

[1] Corresponding Author, Monika Pobiruchin, GECKO Institute, Heilbronn University, Heilbronn, Germany; E-mail: monika.pobiruchin@hs-heilbronn.de.

Thus, medical progress has to be incorporated laboriously into a model, which leads to high development and maintenance costs.

Therefore, the authors propose a new model generation software that offers an interface to real-world medical database systems. As published previously, the structure of the model is derived from the entries in a database system by the help of unsupervised Machine Learning (ML) algorithms [2]. This DM generation software builds on the methodological groundwork of the PROSIT Online Disease Modelling Community [6], hence the name *PrositNG* (Next Generation) has been chosen.

1.2. Requirements

Persons who construct new DMs ('modelers') should be supported during the whole DM generation process. This includes the following steps: i) definition of disease states, ii) calculation of transition probabilities, iii) definition of cost & utility data, iv) performing the cohort simulation. Furthermore, *PrositNG* should enable modelers to 'experiment' with the given data, i.e., evaluate the influence of specific parameters or parameter values on the selection of patient cohorts, transition probabilities or health economic outcomes. This way, new hypotheses could be generated which can then be tested and examined in further steps or even addressed in new studies and clinical trials. Input of *PrositNG* should be configurable to support different kind of data sources or databases, e.g., CSV (comma-separated value) files, databases from disease registries, hospital information systems, etc.

2. State of the art

Today, the conceptualization of new MMs is in most cases done manually by pen and paper. SPMs can be used to implement MM calculations and cohort simulations. However, DMs built with SPMs cannot incorporate real-world patient data and are not connected to database systems from healthcare facilities. On the other hand, they are considered to be most transparent and widely available office software is sufficient when working with them. Specialized modeling software suites like TreeAge offer more features than SPMs and support additional modeling techniques, e.g., decision trees and microsimulation. TreeAge is capable to use patient data stored in an Excel/CSV file for bootstrapping and for input parameters. However, it offers no ML-based support for the selection of input parameters.

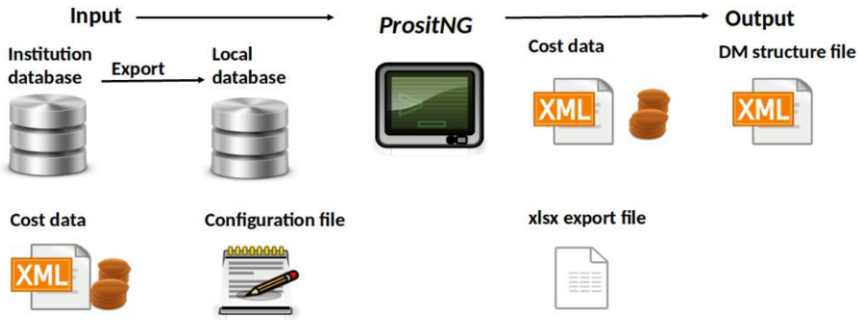

Figure 1. Input and ouput resources of the PrositNG modeling software suite.

3. Concept

Databases from health care providers as well as public/scientific use files should be integrated in *PrositNG*; this is achieved via a configuration file. In addition, the output of *PrositNG*, i.e., derived health economic models and results of cohort simulations should be exportable in an open, non-proprietary format (see Fig. 1). Unsupervised ML algorithms should support the construction steps of the DM: Cluster algorithms can pinpoint to meaningful combinations of parameters, i.e., valid definitions of disease states, in the real-world data set. Next, frequent combinations are used to derived transitions and their respective probabilities (details about the algorithms for model generation can be found in [2]).

4. Implementation

PrositNG was implemented with the programming language Java. PostgreSQL was chosen as local database system (see Fig. 1). The local database consists of i) a 'flat' primary table (one row represents exactly one case) and ii) a secondary table for progress information that is associated with a specific date field given in i). Currently, two data sources were successfully utilized: a) Local Cancer Registry [2], b) Study data from a Type 2 Diabetes screening program. The user interface is implemented with JavaFX components (see Fig. 2). We used the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) implementation of the Weka collection [7] for clustering clinical parameters. DBSCAN allows for the specification of a minimum number of associated cases in a specific cluster.

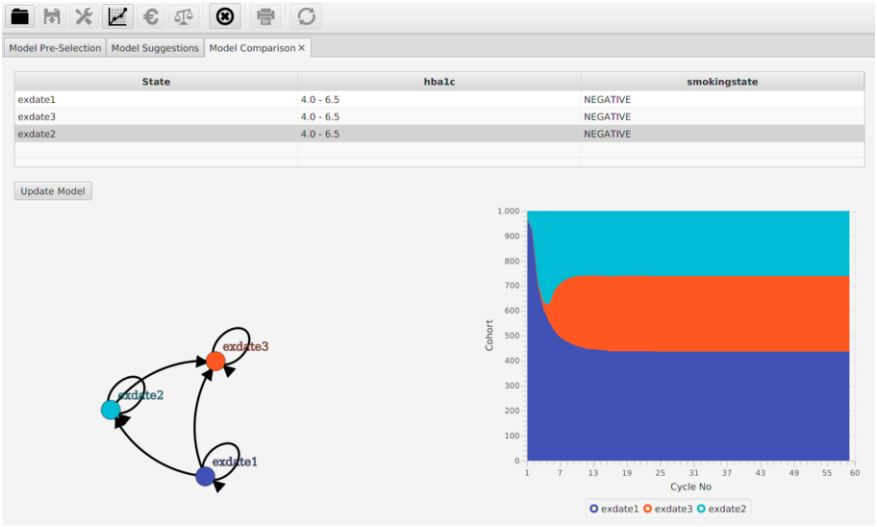

Figure 2. Screenshot of *PrositNG – example result view*: structure of 3-state DM (exdate1, exdate2, exdate3 indicate appointments for screening examinations, lower left corner), cohort simulation (x-axis: monthly cycles, y-axis: cohort members associated with a particular state; lower right corner). Data source: Type 2 Diabetes screening program.

5. Lessons learned (Discussion)

Cluster algorithms, like the used DBSCAN, can support in the selection of parameters based on frequency. Using frequent parameters could add robustness to a model or help less-experienced modelers. However, for an experienced modeler with a strong medical (documentation) background a software-generated suggestion for a cancer state like `T=3, N=0, M=0, surgery=true, radiation=false` might be trivial as this equals well-known cancer stadium definitions. Furthermore, when relying only on frequently documented parameters *PrositNG* could not support the discovery of yet unknown correlations. This is a conflict we are still trying to solve. A good understanding of the local documentation routine and the documentation (software) system is paramount to correctly leverage real-world data for the construction of DMs. Modelers need a basic understanding of the underlying data structure and generation algorithms of *PrositNG* to evaluate the validity of the resulting models. In addition, such knowledge is needed to evaluate if and how new data sources can be leveraged for automatic model generation. Human intelligence is still indispensable.

6. Conclusions

PrositNG supports the process of automatic model generation, i.e., to construct disease states based on real-world data and to derive transitions. We could demonstrate that results of clustering can support modelers in the definition of a disease state. However, there are certain requirements for data sources that have to be met, e.g., precise date information, longitudinal data. Today, many publicly available scientific used files do not meet these criteria. Furthermore, consulting documentation staff is necessary to understand the content and meaning (semantics) of the documented data. Future work could include the development of *PrositNG* as a web-based solution as well as an online showcase for interested modelers.

References

[1] Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T, Vray M, Modelling in economic evaluation: an unavoidable fact of life, Health Econ 6 (1997), 217–227.
[2] Pobiruchin M, Bochum S. Martens UM, Kieser M, Schramm W, A method for using real world data in breast cancer modeling, J Biomed Inform 60 (2016), 385-394.
[3] Kent S, Becker F, Feenstra T, Tran-Duy A, Schlackow I, Tew M et al, The challenge of transparency and validation in health economic decision modelling: a view from Mount Hood, Pharmacoeconomics 37 (2019), 1305-1312.
[4] TreeAge Software, Inc. Home - TreeAge Software [Internet]. Williamstown: TreeAge Software, Inc; 2020 [cited 18-03-2020]. Available from: https://www.treeage.com/.
[5] LibreOffice The Document Foundation. LibreOffice – Free Office Suite – Fun Project – Fantastic People [Internet]. Berlin: The Document Foundation; 2020 [cited 18-03-2020]. Available from: https://www.libreoffice.org/
[6] Schramm W, Sailer F, Pobiruchin M, Weiss C, PROSIT Open Source Disease Models for Diabetes mellitus, Stud Health Technol Inform 226 (2016), 115-118.
[7] Witten IH, Eibe F, Hall MA, Pal CJ, Data mining: practical machine learning tools and techniques, Elsevier, Morgan Kaufmann, Amsterdam, 2017.