eHealth2014 – Health Informatics Meets eHealth
A. Hörbst et al. (Eds.)

doi:10.3233/978-1-61499-397-1-133

Case-based Visualization of a Patient Cohort using SEER Epidemiologic Data

Christian MAIER[a,1], Thomas BÜRKLE[a], Hans-Ulrich PROKOSCH[a,b] and Thomas GANSLANDT[b]
[a] *Chair of Medical Informatics, Friedrich-Alexander-University Erlangen/Nuernberg*
[b] *Department of Medical Information and Communication Technology, Erlangen University Hospital*

Abstract. Data from cancer registries can be used to track the epidemiology of cancer and can potentially serve to guide individual diagnostic and treatment decisions. Even though some cancer registry datasets have been made publicly available for scientific and clinical use, few applications have so far provided direct access to these data from within the patient context of an electronic patient record. The goal of this project was to implement a proof-of-concept integration of the public SEER (Surveillance, Epidemiology and End Results) cancer registry dataset with a digital breast cancer tumor board at a German university hospital and to determine its utility in the clinical settings. The integration was successfully established, using data from routine documentation to provide dynamic visualizations of cohort composition and Kaplan-Meier survival plots. Evaluation feedback was favorable regarding the concept and implementation, but highlighted that important data elements, e.g. receptor status data, were missing in the SEER dataset, limiting clinical value of the system.

Keywords. Decision Support Systems, Clinical; Electronic Health Records; Breast Neoplasms; Kaplan-Meier Estimate; User-Computer Interface.

1. Introduction

According to the German Federal Statistical Office, cancer is the second most common cause of death in Germany [1]. Collecting diagnostic and outcome data of cancer patients in clinical or epidemiologic cancer registries permits to track the epidemiology of cancer and can potentially serve as a guide for individual treatment decisions. But today these resources are still primarily used for epidemiology, quality management, guideline development and policy decisions, but much less on an individual patient level.

On one hand, a reason for this fact may be the restricted public availability of detailed single patient cancer registry data. The American SEER dataset (Surveillance, Epidemiology and End Results) [2], however, is one of the large cancer-registry databases that already offer their data for academic use. But even with availability of a detailed dataset, another challenge is to make this information available at the point of care for clinicians. Today, only a small number of applications exist that apply local patient data to query an epidemiological cancer registry dataset and provide the

1 Corresponding Author: chrisp.maier@arcor.de

clinician with decision support regarding further therapy or prognosis [3-5]. Many of these resources such as the prostate predictive nomograms on the Sloan Kettering website [6] are not directly integrated into clinical information systems and the physicians workflow (requiring manual re-entry of data), and thus less used than desirable, although they might deliver a significant advantage for individual medical decision making. Integration of decision support tools into the clinical workflow has been shown to improve their uptake [7].

Considering these facts, the goal of this project was to design and implement a proof-of-concept integration of a subset of the SEER database in a routine clinical information system in order to supply individualized epidemiology information for breast cancer patients. Relevant steps towards this goal include an analysis of both the local electronic patient record (EPR) as well as the SEER breast cancer dataset, a mapping between related data items from both sources, implementation of a suitable visualization and interviews with clinical users. Breast cancer was chosen due to the large amount of clinical studies performed in this department and a previous series of physician interviews indicating interest in the topic.

2. Methods

2.1. Environment

Erlangen University hospital (EUH) is a 1316 bed academic maximum care facility, as well as a comprehensive cancer center supported by the German Cancer Aid and a German Cancer Society (DKG) certified oncological center. EUH uses the Siemens Soarian Clinicals™ EPR system which is interfaced to a variety of departmental information systems such as laboratory and pathology systems. Soarian Clinicals is a configurable web-based system which supports definition of new documentation forms in a form generator [8]. At EUH this feature has been used to enhance the EPR for the documentation of digital interdisciplinary tumor board meetings that include the collection of a comprehensive set of data elements for different cancer entities [9,10]. The EPR also provides methods to invoke external web-based applications, including the transmission of patient-related data items.

2.2. The EPR dataset, SEER dataset and mapping

The EPR-based tumor board documentation in the gynecology department comprises a set of four online forms to document the patient history, to register a patient to the tumor board, to document the pathology results and to document the tumor board decision. Most relevant for this project is the Pathology form, which is filled by the pathology department with information available before the tumor board meeting (e.g. from biopsies or surgery) and supplemented with additional diagnostic information (e.g. regarding metastases). Fig 1 shows the structured pathology documentation form which is completed by the pathology department prior to the tumor board meeting.

The current “Surveillance, Epidemiology and End Results” (SEER) breast cancer dataset includes more than seven million data records for different cancer entities covering cases from 1973 until 2010 [2]. Every case equals one anonymized data record consisting of 120 data fields comprising items such as diagnosis according to the

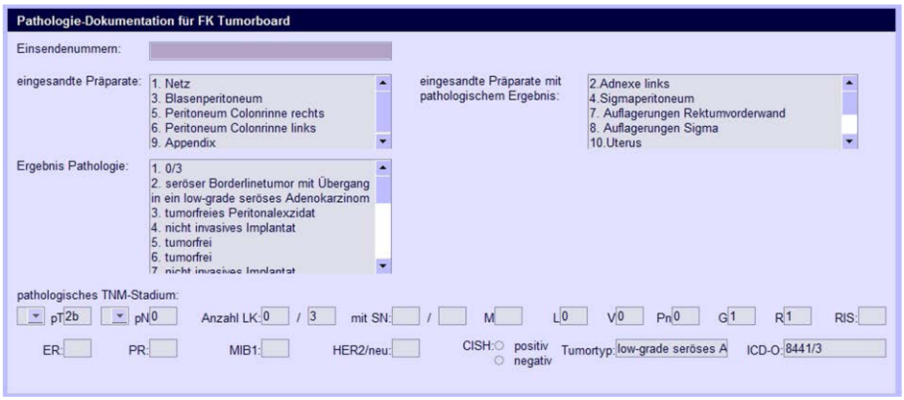

Figure 1.-Screenshot of EPR form for pathology documentation

International Classification of Diseases (ICD, version 10), tumor grading, TNM (tumor patients 1973-2008 comprising more than 436.000 cases was used. Mapping was performed for the smallest common denominator, i.e. those data items available in sufficient quality in the EPR and the SEER dataset. For each item, also the value sets were mapped between EPR and SEER.

2.3. Implementation

The SEER dataset was extracted from its fixed width-based text format. Relevant records and fields were selected and loaded into an Oracle 11g™ database, using the Talend Open Studio™ platform. To provide SEER visualizations linked to EPR patient data, an external application was implemented which is invoked from a dedicated EPR form. Fig. 2 shows the overall application structure.

Since data elements relevant for SEER case selection are spread throughout several gynecology tumor board forms, a dedicated "interface" form was designed that bundles all relevant elements. All data items already available from filled-out tumor board forms are pre-populated, and the remaining fields can be manually complemented. The EPR provides a mechanism to collate all data elements and construct a hyperlink to invoke the external visualization application with all necessary parameters. The visualization frontend parses the parameters contained in the hyperlink and requests aggregated counts of the matching SEER patient cohort from the backend using a webservice call. All available data elements from the interface form are used as selection criteria when querying the SEER dataset. The frontend was implemented using HTML and JavaScript as well as the JSONSS framework [11]. Dynamic pie-chart visualizations were realized with the Highcharts JavaScript library [12]. Kaplan-Meier-plots were implemented with an Oracle stored function to pre-process the relevant data items and a JavaScript function that made use of low-level functions of the Highcharts library.

2.4. Evaluation

The implementation was evaluated by a convenience sample of 5 medical doctors from the EUH gynecology and obstetrics department (unpaid, voluntary). Evaluation consisted of a 10 minute demonstration of the system and a structured interview

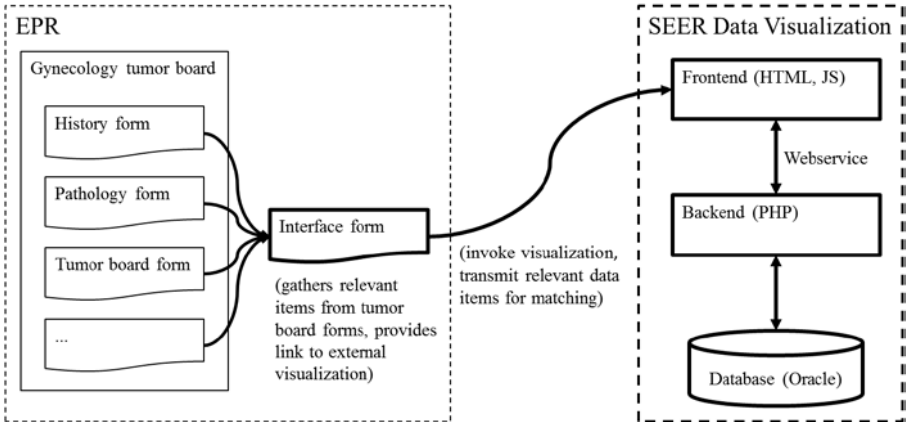

Figure 2. Application structure

including 6 selected questions from the ISONORM 9241/10 usability instrument [13] (using a 5 point Likert scale) as well as 5 open questions addressing specific aspects of the implementation and its possible integration into the clinical workflow.

3. Results

3.1. Data elements and mapping

Comparing both datasets resulted in a common set of attributes that can be used for selecting a patient cohort within SEER matching the selected individual patient, as shown in Table 1.

Even though relevant items were generally available in the SEER dataset, not all of them were documented throughout the whole time period covered by it (1973-2010). E.g. the TNM status was made available only from the year 2004 onwards, making all records before that time unavailable for matching with current EPR data. The remaining usable dataset contained 89.616 records. It would have been desirable to map tumor specific antigens such as human epidermal growth factor receptor 2 (HER2/neu) which may be a guide to use specific chemotherapy agents. Unfortunately, HER2/neu is not available for all cases in SEER and therefore currently not included into the standard SEER dataset today.

Table 1. Matching attributes in both SEER and EPR dataset

Priority	Attribute	Description
1	Age at diagnosis	Available in both datasets
2	TNM stages	Available in both datasets
3	Progesterone receptor status	Available in both datasets
3	Estrogen receptor status	Available in both datasets
4	Grading	Available in both datasets
x	HER2/neu status	Not available in SEER dataset

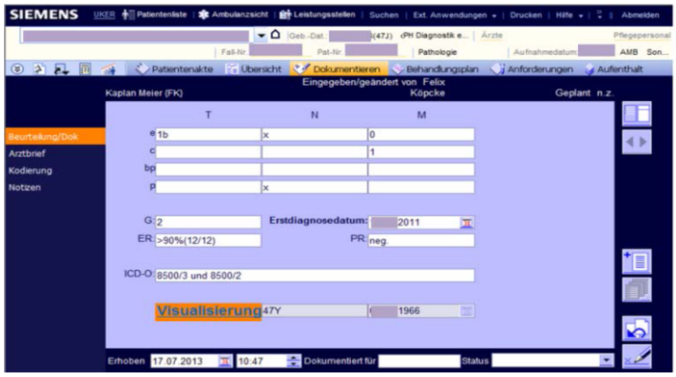

Figure 3. Screenshot of EPR interface form to invoke the visualization

3.2. Implementation

The implementation allows a clinical user to start the workflow by opening the dedicated "interface" form within the EPR (Fig. 3).

All relevant data elements already documented in the routine tumor board forms are used to pre-populate fields in the interface form. The clinical user can change the preset values, e.g. by removing values to achieve a broader selection or manually adding missing values not yet documented in the routine workflow. The visualization is invoked by clicking a hyperlink on the interface form, opening a new browser window.

For the visualization, all records from the SEER breast cancer dataset are selected which match the criteria given in the interface form. Counts are aggregated by age group and by surgical treatment. Pie-Charts are dynamically generated for both variables. Fig. 4 shows a screenshot of the breakdown by surgical treatment.

The size of patient groups as annotated both with percentages and absolute numbers, so the user can take into account the sample size. Small patient groups that cannot by reliably displayed on the pie-chart are automatically grouped into an "Other" category with detailed counts given in the legend.

A Kaplan-Meier survival curve can be generated both for the full cohort as well as dynamically for sub-groups based on surgical treatment group by clicking on the relevant part of the pie-chart (Fig. 5). Confidence intervals can by dynamically included to the chart by the users.

3.3. Evaluation

Five medical doctors took part in the evaluation of the system. Responses were favorable regarding the type of visualization and the re-use of data already available from routine documentation. However, the content was found to be insufficient, as both chemotherapy data as well as HER2/neu status were not available in the published SEER dataset. Both data elements were considered of higher importance by the participants of the evaluation than the surgical therapy data present in the dataset.

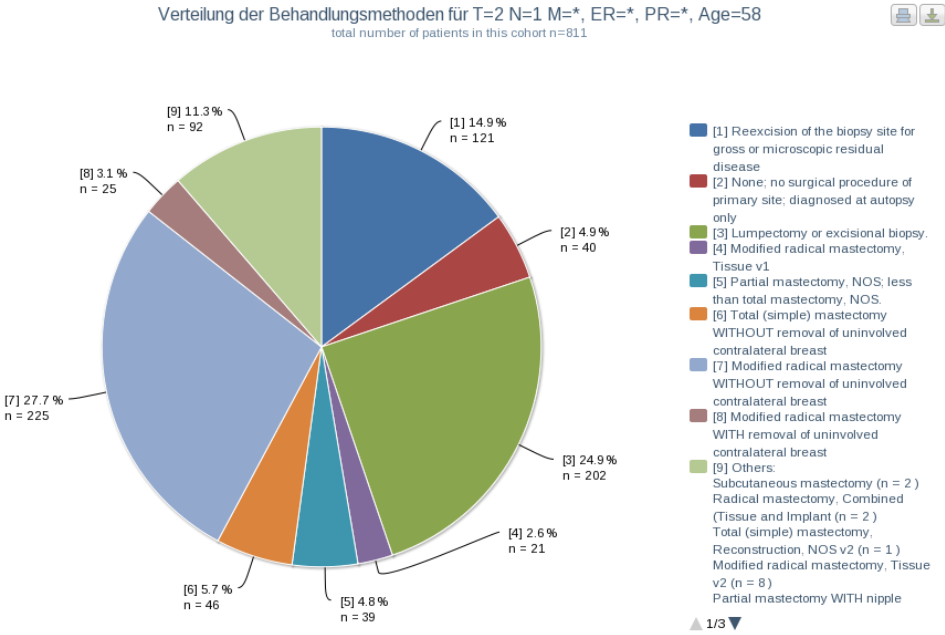

Figure 4. Screenshot of pie-chart visualization of surgical treatment within the selected cohort

4. Discussion

A proof-of-concept integration of an EPR-based online breast cancer tumor board with the SEER cancer registry dataset was successfully implemented, enabling the dynamic visualization of SEER cohorts matched to the current patient context. A successful mapping between local EPR and SEER data elements could also be demonstrated. However, several limitations were identified: Due to TNM data being available in the SEER dataset only from 2004 onwards and limited compatibility between American and International derived cancer staging systems, just a subset of 20.5% of the dataset could be made available for the visualization. When combinations of several selection criteria were applied, the resulting matched groups in the SEER dataset in many cases rapidly dwindled to double- or single-digit sizes of limited value. Therefore further filtering mechanisms to better match population characteristics were not yet applied. Patient groups of special interest (e.g. uncommon cases with advanced tumor stages) were especially affected by this issue. Also, the lack of chemotherapy data or biochemical markers like HER2/neu in the published SEER dataset reduced the value of the system to clinical users. The authors contacted the SEER team about availability of these data items. Although chemotherapy data is now available internally, it is not yet considered for routine publication due to lack of coverage throughout all participating cancer registries. The authors also contacted the local clinical cancer registry of EUH to inquire about using a local dataset instead. Unfortunately, cohort sizes were deemed too small to provide additional value in this context. With the rapid development of novel diagnostic and therapeutic options for cancer, registry data will potentially lag behind in many areas.

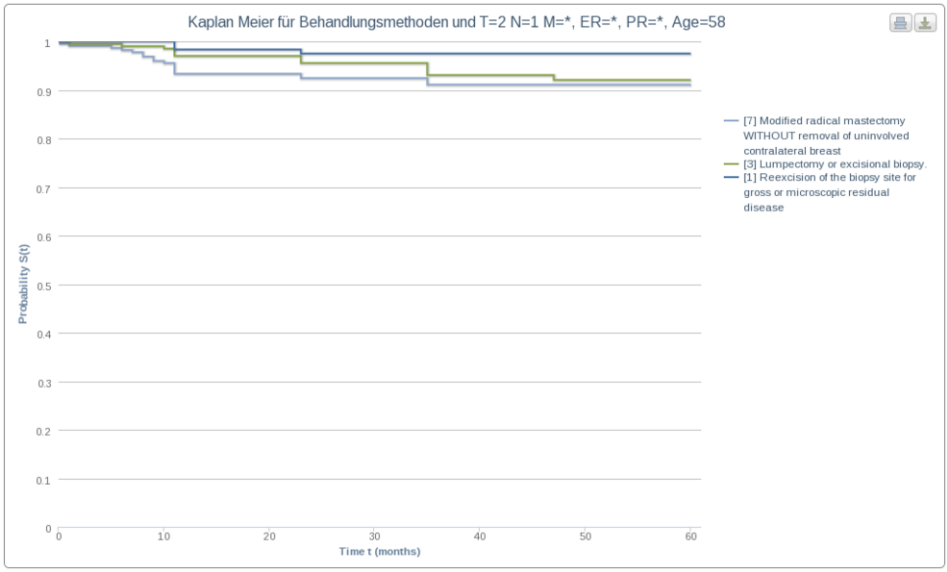

Figure 5. Screenshot of Kaplan-Meier survival curve for three selected types of surgery

However, clinician feedback on the integration of visualizations of this type of epidemiologic data into the routine documentation workflow was positive. Future efforts could consider integrating data from multiple sources to provide a more comprehensive and customizable database. For routine use such a system would need risk assessment and potentially medical product certification. The recent passage of a law establishing and regulating nationwide clinical cancer registries in Germany is a positive development towards this goal and it will be desirable that clinical registers also cover latest biochemical or pharmacogenomic markers at the earliest possible time.

Comprehensive searching in several online libraries showed that there has not yet been a project in the past that allowed a direct linkage of local patient data with a cancer registry database. Projects like Adjuvant! Online [3] or PREDICT [4], are capable of visualizing case-based cohort data which is similar to this project. These tools, however, need to be provided with local patient data by hand which obviously does not integrate in the physician's work-flow flawlessly.

References

[1] Aktuelle Todesursachenstatistik. http://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/Gesundheit/Todesursachen/Aktuell.html (accessed 27.01.2014)

[2] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2010), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2013, based on the November 2012 submission. http://seer.cancer.gov/data/ (accessed 27.01.2014)

[3] Ravdin PM, Siminoff LA, et al. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol. 2001 Feb 15;19(4):980-91.

[4] Wishart GC, Azzato EM, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res. 2010;12(1):R1.

[5] Kattan MW, Eastham JA, et al. A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer. J Natl Cancer Inst. 1998 May 20;90(10):766-71.
[6] Sloan-Kettering Prostate Cancer Prediction Tools. http://www.mskcc.org/cancer-care/adult/prostate/prediction-tools (accessed 27.01.2014)
[7] Moxey A, Robertson J, et al. Computerized clinical decision support for prescribing: provision does not guarantee uptake. J Am Med Inform Assoc. 2010 Jan-Feb;17(1):25-33.
[8] Haux R, Seggewies C, et al. Soarian - workflow management applied for health care. Methods Inf Med. 2003;42(1):25-36.
[9] Prokosch HU, Ries M, et al. IT infrastructure components to support clinical care and translational research projects in a comprehensive cancer center. Stud Health Technol Inform. 2011;169:892-6.
[10] Ries M, Prokosch HU, et al. Single-source tumor documentation - reusing oncology data for different purposes. Onkologie. 2013;36(3):136-41.
[11] JSONSS Framework. http://code.google.com/p/jsonss/ (accessed 27.01.2014)
[12] HighCharts.JS Framework. http://www.highcharts.com/ (accessed 27.01.2014)
[13] Prümper, J. (1997). Der Benutzungsfragebogen ISONORM 9241/10: Ergebnisse zur Reliabilität und Validität. In R. Liskowski, B.M. Velichkovsky & W. Wünschmann (Eds.), Software-Ergonomie '97 - Usability Engineering: Integration von Mensch-Computer-Interaktion und Software-Entwicklung (pp. 253-262). Stuttgart: Teubner.