ATLANTIS PRESS
International Journal of Computational Intelligence Systems, Vol. 12 (2018) 28-38

Case-Based Decision Support System for Breast Cancer Management

Booma Devi Sekar[1], Jean-Baptiste Lamy[2], Nekane Larburu[3, 4], Brigitte Séroussi[2, 5], Gilles Guézennec[2], Jacques Bouaud[2, 6], Naiara Muro [2, 3, 4], Hui Wang[1], Jun Liu[1]

[1]School of Computing, Ulster University, Northern Ireland, UK
E-mail: b.sekar@ulster.ac.uk, h.wang@ulster.ac.uk, j.liu@ulster.ac.uk

[2] Sorbonne Universités, UPMC, Univ Paris 06, INSERM, Université Paris 13, Sorbonne Paris Cité, UMR S 1142, LIMICS Paris, France
[3]eHealth and Biomedical Applications, Vicomtech-IK4, Donostia-San Sebastian, Spain
[4]Biodonostia, Donostia-San Sebastián, Spain
[5]AP-HP, Hôpital Tenon, Départemente de Santé Publique, Paris, France
[6]AP-HP, DRCI, Paris, France
E-mail: jean-baptiste.lamy@univ-paris13.fr, nlarburu@vicomtech.org, brigitte.seroussi@aphp.fr, gilles.guezennec@univ-paris13.fr, jacques.bouaud@aphp.fr, nmuro@vicomtech.org

Received 21 October, 2017

Accepted 12 August, 2018

Abstract

Breast cancer is identified as the most common type of cancer in women worldwide with 1.6 million women around the world diagnosed every year. This prompts many active areas of research in identifying better ways to prevent, detect, and treat breast cancer. DESIREE is a European Union funded project, which aims at developing a web-based software ecosystem for the multidisciplinary management of primary breast cancer. The development of an intelligent clinical decision support system offering various modalities of decision support is one of the key objectives of the project. This paper explores case-based reasoning as a problem solving paradigm and discusses the use of an explicit domain knowledge ontology in the development of a knowledge-intensive case-based decision support system for breast cancer management.

Keywords: knowledge intensive case-based reasoning, breast cancer, ontology, case-based decision support system, jColibri.

1. Introduction

Breast cancer is the most common type of cancer in women worldwide, with the mortality rate being second highest (next to lung cancer) among different types of cancer.[1] As reported by World Health Organization (WHO, 2013),[2] 508,000 women died due to breast cancer in 2011. Perceived as a disease of the developed world, incidence rates vary worldwide, from 19.3, 40, to 89.7 per 100,000 women in Eastern Africa, most of the developing countries, and Western Europe respectively.[2] Whereas, the survival rates vary from 40%, 60%, to 80% in low-income, middle-income, and developed countries respectively, which reflects the lack of early detection programs and adequate diagnosis and treatment facilities in the less-developed countries.

In light of this, both industry and academia are taking mitigating action in providing clinical support solutions and patient-centric healthcare systems for breast cancer diagnosis and management. For instance, the worldwide cancer research organization[3] is funding 30 research projects across the world for breast cancer, among which 23 projects are researched and developed in Western European countries. In UK, various breast cancer related research projects are being funded by Cancer Research UK,[4] National Breast Cancer

Foundation,[5] Breast Cancer UK[6] etc. DESIREE is a European Union funded project,[a] which aims at developing a web-based software ecosystem for the personalized, collaborative, and multidisciplinary management of primary breast cancer (PBC) by multidisciplinary Breast Units (BUs). With the ultimate goal of the system to be deployed in actual BUs in the future, one of the main objectives of the project is to develop an intelligent clinical decision support system (DSS). The idea is to go beyond the limitations of clinical practice guidelines for breast cancer care,[7] and incorporate experience-based and case-based decision support modalities that are based on the evolving knowledge acquired from previous patient cases. In this paper, we present the knowledge intensive case-based reasoning (KI-CBR) model, which incorporates knowledge from explicit domain knowledge we developed and serves as the case-based DSS (CB-DSS) within the DESIREE project. The proposed model conceptualizes and integrates the breast cancer domain knowledge in an ontological format and the case retrieval algorithm using semantic similarity measure.

Case-based Reasoning (CBR) has been a field of great interest for researchers, entrepreneurs, as well as clinicians for over three decades.[8,9,10] It integrates several disciplines of heterogeneous nature such as Cognitive Science, Artificial Intelligence (AI), and Information Science to provide a computational model that is very close to human reasoning. The main advantage is that with CBR, the usually extensive and complex formalization of the problems to be solved is not required. Unlike various other AI approaches such as rule-based reasoning, or neural networks, that generate abstract representations from a set of training examples, CBR methodology adapts instance-based learning and uses previous similar cases as the basis for decision making. In medical domain in particular, the clinicians not only use the rules from clinical practice guidelines, but also use their knowledge and experience to diagnose the disease and provide treatment. In a CBR model, the experience can be incorporated as cases in the case-base, which could include not only typical simple scenarios, but also the complex and exceptional ones. Thus, the model considers them when reasoning and automatically enriches the case-base with parts of changeable knowledge as new case and enables it to perform better with usage. Thus, as the clinical decisions made by physician improves with experience, the performance of CBR model would also improve with usage, which is not possible in a rule-based model.[11] Thus, the fact that the CBR methodology closely resembles the thought process of the clinicians, and the gradual acceptance of advanced decision support systems in clinical practice, suggest the success of CBR in medicine.[12] Some of the CBR systems developed in medicine and health science domains so far, include CASEY[13] to diagnose heart failure patients, MNAOMIA[14] to diagnose and treat eating disorders, PROTOS[15] for hearing disorder diagnosis, MacRad[16] for radiology image classification, GerAmi[17] for Alzheimer's disease management, and GOCBR[18] for breast cancer diagnosis. Many such CBR applications are well summarized by Choudhury *et al*.[11]

When developing a CBR system, the main challenge is to build the case base and implement an effective case retrieval algorithm. In a medical application, acquiring and managing a complete and consistent case base, covering a large number of resolved patient cases with varying diagnosis output, becomes the basis for building a reliable CB-DSS. Then, CBR would adapt a supervised learning algorithm, trained on resolved cases, stored in the case base. Thus, first identifying the most similar cases from the case base, and then adapting the solution of the retrieved cases to build the solution for the new query case, solves a new case. Taking into account of various algorithmic strategies[19] such as root mean square distance, geometrical matching, consensus shapes, weighted Euclidean distance etc., in developing similarity analysis and case retrieval algorithm, which would accurately measure the clinical distance between two patient cases, becomes a part of the problem-solving paradigm. In the present work, we present an ontology-based schema to conceptualize, interpret, and integrate large clinical data of breast cancer patients into a case-base knowledge. Subsequently, benefiting from the knowledge codified in the ontology, we incorporate semantic similarity measure to implement the case retrieval algorithm.

Implementation of CBR algorithm is feasible on various platforms. However, building the CBR framework to execute the complete CBR cycle is a challenge, which

[a] http://www.desiree-project.eu

requires substantial research effort and time. In light of which, with the effort to formalize CBR and provide design and implementation assistance, various CBR frameworks have been distributed as software tools. Some of the popular non-commercial tools include myCBR, jColibri, CAT-CBR, CASPAIN, CBR Shell, FreeCBR, and eXiTCBR. Based on a detailed comparative study performed on these different software tools,[19,21,22] myCBR and jColibri were initially selected, and jColibri2 (Java framework)[23] was finally adopted in this project to design a CB-DSS for breast cancer management.

In the following section, we present the overall framework of the proposed CB-DSS for DESIREE, describe the related functional blocks, and explain the details of the proposed CB-DSS, which include the clinical data integration using case-base knowledge model, the similarity analysis, and the case retrieval algorithm using semantic similarity measure and the patient case representation. Section 3 presents a case study to show the complete operation of the proposed CB-DSS. Finally, section 4 draws the conclusion of the paper.

2. Case-Based Decision Support System for DESIREE

2.1. Framework and Workflow

One of the main objectives of the DESIREE project is to provide decision support for the therapeutic decisions made by BUs, including surgery, radiotherapy, and systemic therapies. With the aim to provide a personalized state-of-the-art clinical decision support system to BUs, the project provides different modalities of decision support, including guideline (GL),[24] experience (EX)[25] and CB-DSS. In this paper, we present the proposed CB-DSS.

The DESIREE platform contains the DESIREE Information Management System (DESIMS), Fast Healthcare Interoperability Resources (HAPI-FHIR)[26] server and the three DSSs, including GL-DSS, EX-DSS and CB-DSS. The DESIMS contains a clinical interface through which the clinical data of patients are entered and stored according to a common data model. The clinical interface also allows clinical partners to access the heterogeneous patient data and retrieve the results and therapeutic propositions from the different DSSs. The clinical data of patient are transferred from DESIMS to the DSSs for analysis through the HAPI-FHIR server, which is an open source standard framework developed for exchanging healthcare data in a systematic manner. Fig. 1 mainly shows the framework and workflow of the CB-DSS in the DESIREE platform.

In order to incorporate decisional criteria beyond the limitations of current guidelines for breast cancer management, the CB-DSS incorporates the experience of clinicians on previous cases, by collecting the description of patients, and the decision made by clinicians, as the case representation. Subsequently, it provides a tool for querying former cases in order to retrieve similar patient cases based on the defined case base and case retrieval algorithm using semantic similarity measure.

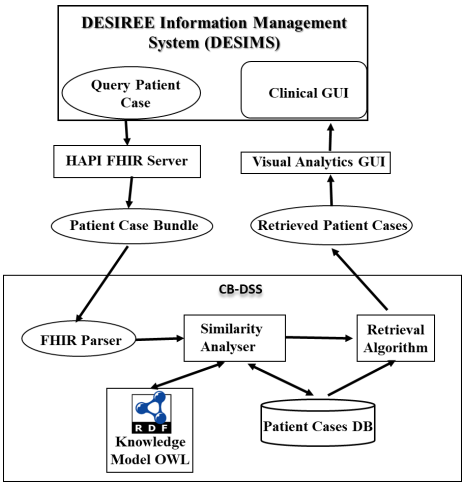

Fig. 1. Framework and workflow of CB-DSS

In the proposed CB-DSS shown in Fig. 1, during runtime, the CB-DSS queries the FHIR server, linked to the DESIMS, for the query patient case. Here, the query patient case represents a new patient case with a number of clinical parameters recorded by BU clinicians using the DESIMS interface. CB-DSS first parses the patient case bundle retrieved from the FHIR server as input attributes to CBR, which includes the similarity analyser and retrieval algorithm. The similarity analyser compares the query case with the patient cases stored in

the database using various similarity functions, whereas, the retrieval algorithm retrieves the most similar patient cases and sends the results to the visual analytics graphical user interface (GUI). The retrieved patient cases are graphically visualized in the DESIMS by comparing the clinical parameters and the clinical decision made on retrieved cases with the information provided for the query case using the rainbow boxes visualization technique developed by Lamy *et al*.[27]

The following sections present a detailed description of the proposed CB-DSS.

2.2. Clinical Data Integration using Knowledge Model Representation

The crucial step in building any clinical DSS allies in first making sense of the large clinical information available, in terms of conceptualizing, preparing, and integrating data. In breast cancer research, relevant clinical data/information could come from various resources, including patient health records, clinical practice guidelines, medical & scientific documents, etc., which are usually in an unstructured and heterogeneous data format. Literature shows that both academia and industry have explored various approaches to data integration.[28,29,30] Among which, we explore semantic technologies and use ontology, where the medical terms are organized based on concepts and the relationship between them. In particular, for CB-DSS, we exploit the use of ontology-based integration to avail the advantage of the semantic similarity measure.

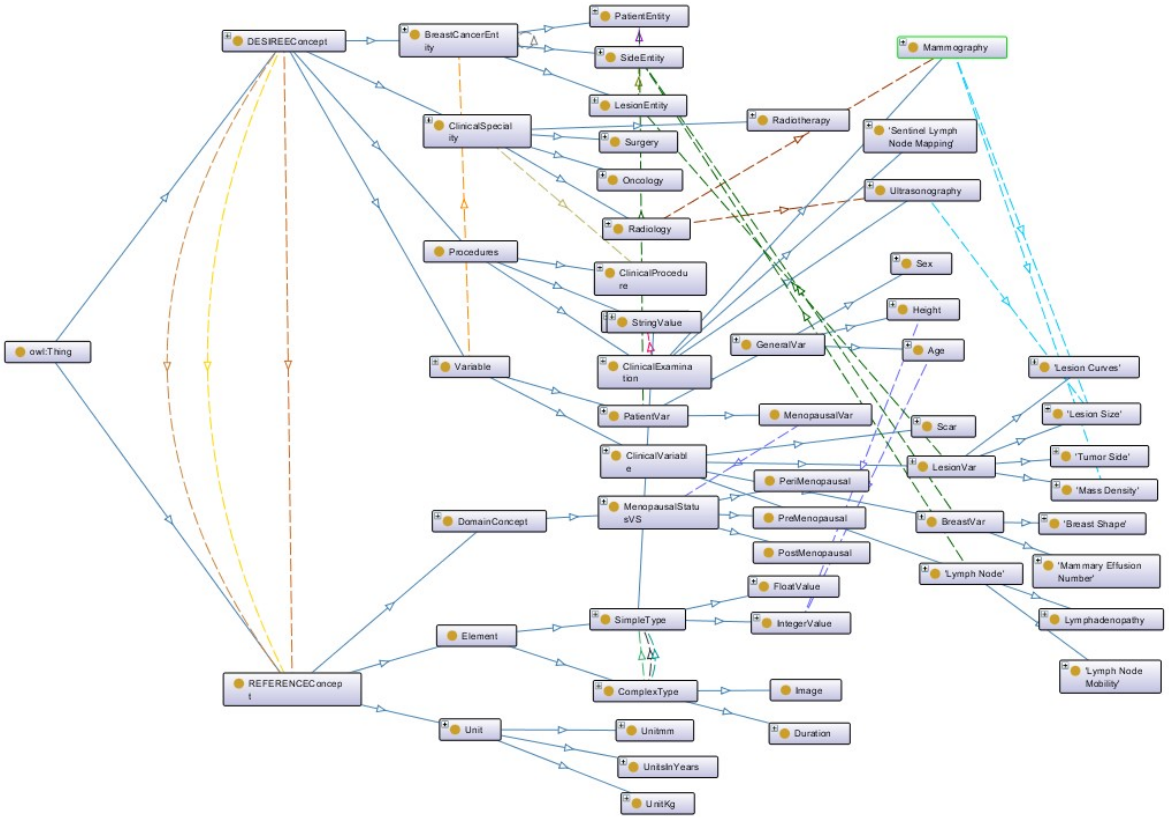

Fig. 2 Semantic representation of breast cancer knowledge model

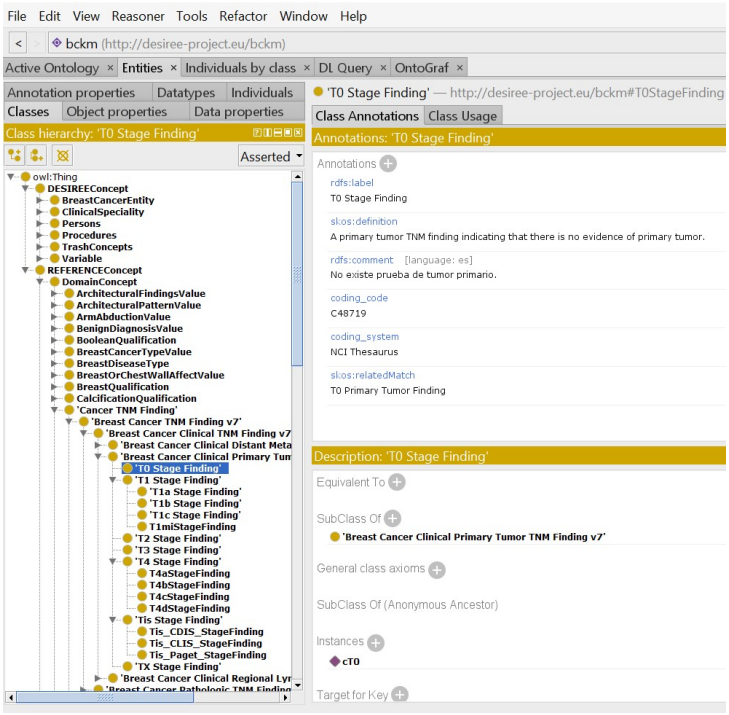

Fig. 3. Example of a concept in the breast cancer knowledge model

From the information provided by clinical partners, clinical practice guidelines, clinical documentation etc., an ontology-based breast cancer knowledge model (BCKM) was defined using the Web Ontology Language (OWL) and Protégé software tool. Although it would be ideal to use established authoritative ontologies such as the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT),[b] which has a greater depth of terminologies for all clinical specialties, the knowledge model was defined to exclusively embed concepts relevant to the breast cancer domain. This specifically avoided the risk of misinterpreting medical terminologies in different care settings and facilitated the analysis of different semantic similarity measures for CB-DSS. However, to establish semantic interoperability beyond the DESIREE project, attributes in the knowledge model were linked with some of the established authoritative terminologies in NCI thesaurus[c] and SNOMED CT.

Fig. 2 illustrates the semantic representation of the BCKM. It shows the top-level concepts, some of the relevant classes and the connections between these concepts using object properties. Fig. 3 shows a screenshot of the BCKM developed using Protégé, which highlights an attribute under the concept "Reference Concept" and its link with the standard terminology in the NCI thesaurus. It also shows the taxonomic hierarchy of the "Cancer TNM Staging". This architecture, allows one to examine the path length, depth and local density factors associated with the ontological taxonomic hierarchy to compute the similarity measure.

As an overall architecture, the BCKM contains two main hierarchies, namely the DESIREE concept and REFERENCE concept. The DESIREE concept contains the entire specifications relevant to the DESIREE environment, namely the concepts relevant to the BUs. The REFERENCE concept contains the potential values for the terminologies under DESIREE concept. The DESIREE concept contains all the clinical data including the relevant clinical procedures, possible examinations, clinical findings, observations, etc. It also characterises the series of parameters and attributes under DESIREE concept into three main entities, namely Patient, Side, and Lesion entities, and allows the

[b] https://www.snomed.org/snomed-ct/
[c] https://ncit.nci.nih.gov/ncitbrowser/

description of breast cancer patient cases. For example, this allows one to identify in a patient case whether the tumour is located on the left or right side of the breast, if a particular attribute is relevant to the patient (e.g. age, gender) or to the lesion (e.g. TNM staging, tumour size).

2.3. Similarity Analyser and Retrieval Algorithm

A high-level description of a generic CBR system is composed of four consecutive processes, namely retrieval, reuse, revise, and retain.[8] With a collection of precedent cases and domain knowledge, one may implement one or more of these four steps in the application.

- Retrieval executes the retrieval algorithm with various similarity matrices to retrieve the case(s) that are most similar to the query case from the precedent ones present in the case base.
- Reuse or adaptation uses the knowledge and information acquired from the retrieved case(s) to solve the query case.
- Revise, through external means, revises the proposed solution.
- Retain learns from the problem-solving experience and stores in the case base the new information/knowledge acquired from the resolution of the query case for solving future new problems.

Among the four steps, as retrieval serves as the basis for the subsequent steps, implantation of an accurate case retrieval algorithm becomes crucial. In the present work, with jColibri software tool, the retrieval algorithm is defined using "local" and "global" similarity functions. Local similarity functions measure the distance between the simple attributes, whereas the global similarity function applies the results from local similarity measures to compare the compound attributes. In the proposed CB-DSS, a patient case is represented as compound attribute, composed of several simple attributes, including physiological and clinical attributes, such as age, gender, BIRADS, histological type, HER2 receptor, tumour size, etc. Thus, local similarity functions are first applied to compute the distance between simple attributes in the query case against the ones characterizing patient cases in the case base. The result of local similarity measures of all simple attributes are then aggregated using the global similarity function to select the patient case(s) that are most similar to the query case from the precedent ones present in the patient case base.

In the case retrieval algorithm, in addition to the simple similarity functions like numeric interval, equal, enumerated distance applicable to compare the simple attributes of the datatype, integer, string and enumerated types, the semantic similarity function is defined to avail the benefit of hierarchical placement of concepts in the domain ontology. Semantic similarity function computes the distance between the simple attributes of a query patient with those of patients retrieved from the patient case base as a degree of taxonomical proximity. Based on the taxonomical structure and location of concepts in the ontology, jColibri presents four semantic similarity functions,[31] which are examined in the proposed CB-DSS. As shown in Eqs. 1 and 2, fdeep_basic and fdeep similarity functions take into account of the depth of the concept in the taxonomical structure of the ontology.

$$f_{deep\_basic}(i_1 - i_2) = \frac{\max\left(prof\left(LCS(i_1,i_2)\right)\right)}{\max\limits_{C_i \in CN}\left(prof\left(C_i\right)\right)} \quad (1)$$

$$f_{deep}(i_1 - i_2) = \frac{\max\left(prof\left(LCS(i_1,i_2)\right)\right)}{\max\left(prof\left(i_1\right), prof\left(i_2\right)\right)} \quad (2)$$

Where, $LCS(i_1,i_2)$ is the set of the least common subsumer concepts of the two individuals, $prof(C_i)$ is the depth of the concept C_i, and $prof(i)$ is the depth of the individual and CN is the set of all the concepts in the knowledge model.

$$\cos ine(i_1,i_2) = sim\left(t\left(i_1\right), t\left(i_2\right)\right) = \frac{\left|\left(\bigcup_{d_i \in t(i_1)}\left(\text{super}\left(d_i, CN\right)\right)\right) \cap \left(\bigcup_{d_i \in t(i_2)}\left(\text{super}\left(d_i, CN\right)\right)\right)\right|}{\sqrt{\left|\left(\bigcup_{d_i \in t(i_1)}\left(\text{super}\left(d_i, CN\right)\right)\right)\right|} \cdot \sqrt{\left|\left(\bigcup_{d_i \in t(i_2)}\left(\text{super}\left(d_i, CN\right)\right)\right)\right|}} \quad (3)$$

Likewise, Eqs. (3) and (4), cosine and detail similarity functions, measure the similarity between two vectors or sets, thus takes into account of the number of superclasses or ancestors in the ontology.

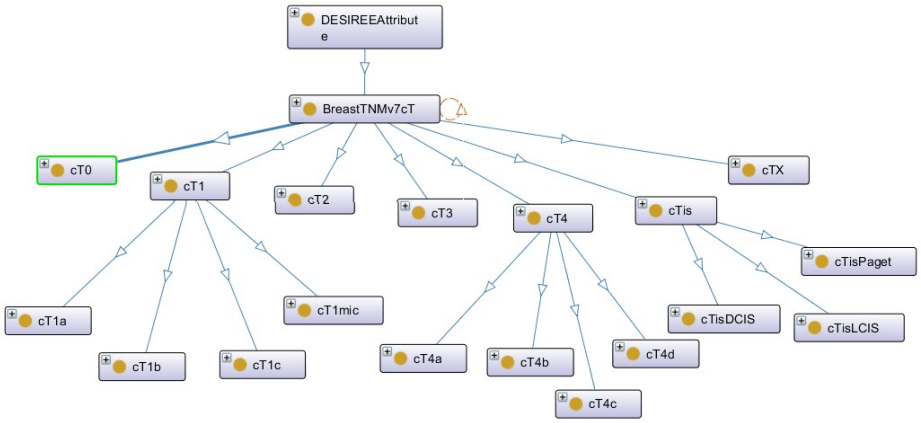

Fig. 4. Hierarchy tree of the DESIREE concept – BreastTNMv7cT

$$\text{detail}(i_1, i_2) = \text{detail}(\text{t}(i_1), \text{t}(i_2)) = 1 - \frac{1}{2 \cdot \left| \left(\bigcup_{d_i \in t(i_1)} (\text{super}(d_i, CN)) \right) \cap \left(\bigcup_{d_i \in t(i_2)} (\text{super}(d_i, CN)) \right) \right|} \quad (4)$$

Where $super(c, C)$ is the subset of concepts in C which are super concepts of c and $t(i)$ is the set of concepts the individual i is an instance of.

Now, with a hierarchy tree of a DESIREE concept: "BreastTNMv7CT" from BCKM presented in Fig. 4, we describe how some of the key terms used in the above equations are computed.

In Eqs. 1 and 2, $LCS(i_1, i_2)$ measures the similarity distance based on the most specific taxonomical ancestor common to i_1 and i_2. The more common the subsumer is, the more similar the terms will be, which ranges from '1' for identical concepts to '0'. In our specific example, assume i_1 is "cT1a" from the query case and i_2 could correspond to any value ranging between "cT0" to "cTx" in the case base. For example, the most specific taxonomical ancestor common to "cT1a" and "cT1b" will be "cT1", whereas for "cT1a" and "cT2" will be "BreastTNMv7cT". Thus, the *LCS*(cT1a, cT1b) will be greater than *LCS*(cT1a, cT2).

In Eqs. 3 and 4, *super*(*c, C*) measures the similarity distance by taking into the account of the amount of shared superconcepts or ancestors of the pair under comparison. Higher the amount of shared superconcepts, shorter the similarity distance between the two concepts. Thus, according to Fig. 4, *super*(cT1a, cT1b) will take into account of the three supperconcepts "cT1", "BreastTNMv7cT" and "DESIREEAttribute", whereas *super*(cT1, cT2) will only take two superconcepts "BreastTNMv7cT" and "DESIREEAttribute" to compute the similarity distance. Therefore, *super*(cT1a, cT1b) will have a lower similarity distance than *super*(cT1, cT1).

Now, with local similarity measures computed for all the simple attributes using Eqs. 1 to 4, *k*-Nearest Neighbour (*k*-NN) is computed as the global similarity function to retrieve the top *k* similar cases from the patient case base. *k*-NN being a non-parametric lazy learning algorithm, it does not make any assumptions on the data distribution or require training data points to do any generalization. Given the query patient case x_q and the local similarity measure computed for the patient cases in the case base y_i, Euclidian distance between the query and case base is computed using Eq. 5.

$$d(x, y) = \sum_{i=1}^{N} \sqrt{x_i^2 - y_i^2} \quad (5)$$

Based on above computation, the *k* nearest patient cases are first located in the case base. The k-NN similarity measure is then computed using Eq. 6, which computes the arithmetic mean output across patient cases in the case base and returns a value between 0 ~ 1, with 0 and 1 indicating the retrieved case being less and most similar to the query case.

$$f(x_q) = \frac{\sum_{i=1}^{k} f(x_i)}{k} \quad (6)$$

where, $f(x_q)$ returns the output value for the query patient case q.

2.4. Patient Case Representation

In a CB-DSS, the most important source of data is the set of resolved cases. Different applications may have different case representational requirements and as the size of the case base increases, it becomes critical that the CBR system accesses the stored cases efficiently. To address such challenges, jColibri provides persistence mechanism and in-memory organization for case base management.[31]

The persistence mechanism provides connectors that allow the CBR system to access the cases from the medium and return them in a systematic manner. The three different connectors include the case base connector, plain text connector, and ontology connector to manage the persistence of cases in the case base, textual files, and ontologies respectively. Secondly, in-memory mechanism deals with the data structure used to organize the cases in memory. To accommodate different data structure, including linear, tree structure, case retrieval nets etc., different CBR case base interfaces, including lineal case base, cached lineal case base, and ID indexed lineal case base are used.

In the present work, patient cases, which are currently simulated data using information from clinical partners as reference are stored using MySQL in a linear structure using lineal case base and the case base connector is implemented to access and retrieve cases from the case base. Hibernate library is internally used to execute the case base connector. To access data from the BCKM for semantic similarity measure computation, ontology connector is used. Internally, OntoBridge library is used to connect to the ontology.

The query patient case is stored outside the CB-DSS, within the DESIMS component. The exchange of patient data between the DESIMS and the CB-DSS is done through the HAPI FHIR server. Using the standard FHIR resources, including Patient, Observation, Condition, BodySite, and Specimen, the FHIR server retrieves and stores the query case from the DESIMS as a patient bundle. The patient bundle is then parsed and decoded by the CB-DSS for further analysis.

The attributes present in the patient case are divided into two case components, namely description component (patient description) and solution component (decision made by the BU). Similarity analysis is performed on the description variables to retrieve similar patient case(s) with description and solution variables. For example, in a patient case, the description component would include variables such as patient's age, TNM staging, BIRADS, tumour size etc., whereas the solution component would include variables such as surgery, endocrine therapy, chemotherapy, and radiotherapy. In the present work, the patient cases present in the case base would include both description and solution components, whereas the query case would include only the description component.

3. A Case Study

To demonstrate the complete operation of the CB-DSS, we present a case study by testing a labeled query case. We consider a woman, aged 50, one pregnancy, one child, breast size = 95A, in premenopausal status. Clinical examination of left breast indicates one nodule of 25 mm at the union of external quadrants. Clinical exam of the ipsilateral axillary area shows one clinical node of size 10mm. Mammography shows the breast tumour size is 30 mm, BIRADS=4. Ultrasound exam shows one breast nodule of dimension 30 mm and one enlarged axillary lymph node of size larger than 10 mm. Microbiopsy of breast nodule indicates ductal invasive carcinoma, Estrogen receptor (ER)= 95%, Progesterone receptor (PR)= 60%, HER2= 0, SBR3, Ki67= 20% and tumoral cells are found in the cytopuncture of the axillary node. TNM staging is cT2, N1, and M0.

During run time, the query case is sent from the DESIMS to the CB-DSS through the FHIR server. The patient bundle from FHIR server is then parsed and decoded for further analysis by the CB-DSS. Attributes are read as the description component of the case. Depending on the data type of these individual attributes, different similarity functions are employed to measure local similarities. For example, for attributes

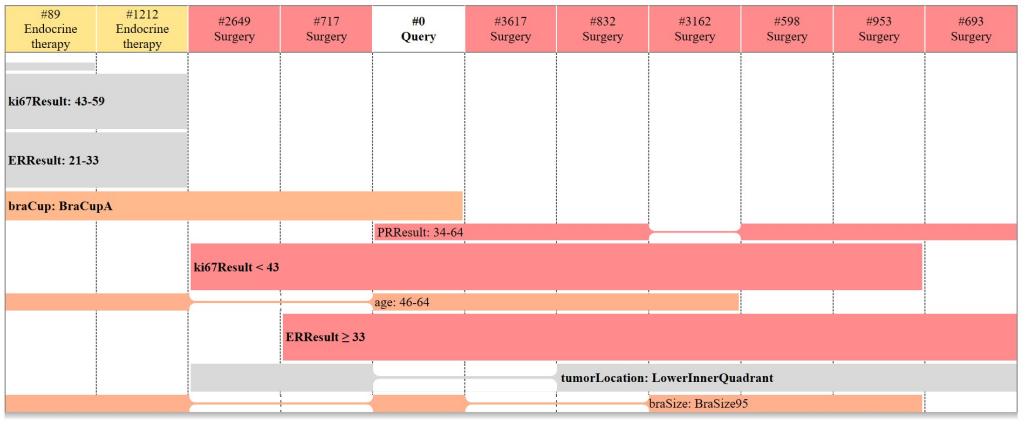

Fig. 5. Visual analytics GUI for CB-DSS: results for the case study

with numeric data type, namely age, tumour size, ER, PR and Ki67, numeric local similarity functions are used. For attributes, which have a straight list of values namely gender, menopausal status, BIRADS category, etc., enumerated similarity function is used. Utilizing the knowledge codified in the BCKM presented in this paper, semantic similarity functions defined in Eqs. (1) ~ (4) are used to measure local similarities, taking into account of the degree of taxonomical proximity in the ontology. The semantic similarity function is applied to attributes, namely breast TNM staging – cT, cN and cM, and histological type. Finally, when all local similarity measures are computed, the global similarity measure is computed using k-NN (Eq. 6), to retrieve the top k similar cases from the case base.

Retrieved patient cases are then compared with the query case through a rainbow box visual analytics GUI and displayed to the clinician/user. The CB-DSS results for the above query case are shown in Fig. 5. In rainbow boxes, each column corresponds to a patient: the "query" column with a white header represents the query patient, and the other columns represent the similar patients retrieved by the CB-DSS. The colour of the column headers indicates the type of treatment prescribed to these similar patients (here yellow for endocrine therapy and red for surgery). We can see that most similar patients were treated by surgery.

The colored boxes below the column headers represent the characteristics shared by several patients. The boxes give the evidence of why patient cases are similar. A given box covers the columns corresponding to the patients sharing the characteristic: *e.g.* patients #89 and #1212 have a Ki67 value between 43 and 59. Holes appear in boxes when patients sharing a given characteristic are not contiguous (*e.g.* the PR result 34-64 box). Columns are placed in order to minimize the number of holes.

Boxes that do not cover the query patient are represented in grey color. The other boxes are colored and their color is the mean of the colors associated with the columns they cover. Thus, a box representing a characteristic present only in patients treated by surgery will be red (e.g., PR result 34-64), and a box mixing patients with surgery and endocrine therapy will be orange (e.g., age 46-64).

In Fig. 5, we can see that most similar patients were treated by surgery, and that most of the characteristics of the query patient (*i.e.* the colored boxes) are associated to red or reddish colors. Therefore, surgery seems the most appropriate type of treatment for the query case.

4. Conclusions

In this work, we have developed the CB-DSS module of the European-funded DESIREE project which aims at providing a web-based software for the therapeutic management of breast cancer. The proposed CB-DSS

provides a tool for querying former cases in order to retrieve similar patient cases based on the defined case base and a case retrieval algorithm using a semantic similarity measure. The paper presents the overall framework of the proposed CB-DSS and systematically describes its workflow from clinical data integration to visualization of results to the user. We defined a BCKM to represent the domain knowledge as an ontology and we integrated a semantic similarity function to improve the similarity analyser and retrieval algorithm. To technically validate the proposed CB-DSS, the workflow of the framework is evaluated on a case study using a labelled query case. However, as the CB-DSS was developed based on simulated data generated using information from clinical partners, the medical relevance of the similarity measure, and the quality of the k most similar cases retrieved could not yet be validated. Further work is required, in terms of assigning weights to the attributes for computing local similarity measures, examining different similarity functions, and testing the relevance of the retrieval algorithm in a larger and reliable case base to clinically validate the proposed model.

Acknowledgements

The DESIREE project has received funding from the European Union´s Horizon 2020 research and innovation program under grant agreement No. 690238.

References

1. World Cancer Research Fund International, Breast Cancer Statistics, Retrieved from http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics on Nov. 2017.
2. World Health Organization, Breast Cancer: Prevention and Control, Retrieved from http://who.int/cancer/detection/breastcancer/en/ on Oct. 2017.
3. World Wide Cancer Research. Retrieved from https://www.worldwidecancerresearch.org/projects/ on Oct. 2017.
4. Cancer Research UK, Retrieved from http://www.cancerresearchuk.org/our-research/our-research-by-cancer-type/our-research-into-breast-cancer on Oct. 2017.
5. National Cancer Center – The Breast Cancer Project, Retrieved from http://nationalcancercenter.org/NCC_BreastCancer.html on Oct. 2017.
6. Breast Cancer UK – Preventing Breast Cancer, Retrieved from http://breastcanceruk.org.uk/grants-and-research-funding on Oct. 2017.
7. D. Lüftner, Clinical practice guidelines for breast cancer: current limitations, *The Breast Journal*. 21(4) (2015) 450 - 453.
8. I. Watson, F. Marir, Case-based reasoning: A review, *The Knowledge Engineering Review,* 9(4) (1994) 327-354.
9. R. Bergmann, K.D. Althoff, M. Minor, M. Reichle and K. Bach, Case-based reasoning introduction and recent developments. *Proc. of Case based RI*, (2008).
10. C. Jian, T. Zhe, L. Zhenxing, A review and analysis of case-based reasoning research. *Proc. of International Conference on Intelligent Transportation, Big Data and Smart City*, eds. R. Bilof (Halong Bay, Vietnam 2015), pp. 51-55.
11. N. Choudhury, S.A Begum, A Survey on Case-based Reasoning in Medicine. *International Journal of Advanced Computer Science and Applications.* 7(8) (2016) 136-144.
12. A. Holt, I. Bichindaritz, R. Schmidt, and P. Perner, Medical applications in case-based reasoning. *The Knowledge Engineering review, Cambridge University Press,* 1(4) (2005).
13. P. Koton, A medical reasoning program that improves with experience. *Computer Methods and Programs in Biomedicine*, 30(2) (1989) 177-184.
14. I. Bichindaritz, Improving case-based reasoning for an application in psychiatry. *Artificial Intelligence in Medicine: Applications of Current Technologies*, Stanford, CA, (1996) 14–20.
15. R. Bareiss, Exemplar-based knowledge acquisition: A unified approach to concept representation, classification, and learning. *Perspectives in Artificial Intelligence*, 2 (1989) 1–169.
16. R.T. Macura and K.J. Macura, Mcrad: Radiology image resource with a casebased retrieval system, in *Lecture Notes in Computer Science* (*Lecture Notes in Artificial Intelligence*), *Case-Based Reasoning and Development*. *ICCBR* 1995, eds. M. Veloso and A. Aamodt (Springer, Berlin, Heidelberg, 1995) pp. 43–54.
17. J. M. Corchado, J. Bajo and A. Abraham, GerAmi: Improving healthcare delivery in geriatric residences. *Intelligent Systems*, 23(2) (2008) 19-25.
18. H. Ahn and K. J. Kim, Global optimization of case-based reasoning for breast cytology diagnosis. *Expert Systems with Applications*, 36(1) (2009) 724-734.
19. M. Li, X. Chen, X. Li, B. Ma and P.M.B. Vitanyi, The similarity metric. *IEEE Transactions on Information Theory, 50*(12) (2004) 3250 – 3264.
20. P. ElKafrawy and R.A. Mohamed, Comparative study of case based reasoning software. *International Journal of Scientific Research and Management Studies,* 1(6) (2015) 224-233.

21. A. Atanassov, Comparative analysis of contemporary case based reasoning software frameworks, in *Proc. of 12th International Congress: Machine Technologies, Materials,* (Varna, Bulgaria 2015) pp. 27-30.
22. A. Atanassov and L. Antonov, Comparative analysis of case based reasoning software frameworks jColibri and myCBR. *Journal of the University of Chemical Technology and Metallurgy.* 47(1) (2012) 83-90.
23. J.A. Recio-García, P.A. Gonzálex-Calero and B. Díaz-Agudo, jColibri2: A framework for building Case-based reasoning systems. *Science of Computer Programming.* 79 (2014) 126-145.
24. B. Séroussi, G. Guézennec, J.B. Lamy, N. Muro, N. Larburu, B.D. Sekar, C. Prebet and J. Bouaud, Reconciliation of multiple guidelines for decision support: a case study on the multidisciplinary management of breast cancer within the DESIREE project, in *Proc. AMIA Annual Symposium,* (Washington DC, USA 2017) pp. 1527-1536.
25. N. Larburu, N. Muro, I. Macía, E. Sánchez, H. Wang, J. Winder, J. Boaud and B. Séroussi, Augmenting guideline-based CDSS with experts' knowledge, in *Proc. of* 10th *International Conference on Health Informatics*, (Porto, Portugal 2017) Vol. 5, pp. 370-376.
26. HL7.org. FHIR Release 3 (STU). Retrieved from http://hl7.org/fhir/summary.html on Oct 2017.
27. J. B. Lamy, H. Berthelot, C. Capron and M. Favre, Rainbowboxes: a new technique for overlapping and two applications in the biomedical domain. *Journal of Visual Language and Computing*, 43 (2017) 71-82.
28. J. D. Ullman, Information integration using logical views, in *Proc. of ICDT* of Lecture noes in Computer Science, (Springer, Berlin, Heidelberg 1997) vol. 1186, pp. 19-40.
29. G. De Giacomo, D. Lembo, M. Lenzerini, A. Poggi and R. Rosati, Using Ontologies for Semantic Data Integration, in *A Comprehensive Guide Through the Italian Database Research Over the Last* 25 *Years*. *Studies in Big Data*, eds. S. Flesca, S. Greco, E. Masciari, D. Saccà (Springer, Cham 2018) vol. 31, pp. 187-202.
30. V. Gligorijević and N. Pržulj, Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface*, 12(112) (2015).
31. J.A. Recio-Garía, B. Díaz-Agudo Ontology based CBR with jCOLIBRI, in *Applications and Innovations in Intelligent Systems* XIV, eds. R. Ellis, T. Allen and A. Tuson (Springer, London 2007) pp.149-162.