Mining Careflows of Breast Cancer Patients

Lucia Sacchi, Arianna Dagliati, Annagiulia Pedrazzoli, Alberto Zambelli, and Riccardo Bellazzi

***Abstract*— In 2010, the University of Pavia in collaboration with the IRCCS Fondazione Maugeri hospital developed ONCO-i2b2, an IT platform to integrate clinical and research oncology data based on the Informatics for Integrating Biology and the Bedside framework (i2b2). This system includes the data of 28.838 oncology patients and it gives the possibility to extract heterogeneous clinical, process and research data in a standardized way. In this paper we explore a way to exploit the data collected in the ONCO-i2b2 system for research. In particular, we focus on the development of a methodology to mine frequent patterns of care (or *careflows*) depicting the histories of the oncology patients managed by the hospital. We present the results obtained on the subset of patients with a diagnosis of breast cancer. These careflows highlight the most frequent pathways that breast cancer patients undergo during their follow-up.**

I. Introduction

In 2010, the University of Pavia in collaboration with the IRCCS Fondazione Maugeri hospital (FSM) developed an IT platform to integrate clinical and research data based on the Informatics for Integrating Biology and the Bedside framework (i2b2, https://www.i2b2.org/). This project, named ONCO-i2b2 and funded by the Lombardia region, aims at supporting translational research in oncology. The project exploits the software solutions implemented by the i2b2 research center, an initiative funded by the NIH Roadmap National Centers for Biomedical Computing and headed by Partners HealthCare Center in Boston [1]. The i2b2 project provides a data warehouse (DW) system and a set of software tools based on an architecture (the *hive*) that has different software cells devoted to data extraction, data manipulation and data analysis tasks [2].

Within ONCO-i2b2, the i2b2 infrastructure has been integrated with FSM information system (HIS) and with a cancer biobank that manages both plasma and cancer tissues. The integration with the HIS provides access to all the electronic medical records of cancer patients. As the majority of the data collected in the FSM HIS is represented by textual reports, a Natural Language Processing (NLP) module was developed and integrated inside the system architecture. This allows extracting important information and clinical tests results, such as patients' histological reports [3,4]. The oncology biobank provides bio-specimens prepared from a collection of blood and tissue samples, taken with the informed consent of healthy individuals and oncology patients.

The ONCO-i2b2 system currently collects data about several aspects of a patient's history. In particular, data coming from the HIS allow recording: all the accesses to the hospital and their type, all the performed diagnoses and procedures (coded by ICD9-CM codes), and the drugs (identified through the ATC code system) administered to a patient during such accesses. These are process data, as they reflect the events a patient undergoes during his clinical history, without collecting any quantitative information. To represent information related to process data in a standardized way, the ICD9-CM ontology was integrated in the i2b2 platform relying on the web services provided by the NCBO BioPortal (http://bioportal.bioontology.org) [3]. From the biobank, the information about the histological specimens and their readings are stored and classified thanks to the SNOMED system

The aim of this paper is to propose a way to exploit the ONCO-i2b2 DW for research. Among all the available oncology patients, we selected those diagnosed with breast cancer. During her disease history, each patient will undergo several clinical events (including all the data listed above). These events develop over time, creating a temporal sequence representing the flow of care (*careflow* [5,6]) for each specific subject. Herein we will present a methodology to mine frequent careflows from the data stored in the ONCO-i2b2 DW. Once mined from data, such behaviors will give a picture of the most common pathways the patients undergo during the history of their disease. This could give hints on two aspects: on the one hand, the process of care operated by the structure could be highlighted, and on the other we could gain a better insight on the features of the patients undergoing each predefined pattern.

The paper is organized as follows: in Section II we describe the methodology we have developed to mine frequent careflows from data. In Section III we describe the results obtained for the breast cancer population extracted from the ONCO-i2b2 system.

II. Methods

The methodological approach we defined to extract the most frequent careflows from data is inspired from both Temporal Data Mining and Process Mining techniques. In particular, we have developed a specific method to extract frequent temporal histories derived from the methodologies for sequential patterns mining [7,8], which creates an output

L.S., A.D and A.P. Authors are with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy (corresponding author: L.S. phone: +39 0382 985981; fax: +39 0382 985060; e-mail: lucia.sacchi@unipv.it. Other Authors' email: arianna.dagliati@unipv.it, annagiulia.pedrazzoli01@universitadipavia.it, riccardo.bellazzi@unipv.it).

R.B. Author is with the Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy and with the IRCCS Fondazione Salvatore Maugeri, Pavia, Italy (email: riccardo.bellazzi@unipv.it)

A.Z. Author is with the IRCCS Fondazione Salvatore Maugeri, Pavia, Italy (e-mail: alberto.zambelli@unipv.it).

that can be visualized using the process mining suite ProM [9]

To extract frequent careflows we use the algorithm specified in the following.

The data are arranged in a structure where each row represents a patient and each column is a specific clinical event. We use this structure as the starting point for our mining approach. We first consider the set made up of all the starting events of the clinical histories and, for each of these events, the *support* is computed. The support is an indicator that allows quantifying the frequency of an event (or of a history) in the dataset. In this context, to define the support we take into account two aspects. On the one hand, we compute the proportion of cases that experience a clinical event (or an history) over the total number of cases. On the other, we consider the raw number of cases that experience a specific event (or history). Thresholds defined over these two quantities are used to guide the search to extract only the most frequent patterns among the patients population. The thresholds on support are first used to select frequent starting events.

Then an iterative process starts:

1. for each selected frequent event, consider all the temporal histories starting with that event. At this step, we have a set of temporal histories all starting with the same event;
2. consider the following events of the history. For each of these events, compute the support of each history made up of the starting and the second event. Discard the first event of the history and consider the second events as the starting points;
3. Return to step 1 until no more frequent histories can be extracted.

The clinical histories extracted using this algorithm are formatted in a way that is suitable for visualization with the well-known process mining suite ProM.

A very interesting application of this careflow mining technique is the one related to comparison of careflows in different patients groups. This may for example enable to highlight differences in the management of different risk profiles, which wouldn't be possible when considering the overall patients population.

III. Results

ONCO-i2b2 currently includes data for 28.838 oncology patients. Querying the system for patients with a diagnosis of malignant breast neoplasm (ICD9-CM code 174.0-174.9) or personal history of malignant neoplasm (ICD9-CM V10.3), a total of 8.969 subjects was obtained.

To analyse the flows of care, together with our medical expert we decided to focus on the clinical events related to hospital admissions. The reason underlying this choice is that, according to medical knowledge, breast cancer patients usually experience a pretty well-defined set of processes of care, depending of course on their overall clinical condition. This offers a good testing environment for our proposed approach. The clinical events we have considered are the ward of admission and the admission regimen for each event. The admission regimen can be an in-hospital stay, typically lasting more than one day, or a day hospital, lasting one day.

Interesting careflows are intuitively those made up of at least two clinical events. Analyzing our dataset to discard those histories made up of just one event led us though to find out some interesting information. Out of the 2.282 patients who underwent a single hospital admission, the 88% were staying in the breast surgery division. At FSM, this division is certified by the European Society of Breast Cancer Specialists (www.eusoma.org) and is a European reference center. This leads to a considerable number of patients (the 26% of our dataset) accessing the ward for the intervention and then being followed in other hospitals, especially out of the region.

After discarding patients with only one event in their history and patients with complete clinical histories ending before 2000, when the organization of the wards was different from the current one, we were left with a set of 4.533 subjects, who were analyzed with the methodology proposed in section II.

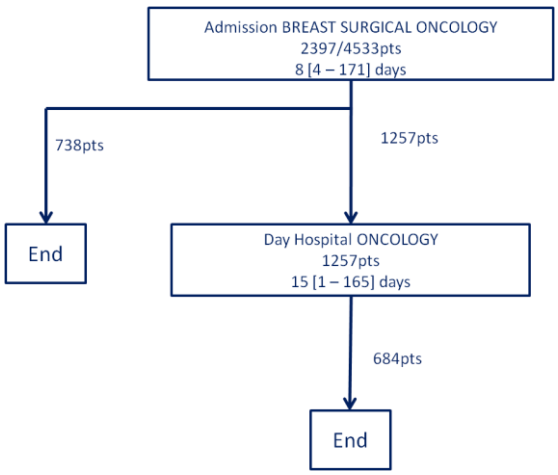

Figure 1. The most frequent careflow mined from the dataset

Figure 1 shows the most frequent careflow extracted from our dataset. Each clinical event of the history is represented in a rectangle, while arrows are transitions between events. The name of the event is specified in the rectangle, together with the number of patients undergoing that event. The starting event indicates also the total number of patients involved in the analysis. Numbers on the arrows represent the patients who go through that flow. Patients who exit the flow are indicated on the arrows going towards the "End" blocks. Besides the information on the number of patients, also the information about the duration of the events is presented. In particular, the median length of stay and its interquartile range are shown.

The figure in this case represents a flow of care starting with an admission to the breast surgical oncology ward. This is the starting event for the 53% of the patients in our dataset. Out of 2397 patients, 738 exit the flow, this meaning

they are no more followed by the FSM hospital. As patients undergoing a single hospitalization had already been removed before running the algorithm, these 738 subjects are those who undergo a set of following surgical interventions in the ward and then are treated elsewhere. Performing multiple surgical interventions is a frequent procedure in surgical oncology, where separate surgeries are performed for biopsy, cancer removal and breast reconstruction.

The patients who stay in the flow undergo one or more day hospitals in the oncology ward. The fact that this event can be single or repeated can be inferred by the information on the duration. Patients who undergo only one day hospital are those for whom the event lasts 1 day (first quartile, in brackets). Of the 1257 patients undergoing the day hospitals, 684 then end their history. As we will see in the following, the remaining part of the patients is involved in different flows.

From a clinical perspective, the flow we have just described represents the standard process of diagnosis and care for a breast cancer patient without complications in the Italian healthcare system. In particular, these patients are diagnosed either following the screening programs made available by the Italian national healthcare service (ultrasounds and mammography) or thanks to a self-evaluation that triggers further investigations. The following step is then a visit to the primary care practitioner, who directs the patients to the surgery, thus making the process represented in Figure 1 start. The set of day hospitals following breast surgery represents the therapeutic flow. In fact, chemotherapy drugs are usually administered in a day hospital regimen, where patients are followed for one day at the ward to control the therapy effects, and then discharged.

Another interesting process is the one reported in Figure 2. The grey part of the careflow is common to the one of Figure 1. Patients belonging to the careflow in Figure 2, though, continue their history either with a hospitalization in the oncology ward or with an additional surgical intervention at the breast surgery ward. These patients are the ones who start their history with the usual diagnostic process but then have some complications. Additional hospitalizations can be due for example to therapy toxicity events or to unstable conditions that lead to a progression of the disease.

Besides the data on hospitalizations, ONCOi2b2 offers the possibility of using also other clinical information to deepen the analysis details. We exploited this unique feature of the system to understand whether there exist some differences in terms of clinical variables among patients who experience the history in Figure 1 and the ones in Figure 2. To this end, we have considered data related to blood test results, which are available in ONCOi2b2 thanks to the information retrieved from FSM HIS. The DW stores the information on all the blood tests that the patients undergo during their history. Together with the clinical expert, we have selected a subset of them, particularly interesting from an oncology perspective. These are shown in Table 1. Blood test results come as raw values in the DW. For properly processing them with our mining algorithm, we discretized them using the thresholds shown in the third column of the table. Such thresholds were defined by the clinical expert as well.

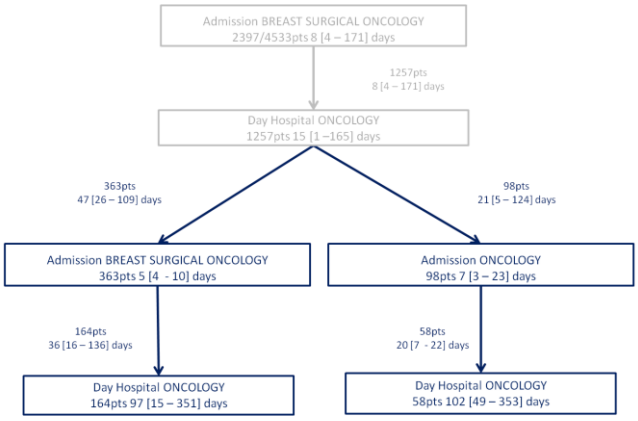

Figure 2. The careflow depicting the history of patients with complications.

To exploit the information on blood tests we analyzed each transition of the extracted careflows and considered the results for the patients who actually verify that path. Interestingly, as transitions duration might be different for different patients, what it usually happens is that one patient might have one or more recording for each blood test in the period. To deal with this issue, in this analysis we applied a specific data mining technique, Temporal Abstraction (TA) [10,11,12], which allows extracting intervals of validity of each pattern (for example intervals of LOW white blood cells count).

We found some differences in the values of the two tumor markers CEA and CA15-3 in patients with and without complications. In more detail, the majority of the patients who undergo one or more hospitalizations in the oncology ward had both positive CEA and CA15-3, while the patients who are not re-hospitalized have normal values for both the markers. This confirms that patients who require to change the careflow shifting from routinely day hospitals to hospitalizations in the oncology ward are those for whom the disease progresses.

TABLE I. THE ANALYZED BLOOD TESTS

DESCRIPTION	UNIT	THERSHOLD
White Blood Cells Count	$x10^3/mm^3$	LOW if < 4
Hemoglobin	g/dl	LOW if < 8
Platelet Count	$x10^3/mm^3$	LOW if < 200
Carcino Embryonic Antigen (CEA)	ng/ml	POSITIVE if > 5
Cancer Antigen15-3 (CA15-3)	u/ml	POSITIVE if > 35

The last interesting careflow we will analyze in this paper is shown in Figure 3. In this case, the starting event of the flow is a day hospital in the oncology ward. The patients who enter this history are 751/4533 (16% of the patients). The careflow then develops through three branches. The first refers to those patients (344, or 45,8%) who exit the pathway. Another branch joins to the typical flow of non-

complicated patients of Figure 1 (140 patients, or 18,6%). The last one is instead characterized by an alternation of day hospitals and hospitalizations (152 patients, or 20,2%).

From a clinical perspective, this last careflow is the one followed from those patients who don't go through the traditional diagnosis course, but show a more complex diagnostic itinerary. These can be for example those patients for whom cancer is diagnosed at a more advanced stage, who require a therapeutic treatment to be done prior to surgery or to any other clinical intervention. Patients who alternate day hospitals and inpatient stays are the most severe patients, who are very unstable and require frequent re-hospitalizations within treatments.

To confirm the observations above regarding the starting event of the careflow (a day hospital in the oncology ward), we have taken into account the information about the drugs that have been prescribed to this group of patients during the period they underwent the initial day hospitals. This further in-depth analysis was possible thanks to the information stored in the DW, which includes also the one related to drug prescriptions. Interestingly, the 66% of these patients underwent injection or infusion of chemotherapy substances during these events. This confirms that a therapy is performed on these subjects prior to start any other intervention.

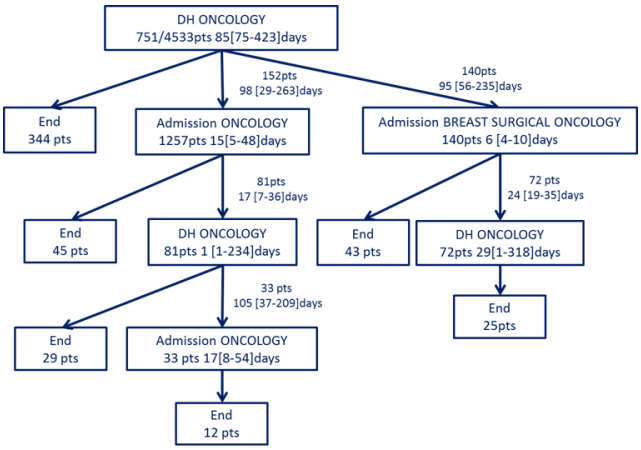

Figure 3. The careflow depicting the history of patients with complications.

IV. Conclusion

In this paper we have presented a methodology to mine frequent careflows from clinical data stored in the ONCO-i2b2 system. The methodology is able to take into account events of different nature and the variety of data collected in the system allows further detailing the extracted careflows with additional information. In the future, we will investigate further aspects, focusing in particular on the stratification of patients on the basis of the information on the specimens. This will allow understanding whether patients showing different tumor types undergo different care flows. Related to this point, we will explore similarity metrics that allow formally comparing histories mined on different groups of patients. Finally, we will focus on better enhancing the histories to fully exploit the rich information included in the DW.

References

[1] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S. et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2) Journal of the American Medical Informatics Association. 2010;17:124–130. doi: 10.1136/jamia.2009.000893

[2] https://www.i2b2.org/software/projects/datarepo/CRC_Design_Doc_13.pdf

[3] Segagni D, Tibollo V, Dagliati A, Zambelli A, Priori SG, Bellazzi R. An ICT infrastructure to integrate clinical and molecular data in oncology research. BMC Bioinformatics. 2012 Mar 28;13 Suppl 4:S5. doi: 10.1186/1471-2105-13-S4-S5. PubMed PMID: 22536972; PubMed Central PMCID: PMC3303735.

[4] Segagni D, Tibollo V, Dagliati A, Malovini A, Zambelli A, Napolitano C, Priori SG, Bellazzi R. Clinical and research data integration: the i2b2-FSM experience. AMIA Summits Transl Sci Proc. 2013 Mar 18;2013:239-40. PubMed PMID: 24303274; PubMed Central

[5] Quaglini S, Stefanelli M, Cavallini A, Micieli G, Fassino C, Mossa C. Guideline-based careflow systems. Artif Intell Med. 2000 Aug;20(1):5-22. PubMed PMID: 11185420.

[6] Quaglini S, Stefanelli M, Lanzola G, Caporusso V, Panzarasa S. Flexible guideline-based patient careflow systems. Artif Intell Med. 2001 Apr;22(1):65-80. PubMed PMID: 11259884.

[7] Agrawal R., Srikant R. (1995) Mining Sequential Patterns. In: Yu PS, Chen ALP (eds) Proceedings of the 11th International Conference on Data Engineering. IEEE Computer Society, pp. 3-14

[8] Zaki M.J. (2001) SPADE: An Efficient Algorithm for Mining Frequent Sequences. Machine Learning 42(1-2) 31-60

[9] H.M.W. Verbeek, B.F. van Dongen, J. Mendling, W.M.P. van der Aalst. Interoperability in the ProM Framework. In: CAiSE 2006 Workshop Proceedings - Open INTEROP Workshop on Enterprise Modelling and Ontologies for Interoperability (EMOI-INTEROP 2006), 619-630, June, 2006

[10] Y. Shahar, A framework for knowledge-based temporal abstraction, Arificial Intelligence 90 (1997) 79-133

[11] Batal I., Sacchi L., Bellazzi R., Hauskrecht M. (2009) Multivariate Time Series Classification with Temporal Abstractions. Int J Artif Intell Tools 22 344-349.

[12] Sacchi L., Larizza C., Combi C., Bellazzi R. (2007) Data mining with Temporal Abstractions: learning rules from time series. Data Mining and Knowledge Discovery 15(2) 217-247