*Integrated Citizen Centered Digital Health and Social Care*
*A. Värri et al. (Eds.)*


*doi:10.3233/SHTI200718*

# Creating Synthetic Patients to Address Interoperability Issues: A Case Study with the Management of Breast Cancer Patients

Akram REDJDAL[a1], Jacques BOUAUD[b,a], Gilles GUÉZENNEC[a], Joseph GLIGOROV[c,d] and Brigitte SEROUSSI[a,c]
[a] *Sorbonne Université, Université Sorbonne Paris Nord, INSERM, UMR S_1142, LIMICS, Paris, France*
[b] *AP-HP, DRCI, Paris, France*
[c] *AP-HP, Hôpital Tenon, Paris, France*
[d] *Sorbonne Université, Institut Universitaire de Cancérologie, Paris, France*

**Abstract.** Interoperability issues are common in biomedical informatics. Reusing data generated from a system in another system, or integrating an existing clinical decision support system (CDSS) in a new organization is a complex task due to recurrent problems of concept mapping and alignment. The GL-DSS of the DESIREE project is a guideline-based CDSS to support the management of breast cancer patients. The knowledge base is formalized as an ontology and decision rules. OncoDoc is another CDSS applied to breast cancer management. The knowledge base is structured as a decision tree. OncoDoc has been routinely used by the multidisciplinary tumor board physicians of the Tenon Hospital (Paris, France) for three years leading to the resolution of 1,861 exploitable decisions. Because we were lacking patient data to assess the DESIREE GL-DSS, we investigated the option of reusing OncoDoc patient data. Taking into account that we have two CDSSs with two formalisms to represent clinical practice guidelines and two knowledge representation models, we had to face semantic and structural interoperability issues. This paper reports how we created 10,681 synthetic patients to solve these issues and make OncoDoc data re-usable by the GL-DSS of DESIREE.

**Keywords.** Health information interoperability, Knowledge representation, Clinical decision support systems, Breast cancer.

## 1. Introduction

Today, it is common for health care to be delivered across multiple settings. Each stay generates a record, but due to the lack of interoperability between these records, quality of care can be put at risk when patients are transferred from one organization to another. Thus, cross-organizational healthcare data sharing is a major issue, and improving healthcare interoperability is a top priority for health organizations. Indeed, interoperability issues are currently common, and reusing data generated from a system by another system, for instance a clinical decision support system (CDSS), in a new organization is a complex task due to recurrent problems of alignment between data

[1] Corresponding Author, Akram Redjdal, LIMICS UMRS_1142, 15 rue de l'Ecole de Médecine, Paris, France; E-mail: redjdalakram300@gmail.com

models and semantics. Solutions have been proposed like the OMOP common data model or the FHIR exchange format, while sharing common reference terminologies (e.g., SNOMED-CT, ICD10, UMLS, etc.). But, the source and the target systems often share the same conceptual model. Thus, it remains complex to smoothly integrate existing data sources into other systems.

DESIREE[2] is a recent European-funded project which aimed at developing a web-based platform to improve the management of primary breast cancer patients. Among other services, DESIREE includes a guideline-based decision support system (GL-DSS) that the authors of this article have developed [1]. OncoDoc is another CDSS that the authors also developed previously for the management of breast cancer patients. OncoDoc has been routinely used by the multidisciplinary tumor boards (MTBs) of the Tenon hospital (Paris, France) during three years proposing guidance for 1,861 decisions [2]. As part of the final deliverable of the DESIREE project, we had to evaluate the GL-DSS. Since we were lacking a large sample of clinical data, we decided to reuse the database of clinical cases resolved with OncoDoc.

Given the two CDSSs use two different domain knowledge models and two different formalisms to represent breast cancer guidelines, the aim was to develop and implement a model transformation from OncoDoc to the GL-DSS of DESIREE that accounts for both semantic and structural interoperability issues. This paper reports the solution we implemented to deal with interoperability issues by creating synthetic patients.

## 2. Material and Methods

### *2.1. Two CDSSs, two knowledge models, two guideline representation formalisms*

OncoDoc has been developed in a documentary approach of decision support. The knowledge base is structured as a decision tree within which the user navigates while interactively answering questions that instantiate a patient clinical profile. Nodes represent decision variables and edges represent their modalities. OncoDoc data sample is made of clinical cases resolved when using OncoDoc during MTBs. Each recorded decision is attached to a "breast side" and includes a description of the patient profile as a list of instantiated clinical parameters corresponding to decision variables that are all qualitative (e.g., "tumor size" has three values, "less than 2 cm", "between 2 and 4 cm", or "more than 4 cm"), and the decision actually made by MTB physicians.

The GL-DSS of DESIREE relies on a Breast Cancer Knowledge Model (BCKM) formalized as an ontology. The BCKM allows for rule-based and subsumption-based reasoning to provide best patient-centered therapeutic recommendations. It combines a data model based on the generic Entity-Attribute-Value (EAV) model [1], the main entities being the patient, the breast side, and the lesion, each entity having attributes, and each attribute having a value that can be primitive or hierarchical (e.g., the clinical T of the TNM classification is an attribute of the side entity, and has values among cT1, cT2, cT3, cT4, or cTx).

[2] The DESIREE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690238.

### 2.2. Model Transformation

We started with the identification of correspondences between the two CDSS models, then we developed the mapping of concepts, and we finished with the comparison of the recommendations issued by both OncoDoc and the GL-DSS.

#### 2.2.1. Identification of correspondences

We identified three types of alignment between Oncodoc and BCKM concepts:

- **1-to-1 correspondences** when a variable in OncoDoc has a unique equivalent concept in the BCKM. Several distinctions can be made, as reported in Figure 1:
  - Exact matching: OncoDoc variables and BCKM concepts and their values are equivalent in both models
  - Partial matching: several OncoDoc variables are aligned with a unique BCKM concept but some values of OncoDoc variables do not have correspondence in the BCKM
  - Conditional matching: several OncoDoc variables are aligned with a unique BCKM concept and all values of OncoDoc variables do have correspondence in the BCKM

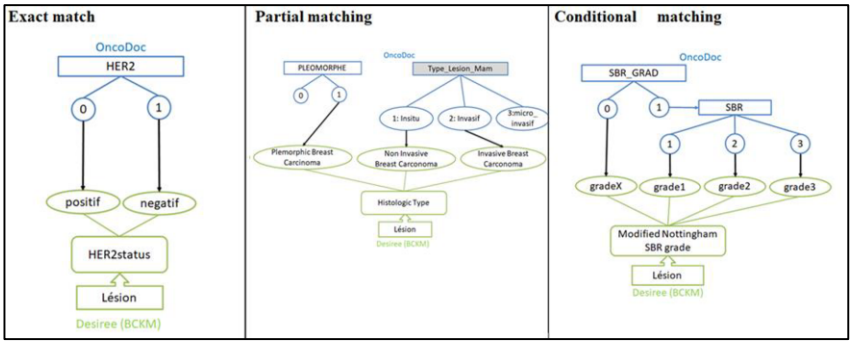


**Figure 1.** The three types of 1-to-1 correspondences

- **n-to-1 correspondences** when a variable in OncoDoc is a macro variable that relies on different sub-variables. For instance, the variable "lumpectomy contraindicated" in OncoDoc is described by different subvariables (radiotherapy contra-indicated, widespread microcalcifications, local recurrence) that have to be taken into account in the correspondence with the concept of contra-indicated lumpectomy in the BCKM.
- **1-to-n correspondences** when a value in OncoDoc has multiple correspondences in the BCKM (the tumor size & the lymph node invasion). For instance, the variable "tumor size" has three values in OncoDoc, while its BCKM equivalent concept is captured by the clinical T of the TNM classification, and correspondences are not exact as displayed in the Figure 2. For a patient with "tumor size" = "> 4 cm" in OncoDoc, there are two possible BCKM values, cT2 (which means the tumor size is more than 2cm but no more than 5cm) or cT3 (which means the tumor size is larger than 5cm). To address these semantic issues, we generated for each OncoDoc clinical case, several synthetic patients to represent all possible values of this kind of concepts in the BCKM.

### *2.2.2. Creation of synthetic patients*

The first step was to identify which variables in OncoDoc were involved in a 1-to-n correspondence. These variables were related in the BCKM either to the clinical and pathological T of TNM or the clinical and pathological N of TNM. Then we identified all patients that had at least one of these variables in their profile as recorded in the OncoDoc database and we implemented an algorithm to create synthetic patients for each of them depending on tumor size and lymph nodes invasion, e.g., if a patient had "tumor size" = "> 4cm" and "MoreThan2N" = "false" (false in OncoDoc is aligned with cN0 or cN1 in the BCKM), this patient would have 4 synthetic patients as displayed in Figure 2. The creation of synthetic patients is performed through the combinatory combinations of T and N values.

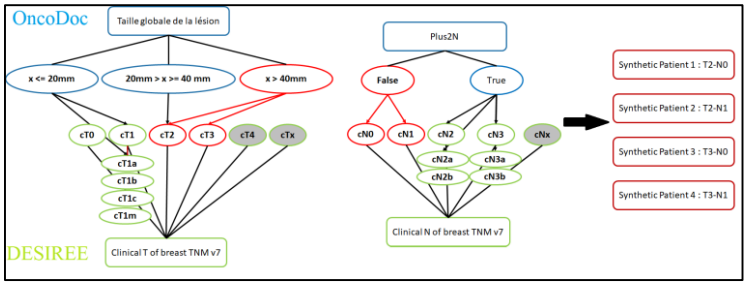


**Figure 2**. Example of two 1-to-n correspondences generating the creation of four synthetic patients.

## 3. Results

OncoDoc database included 1,861 resolved clinical cases described by a set of 61 variables. After identifying correspondences, 30 OncoDoc variables had an exact matching in the BCKM, eight had a partial matching, and five had a conditional matching. For these variables, there was no need to create synthetic patients.

We identified 18 "1-to-n" correspondences leading to the creation of synthetic patients. They were related to four main concepts in the BCKM that were added as variables in OncoDoc to be used by the algorithm implemented:

- Clinical T of TNM: this BCKM concept matches with eight OncoDoc variables related to the clinical size of the lesion. Besides, there is an additional Boolean variable "TUM-Operable" that specifies whether a tumor is operable or not. It corresponds to cT4 when the tumor is not operable, and other cT values when the tumor is operable.
- Pathologic T of TNM: this BCKM concept matches with seven OncoDoc variables describing the pathologic size of the lesion (after surgery), and depending on the cancer type (ductal or lobular carcinoma).
- Clinical N of TNM: as displayed in Figure 2, the OncoDoc variable "MoreThan2N" is related to the clinical N of TNM in the BCKM.
- Pathologic N of TNM: this BCKM concept is matched with the OncoDoc variable "LymphNodesInvasion" which refers to the result of the axillary lymph node dissection (N-, 1-to-3N+, or >4N+).

We finally created 12,542 synthetic patients, from 1,861 resolved clinical cases in OncoDoc. These BCKM-compliant patients represent all the possible representations of

OncoDoc clinical cases. Table 1 displays the distribution of synthetic patients according to their referent OncoDoc clinical cases. The average number of synthetic patients is 206. The max number of synthetic patients created for a clinical case is 35 coming from the combination of seven pN≥2 (pN2, pN2a, pN2b, pN3, pN3a, pN3b, pN3c), and five pT1 (pT1, pT1a, pT1b, pT1c, pT1mic). The category of patients with the most repetitions (766) corresponds to patients that have a unique N or no information about the N of TNM. In this case, synthetic patients are created only because of the T of TNM, with cT1, cT4, pT1 or pT4, values, thus leading to five synthetic patients (i.e., cT1, cT1a, cT1b, cT1c, cT1mic, for cT1).

**Table 1.** Distribution of synthetic patients according to OncoDoc Clinical cases.

| # synthetic patients/clinical cases | 1 | 2 | 4 | 5 | 7 | 10 | 14 | 25 | 35 |
|---|---|---|---|---|---|---|---|---|---|
| # clinical cases | 207 | 274 | 12 | 766 | 74 | 379 | 14 | 132 | 3 |

## 4. Discussion and Conclusions

We have developed an algorithm that creates synthetic patients to make the clinical cases resolved with one CDSS (OncOdoc) reused to be solved by another CDSS (GL-DSS of DESIREE). We first considered aligning OncoDoc data to OMOP [4], in order to use the common OMOP data model as a transient model, and then develop ETL tools to map concepts from OMOP to the BCKM ontology. However, matching OncoDoc to OMOP was complex because of semantic issues, and we decided to use synthetic patients to cover the missing matches.

The lack of clinical data is a common problem in health information technology. It has hindered innovation and raised the barrier of entry into the industry which lags behind other industries involving information technology, data exchange, and interoperability. The main reason comes from data privacy and relies on the problem of re-identification. Approaches and tools have been proposed to generate synthetic data [4] and some tools were validated [5]. To evaluate the GL-DSS of DESIREE, the next step is to enrich the BCKM ontology and add all concepts related to OncoDoc as attributes with their values to be able to run the GL-DSS on all the cohort of synthetic patients, and compare the recommendations produced by the GL-DSS and the decision taken by MTB physicians with OncoDoc.

## References

[1] Bouaud J, Guézennec G, Séroussi B. Combining the Generic Entity-Attribute-Value Model and Terminological Models into a Common Ontology to Enable Data Integration and Decision Support. Stud Health Technol Inform. 2018;247:541-545.

[2] Séroussi B, Laouénan C, Gligorov J, Uzan S, Mentré F, Bouaud J. Which breast cancer decisions remain non-compliant with guidelines despite the use of computerised decision support?. Br J Cancer. 2013;109(5):1147-1156.

[3] Chakrabarti S, Sen A, Huser V, et al. An Interoperable Similarity-based Cohort Identification Method Using the OMOP Common Data Model version 5.0. J Healthc Inform Res. 2017;1(1):1-18.

[4] Walonoski J, Kramer M, Nichols J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record [published correction appears in J Am Med Inform Assoc. 2018;25(7):921]. J Am Med Inform Assoc. 2018;25(3):230-238

[5] Chen J, Chun D, Patel M, Chiang E, James J. The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med Inform Decis Mak*. 2019;19(1):44. doi:10.1186/s12911-019-0793-0