# Interpretation of Best Medical Coding Practices by Case-Based Reasoning — A User Assistance Prototype for Data Collection for Cancer Registries

Michael Schnell[1,2], Sophie Couffignal[1], Jean Lieber[2], Stéphanie Saleh[1], and Nicolas Jay[2,3]

[1] Department of Population Health, Luxembourg Institute of Health, 1A-B, rue Thomas Edison, L-1445 Strassen, Luxembourg, firstname.lastname@lih.lu
[2] UL, CNRS, Inria, Loria, F-54000 Nancy, firstname.lastname@loria.fr
[3] Service d'évaluation et d'information médicales, Centre Hospitalier Régional Universitaire de Nancy, Nancy, France, n.jay@chru-nancy.fr

## 1 Introduction

There are numerous cancer registries around the world collecting data about cancers diagnosed and/or treated in a given area. This data is used to monitor cancer (incidence rates, survival rates, etc.) and to evaluate cancer care (diagnosis, treatment, etc.). To produce comparable data, common definitions (e.g. terminologies like the International Classification of Diseases (ICD)) and coding practices [5] have to be followed. However, the broadness and complexity of these standards make the work of the medical staff in charge of coding (operators) more difficult.

The aim of this research is to address this complexity, by assisting both operators and coding experts in the interpretation of coding best practices.

As an illustrating example, let us consider the case of a particular woman. In 2016, multiple pulmonary opacities were discovered within her right lung lobe. A CT scan indicated no mediastinal adenopathy.[4] A histological analysis of a sample identified the morphology[5] of the cancer as adenocarcinoma. The TTF1 marker test was positive. After further testing, another tumor is found in the ovaries. An operator might wonder which topography[6] should be coded (lung or ovaries?) and request help to answer the question. For the Luxembourg National Cancer Registry (NCR), operators ask their questions using an online ticketing system. With the free text description provided by operators, coding experts provide a solution, i.e. an answer with their reasoning in the form of a motivated argument.

Section 2 describes an approach to assist the data collection process for cancer registries and how case-based reasoning (CBR [1]) is applied. In Section 3, a

[4] An adenopathy is an enlargement of lymph nodes, likely due to cancer.
[5] The morphology describes the type and behavior of the cells that compose the tumor.
[6] The topography is the location where the tumor originated.

prototype and preliminary results are discussed. Section 4 presents a conclusion and points out what further efforts need to be undertaken in the future.

## 2 Case-based interpretation of best practices

This article summarizes the work presented in [8] and adds a description of the developed prototype and some preliminary results.

*Preliminaries.* A case ($\texttt{srce}, \texttt{sol}(\texttt{srce})$) is composed of two parts: 1) a patient record and a question, and 2) a solution. The patient record represents the data from the hospital patient record (patient features, tumors, exams, treatments, etc.) needed to answer the question. The relevant data depends on the subject and is defined by coding experts. The patient record is represented by an RDFS graph [3]. Body parts and cancer morphologies use classes from the SNOMED Clinical Terms[7] ontology. The question indicates the subject (incidence date, topography, tumor nature, etc.). In the example, the question is about the topography. The solution contains the answer to the question and the most important arguments in favor of (**pros**) and against (**cons**) this answer. In the example, the answer is to consider the topography to be the ovaries. The presence of multiple pulmonary opacities is an argument in favor, as they are indicative of a metastasis and thus the tumor is unlikely to have originated in the lungs.

The arguments have two uses. They help explain the answer to operators and serve as a reminder for coding experts. They are also used in the proposed approach during the retrieval step. Three types of arguments will be considered: strong pros, weak pros and weak cons. The difference between a strong and a weak argument comes from their reliability for a given conclusion. A strong argument is considered to be a sufficient justification for an answer, unlike a weak argument which is more of an indication or clue. It can be noted that there are no strong cons in the source cases. Indeed, such an argument would be an absolute argument against the given answer. Formally, an argument is a function $\texttt{a}$ that associates a Boolean to a case and is stored as a SPARQL ASK query.

*Global architecture.* The proposed approach uses a 4-R cycle (retrieve, reuse, revise, retain) adapted from [1] and four knowledge containers [7] (case base, domain knowledge, retrieval knowledge, adaptation knowledge).

*Retrieve.* The proposed approach relies on arguments to find similar cases. Indeed, similar answers should have similar reasoning and thus the same arguments should apply. Our method checks the applicability of the arguments from the source cases on the target problem $\texttt{tgt}$ and uses this to decide which source case is the most appropriate to solve $\texttt{tgt}$. This comparison between two source cases $\texttt{i}$ and $\texttt{j}$ relies on three criteria, one for strong arguments $\Delta^{\texttt{s}}_{\texttt{i,j}}$, one for weak arguments $\Delta^{\texttt{w}}_{\texttt{i,j}}$ and one for patient record similarity $\Delta^{\texttt{dist}}_{\texttt{i,j}}$. For the strong arguments, the source case with the most applicable strong arguments is preferred. For the

[7] https://bioportal.bioontology.org/ontologies/SNOMEDCT

weak arguments, a combination of pros and cons is used. The more weak pros and the less weak cons are applicable, the more suited the source case. For the last criterion, the patient record similarity with the target problem is used (using a graph edit distance [4]). The three criterion are considered lexicographically, first $\Delta^{\mathtt{s}}_{\mathtt{i,j}}$, then $\Delta^{\mathtt{w}}_{\mathtt{i,j}}$ and finally $\Delta^{\mathtt{dist}}_{\mathtt{i,j}}$ (see [8]).

*Reuse.* Once an appropriate source case has been found, the solution associated to the source case is copied: $\mathtt{sol(tgt)} := \mathtt{sol(srce)}$. The arguments that do not apply to the target problem, if any, are removed.

*Revise and retain.* The newly formed case $(\mathtt{tgt}, \mathtt{sol(tgt)})$ can be reviewed by a coding expert, to modify the answer, the arguments and/or the patient record. A coding expert may choose to remove unnecessary information form the patient record, removing unwanted specificity. Thus, $(\mathtt{tgt}, \mathtt{sol(tgt)})$ is substituted by $(\mathtt{tgt'}, \mathtt{sol(tgt')})$, where $\mathtt{tgt'}$ is more general than $\mathtt{tgt}$. $(\mathtt{tgt'}, \mathtt{sol(tgt')})$ is a generalized case that has a larger coverage than $(\mathtt{tgt}, \mathtt{sol(tgt)})$ [6].

## 3 Prototype and preliminary results

The prototype designed for the NCR serves as a ticketing system, where operators ask coding questions and experts provide answers. It assists operators in structuring questions, making it easier for the NCR and coding experts to find similar questions later. For topography questions, it will also provide a tentative answer. This answer is calculated using the approach described in [8]. All the answers are reviewed by experts. The prototype presents itself as a single page application built using Angular[8] with a backing REST API built with Go[8] and the Gin framework.[9] The data is stored in a triple store Apache Jena[10] and exposed as a SPARQL endpoint using Apache Fuseki.[11]

The prototype was tested internally, to perform a first assessment of its usability and utility. Some old cases concerning the topography were formalized and coded, with some domain knowledge. For the arguments, great care was given during modeling in order to make them more broadly applicable. Then new questions were presented to the system, and the proposed solution compared with the expected ones. While the prototype answered every question, not all of them were correct. The main reasons for the difference were the small amount of cases (15 originally, however the case base will be enriched by routine usage) and the simple reuse method used at this stage. Indeed, as the arguments have been formalized to be more general, some of the provided answers might be slightly incorrect (e.g. answering upper lung lobe instead of lower lung lobe). Despite this, as the prototype displays the reused source case, an operator should be able to make the necessary adaptation to the provided solution.

[8] https://angular.io
[9] https://golang.org, https://github.com/gin-gonic/gin
[10] https://jena.apache.org/
[11] https://jena.apache.org/documentation/fuseki2/

For the questions concerning other subjects, the prototype relies entirely on the coding experts to provide answers.

## 4 Conclusion

Recently there has been a growing interest for case-based reasoning applications in health sciences [2]. In this paper, an approach to assist operators in the interpretation of best medical coding practices has been proposed. This approach is based on discussions with operators and coding experts on actual coding problems. A dozen tricky problems were discussed in detail, among a hundred simpler problems. The coding questions asked by the operators are compared to previous questions and solved by reusing the pros and cons of previously given solutions. The results discussed are only preliminary and a more thorough evaluation, including the operators and coding experts, is planned.

At the moment the reasoning process is only partial. Arguments are only a part of a more complex reasoning process. The formalization of this process and the eventual integration of the coding standards remains an interesting avenue for future work.

After the prototype has been validated and improved by routine usage, a second version will be designed that is less domain-dependent. The objective is to build a generic system for argumentative case-based reasoning using semantic web standards.

***Acknowledgments.*** The first author would like to thank the Fondation Cancer for their financial support.

## References

1. Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches (1994)
2. Bichindaritz, I., Marling, C., Montani, S.: Case-based Reasoning in the Health Sciences. In: Workshop Proceedings of ICCBR (2015)
3. Brickley, D., Guha, R.V.: RDF Schema 1.1, https://www.w3.org/TR/rdf-schema/, W3C recommendation, last consultation: March 2017 (2014)
4. Bunke, H., Messmer, B.T.: Similarity measures for structured representations. In: European Workshop on Case-Based Reasoning. pp. 106–118. Springer (1993)
5. European Network of Cancer Registries and Tyczynski, Jerzy E and Démaret, D and Parkin, D Maxwell: Standards and guidelines for cancer registration in Europe: the ENCR recommendations. International Agency for Research on Cancer (2003)
6. Maximini, K., Maximini, R., Bergmann, R.: An investigation of generalized cases, pp. 261–275. Springer (2003)
7. Richter, M.M., Weber, R.O.: Case-based reasoning: a textbook. Springer Science & Business Media (2013)
8. Schnell, M., Couffignal, S., Lieber, J., Saleh, S., Jay, N.: Case-Based Interpretation of Best Medical Coding Practices — Application to Data Collection for Cancer Registries. In: Conference Proceedings of ICCBR (2017)