Journal of Biomedical Informatics 60 (2016) 385–394

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A method for using real world data in breast cancer modeling

Monika Pobiruchin [a,*], Sylvia Bochum [b], Uwe M. Martens [b], Meinhard Kieser [c], Wendelin Schramm [a]

[a] Heilbronn University, GECKO Institute for Medicine, Informatics and Economics, Max-Planck-Str. 39, 74081 Heilbronn, Germany
[b] SLK-Hospitals, Cancer Center Heilbronn-Franken, Am Gesundbrunnen 20-26, 74078 Heilbronn, Germany
[c] University of Heidelberg, Institute of Medical Biometry and Informatics, Im Neuenheimer Feld 305, 69120 Heidelberg, Germany

ARTICLE INFO

Article history:
Received 7 May 2015
Revised 23 January 2016
Accepted 31 January 2016
Available online 8 February 2016

Keywords:
Secondary use
Disease model
Real world data
Cancer registry
Markov model

ABSTRACT

Objectives: Today, hospitals and other health care-related institutions are accumulating a growing bulk of real world clinical data. Such data offer new possibilities for the generation of disease models for the health economic evaluation. In this article, we propose a new approach to leverage cancer registry data for the development of Markov models. Records of breast cancer patients from a clinical cancer registry were used to construct a real world data driven disease model.

Methods: We describe a model generation process which maps database structures to disease state definitions based on medical expert knowledge. Software was programmed in Java to automatically derive a model structure and transition probabilities. We illustrate our method with the reconstruction of a published breast cancer reference model derived primarily from clinical study data. In doing so, we exported longitudinal patient data from a clinical cancer registry covering eight years. The patient cohort ($n = 892$) comprised HER2-positive and HER2-negative women treated with or without Trastuzumab.

Results: The models generated with this method for the respective patient cohorts were comparable to the reference model in their structure and treatment effects. However, our computed disease models reflect a more detailed picture of the transition probabilities, especially for disease free survival and recurrence.

Conclusions: Our work presents an approach to extract Markov models semi-automatically using real world data from a clinical cancer registry. Health care decision makers may benefit from more realistic disease models to improve health care-related planning and actions based on their own data.

1. Introduction

Breast cancer is the most leading cause of cancer death in women worldwide. More than one million patients are diagnosed with breast cancer every year [1]. In the last decades, targeted therapies for HER2 (human epidermal growth factor receptor 2)-positive cancer were developed to improve the prognosis for these patients. HER2 is a transmembrane receptor tyrosine kinase expressed in epithelial cells as found in breast tissue. Overexpression of the HER2 protein or amplification on the HER2 gene occurs in about 15–25% of all breast cancer cases [2]. Both indicate aggressive growth and spreading of the tumor. Several health economics evaluations assessed the cost-utility and cost-effectiveness of Trastuzumab, a humanized monoclonal antibody, in the treatment of HER2-positive women [3–6].

In health economics, disease models (DM) are widely used to conduct evaluations of new treatment plans, medications or prevention programs [7]. Health care facilities use an increasing number of information systems to document the treatment of patients. However, these systems also collect data for "such activities as analysis, research, quality and safety measurement, public health, payment, provider certification or accreditation, marketing, and other business applications, including strictly commercial activities" [8]. This approach is referred to as *secondary use of data*, i.e., data are re-used in a different context. It is general consensus that such clinical data have "significant potential to facilitate research, improve quality of care for individuals and populations, and reduce healthcare costs" [9]. Before secondary use can tap its full potential issues like data stewardship principles, reduction and elimination of data silos and guaranteed patient privacy must be addressed [9].

Several US [10] and European [11] research initiatives also from scientific societies such as ISPOR (International Society for Pharmacoeconomics and Outcomes Research) focus on the usage of "real

* Corresponding author. Tel.: +49 (0)7131 504 633.
E-mail addresses: monika.pobiruchin@hs-heilbronn.de (M. Pobiruchin), sylvia.bochum@slk-kliniken.de (S. Bochum), uwe.martens@slk-kliniken.de (U.M. Martens), meinhard.kieser@imbi.uni-heidelberg.de (M. Kieser), wendelin.schramm@hs-heilbronn.de (W. Schramm).

http://dx.doi.org/10.1016/j.jbi.2016.01.017
1532-0464/

world data" for comparative effectiveness research and health services research [12,13]. These initiatives demand for real world data to assess the impact of new medical interventions in routine care settings. Especially in disease modeling, an increased usage of real world data could bridge the gap between efficacy and effectiveness, i.e., the relation of cost to (added) value under routine conditions [14].

This article presents a 4-step model generation process (Definition – Selection – Transformation – Generation) to derive a Markov model from a cancer registry in a semi-automatic manner. We illustrate our approach in a validation study by re-constructing a previously published DM by Blank et al. [3] with data from routine care.

We demonstrate how inclusion and exclusion criteria from the randomized Herceptin Adjuvant Trial (HERA) are used to select a suitable cohort of patients [15]. Moreover, definitions of disease states are mapped to database tables from a tumor documentation system. To assess differences between the results obtained from real world data and data collected under optimal conditions of a randomized clinical trial (RCT), transition probabilities of the computed model are compared with probabilities derived from a breast cancer RCT.

2. Background

In general, the structure of a Markov DM consists of disease states and transitions. Typically, the probabilities of these transitions should be derived from the most representative data sources, e.g., population-based epidemiological studies, systematic-reviews, or meta-analyses. If none of the above are available, RCTs are being used regularly [16].

A DM based on patient records captured in a large clinical database instead of RCTs could improve DMs to better reflect the clinical reality. That means the efficacy of interventions can be confirmed with data from clinical routine conditions [17]. Health care decision makers could improve the development and allocation of measures to the health care system based on more realistic forecasts [12]. As a consequence, hospitals and disease registries should not use their data for the single purpose of patient care but to "dig for hidden gold" and "uncover the treasures buried in hospital medical records" [18]. This could lead to continuous improvement in health and health care what is commonly referred to as "the learning health system" [19]. The approach presented in this work supports this paradigm as it makes use of a real world, health care data collection.

3. Materials and methods

In health economics, the most common model types are decision trees and Markov models [20]. Our study concentrates not on decision trees but on Markov models because these can handle more complex and time-depending scenarios. The structure of a Markov model consists of $S = \{s_1, s_2, \ldots, s_n\}$ disease states and a maximum of n^2 transitions between them. States are specified as mutually exclusive, i.e., every patient is only allocated to clearly distinctive disease conditions ("no overlaps"). Moreover, the states are collectively exhaustive, i.e., the complete disease history is illustrated in the model structure ("no gaps"). While passing through the model every patient can be allocated to only one particular disease state at any given point in time. In health economic modeling every state is assigned with specific costs and utility values, e.g., quality of life. Cohort simulations are performed to capture all costs and utilities for a cohort of patients which is distributed across all the states during the desired time horizon H [16]. While passing through the different states of the DM the cohort collects costs and utilities. These values can be summed up for further analysis, e.g., cost-effectiveness evaluation.

3.1. Cancer registry

The clinical cancer registry of the Cancer Center (CC) Heilbronn-Franken is operated by the SLK-Hospital Holding, located in southern Germany. Its enrollment population amounts to about 1 million people. In the region of Heilbronn-Franken, the SLK-Hospital Holding is the largest hospital owner. This implies that most cancer cases in the administrative district are treated there [21].

Documentation in the CC started in 1986. Today, the database comprises about 60,000 cases. Since 2001, the Giessen Tumor Documentation System (GTDS) [22] is used in the CC. GTDS was developed with funding support by the Federal Ministry of Health of Germany in the early 1990s. It is used in about 60 cancer registries and is updated regularly to meet new standards and requirements [23]. GTDS is managed via a graphical user interface which is used in daily routine at the CC or for simple analyses. For more complex reports or record exports, a command-line interface is available. Both ambulatory and stationary follow-up assignments are recorded in the GTDS-system.

3.2. Model generation process

We defined a process to generate a Markov model what we refer to as 'semi-automatically', see Fig. 1. In this context, semi-automatically means that the selection of meaningful clinical parameters is done manually, see Step 1. All other steps in the generation process are performed automatically, e.g., generation of database queries or probability calculations. A detailed description of the process is given in the subsequent sections The model generation software was developed in Java to support the process and to visualize the results.

3.2.1. Step 1 – definition

At first, the disease states of the Markov model have to be defined. This is based on physiological knowledge about the disease, its progression and best practices. In order to use a database for the generation of a Markov model structure, the medical definition of a state hast to be mapped to database columns and their specific values. An example for such a mapping can be found in Table 2.

It is mandatory to confer with physicians, documentation officers, study nurses, etc., to obtain the best mapping between medical state description and database structure.

3.2.2. Step 2 – selection

The result of Step 1, i.e., the previously defined disease state mapping, is used to automatically generate SQL (Structured Query Language) statements to query the related dataset. Each disease state is reflected in one specific SQL statement. If a patient is selected by a query, the related patient record applies to the associated disease state. Only those selected patients are considered in Step 3 and Step 4. Patients can be matched to $[1 \ldots n]$ states, which reflect the progression of the disease over time. For example, if a patient is declared as disease free, he/she may have a relapse and may die after another five years. This procedure would then yield three matching disease states (Disease free – Recurrence – Death) for this particular patient.

3.2.3. Step 3 – transformation

Selected patients from Step 2 and their associated $[1 \ldots n]$ disease states are sorted by the date of the observation. This way, a sequence of events is derived for every patient. Usually these events are scattered over the entire time horizon H of the model