

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

Michael Schnell, Sophie Couffignal, Jean Lieber, Stéphanie Saleh, Nicolas Jay

▶ **To cite this version:**

Michael Schnell, Sophie Couffignal, Jean Lieber, Stéphanie Saleh, Nicolas Jay. Interpretation of Best Medical Coding Practices by Case-Based Reasoning - A User Assistance Prototype for Data Collection for Cancer Registries. JWAIH 2018 - Joint Workshop on Artificial Intelligence in Health, Jul 2018, Stockholm, Sweden. hal-01907093

HAL Id: hal-01907093
https://hal.science/hal-01907093

Submitted on 28 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Abstract. In the fight against cancer, cancer registries are an important tool. At the heart of these registries is the data collection and coding process. This process is ruled by complex international standards and numerous best practices, which can easily overwhelm (coding) operators. In this paper, a system assisting operators in the interpretation of best medical coding practices and a short evaluation are presented. By leveraging the arguments used by the coding experts to determine the best coding option, the proposed system answers coding questions from operators and provides a partial explanation for the proposed solution.

Keywords: interpretation of best practices · interpretive case-based reasoning · coding standards · cancer registries · user assistance · decision support

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 denoted by `exmpl` of a particular woman. In 2016, multiple pulmonary opacities were discovered within

her right lung. 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 can request help. For the Luxembourg National Cancer Registry (NCR), operators ask their questions using an online ticketing system. With 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 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 [9] and adds a description of the developed prototype and some preliminary results.

2.1 Preliminaries

RDFS[7] is a knowledge representation language of the semantic web. SPARQL[7] is a query language for RDFS web.

A case ($\texttt{srce}, \texttt{sol}(\texttt{srce})$) is composed of two parts: (1) `srce` is a problem given by a question (i.e. a subject) and a patient record, and (2) $\texttt{sol}(\texttt{srce})$ a solution for the problem `srce`.

The question indicates the subject (incidence date, topography, tumor nature, etc.). In the example, the question is about the topography.

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] (see figure 1). Body parts and cancer morphologies use classes from the SNOMED Clinical Terms[8] ontology.

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 lung 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

[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.
[7] https://www.w3.org/TR/rdf-schema/ and https://www.w3.org/TR/sparql11-query
[8] https://bioportal.bioontology.org/ontologies/SNOMEDCT

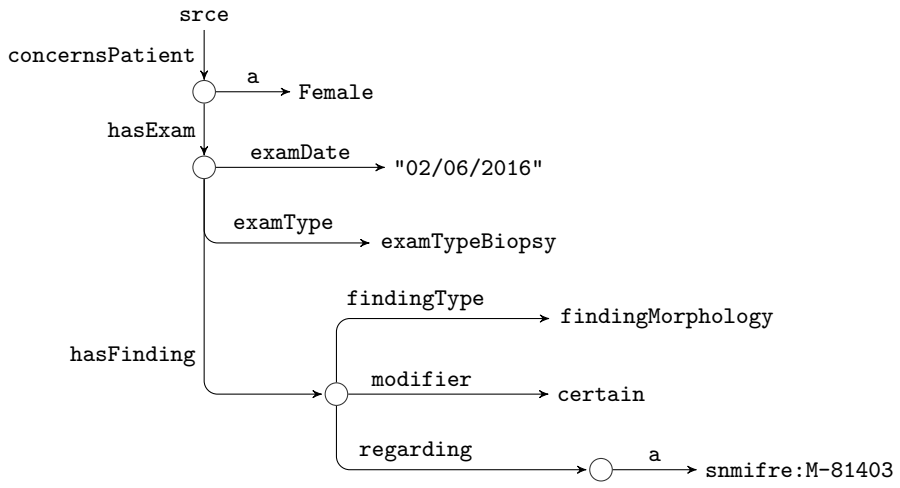


Fig. 1. Short patient record in RDFS. This graph represents a woman with a single biopsy (exam), identifying the tumor as adenocarcinoma (which is coded as M-8140/3). The circles represent blank nodes.

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 that associates a Boolean to a case and is stored as a SPARQL ASK query. The following shows an argument `arg`, followed by an explanation:

$\texttt{arg(case)} =$

```
ASK {
  case concernsPatient ?patient .
  ?patient hasExam ?exam_morpho .
  ?exam_morpho hasFinding ?finding .
  ?finding findingType findingTypeFindMorphology .
  ?finding modifier certain .
  ?finding regarding [ a snmifr:M-81403 ] .
  ?patient hasExam ?exam_ttf .
  ?exam_ttf hasFinding ?finding .
  ?finding findingType findingTypeFindTTF1Marker .
  ?finding present yes .
}
```

`arg` says that a TTF1 positive adenocarcinoma is in favor of a primitive lung cancer. The argument checks that the morphology of the tumor is of type adeno-

carcinoma and that the tumor is positive for the TTF1 marker. This argument applies for the example described in the introduction, i.e. `arg(exmpl) = TRUE`.

2.2 Global architecture

The proposed approach uses a 4-R cycle (retrieve, reuse, revise, retain) adapted from [1] and four knowledge containers [8] (case base, domain knowledge, retrieval knowledge, adaptation knowledge), as shown in figure 2.

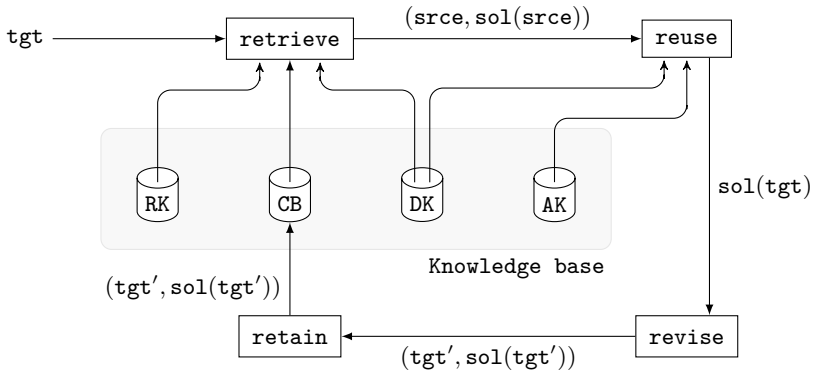


Fig. 2. Adapted 4-R cycle and knowledge containers for the proposed approach.

2.3 Retrieve

The proposed approach relies on arguments to find similar cases. Indeed, similar answers should be based on similar reasoning and thus the same arguments should apply. Our method checks the applicability of arguments from source cases on the target problem `tgt` and uses this to determine the preferred source case to solve `tgt`. This preference relation is denoted by the preorder $\preccurlyeq_{\mathtt{tgt}}$. The comparison between two source cases `i` and `j` relies on three criteria, $\mathcal{C}_{\mathtt{s}}$ for strong arguments, $\mathcal{C}_{\mathtt{w}}$ for weak arguments and $\mathcal{C}_{\mathtt{dist}}$ for patient records.

An argument `arg` is applicable for a case `c` if the preconditions of the argument are met in the patient record of `c`. For the argument `arg` described in the preliminaries, `arg` applies for a case if the patient record contains at least two exams, one identifying the morphology as adenocarcinoma and another exam reporting a TTF1 positive tumor. Formally an argument `arg` is applicable for a case `c` if `arg(c) = TRUE`.

For the criterion $\mathcal{C}_{\mathtt{s}}$, the source case with more applicable strong arguments is preferred. Formally, $\mathcal{C}_{\mathtt{s}}$ is met if $\Delta^{\mathtt{s}}_{\mathtt{i},\mathtt{j}} > 0$, where $\Delta^{\mathtt{s}}_{\mathtt{i},\mathtt{j}}$ is defined as

$$\Delta^{\mathtt{s}}_{\mathtt{i},\mathtt{j}} = \mathcal{N}^{\mathtt{sp}}(\mathtt{srce}_{\mathtt{i}}, \mathtt{tgt}) - \mathcal{N}^{\mathtt{sp}}(\mathtt{srce}_{\mathtt{j}}, \mathtt{tgt})$$

and $\mathcal{N}^{\texttt{args}}(\texttt{srce}, \texttt{tgt})$ denotes the number of arguments of type $\texttt{args}$ of a the source case $\texttt{srce}$ which are applicable for a case $\texttt{tgt}$, i.e.

$$\mathcal{N}^{\texttt{args}}(\texttt{srce}, \texttt{tgt}) = |\{\texttt{a} \in \texttt{args}(\texttt{srce}) \mid \texttt{a}(\texttt{tgt}) = \texttt{TRUE}\}|$$

and $\texttt{args} \in \{\texttt{sp}, \texttt{wp}, \texttt{wc}\}$ is an argument type. $\texttt{sp}(\texttt{srce})$ is the set of strong pros, $\texttt{wp}(\texttt{srce})$ the set of weak pros and $\texttt{wc}(\texttt{srce})$ the set of weak cons of $\texttt{srce}$.

For the criterion for weak arguments $\mathcal{C}_{\texttt{w}}$, a combination of pros and cons is used. Intuitively, if more weak pros and less weak cons are applicable, the source case is preferred. Formally, $\mathcal{C}_{\texttt{w}}$ is met if $\Delta^{\texttt{w}}_{\texttt{i,j}} > 0$, where $\Delta^{\texttt{w}}_{\texttt{i,j}}$ is defined as

$$\begin{aligned}\Delta^{\texttt{w}}_{\texttt{i,j}} = \quad & \lambda_{\texttt{p}} * (\mathcal{N}^{\texttt{wp}}(\texttt{srce}_{\texttt{i}}, \texttt{tgt}) - \mathcal{N}^{\texttt{wp}}(\texttt{srce}_{\texttt{j}}, \texttt{tgt})) \\ & - \lambda_{\texttt{c}} * (\mathcal{N}^{\texttt{wc}}(\texttt{srce}_{\texttt{i}}, \texttt{tgt}) - \mathcal{N}^{\texttt{wc}}(\texttt{srce}_{\texttt{j}}, \texttt{tgt}))\end{aligned}$$

where $\lambda_{\texttt{p}}$ and $\lambda_{\texttt{c}}$ are two nonnegative coefficients that are currently fixed to $\lambda_{\texttt{p}} = 3$ and $\lambda_{\texttt{c}} = 2$. When more data are available, these parameters values will be reevaluated.

For the criterion $\mathcal{C}_{\texttt{dist}}$, a graph edit distance between patient record RDFS graphs is used [4]. Formally, $\mathcal{C}_{\texttt{dist}}$ is met if $\Delta^{\texttt{dist}}_{\texttt{i,j}} \geq 0$, where $\Delta^{\texttt{dist}}_{\texttt{i,j}}$ is defined as

$$\Delta^{\texttt{dist}}_{\texttt{i,j}} = \texttt{dist}(\texttt{srce}_{\texttt{j}}, \texttt{tgt}) - \texttt{dist}(\texttt{srce}_{\texttt{i}}, \texttt{tgt})$$

The three criteria are considered lexicographically, first $\mathcal{C}_{\texttt{s}}$, then $\mathcal{C}_{\texttt{w}}$ and finally $\mathcal{C}_{\texttt{dist}}$ (see [9]). $\texttt{srce}_{\texttt{i}}$ is preferred over $\texttt{srce}_2$, i.e $\texttt{srce}_{\texttt{i}} \preccurlyeq_{\texttt{tgt}} \texttt{srce}_{\texttt{j}}$, ifq

$$\Delta^{\texttt{s}}_{\texttt{i,j}} > 0 \texttt{ or } (\Delta^{\texttt{s}}_{\texttt{i,j}} = 0 \texttt{ and } (\Delta^{\texttt{w}}_{\texttt{i,j}} > 0 \texttt{ or } (\Delta^{\texttt{w}}_{\texttt{i,j}} = 0 \texttt{ and } \Delta^{\texttt{dist}}_{\texttt{i,j}} \geq 0)))$$

2.4 Reuse

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

2.5 Revise and retain

The newly formed case $(\texttt{tgt}, \texttt{sol}(\texttt{tgt}))$ 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 from the patient record, removing unwanted specificity. Thus, $(\texttt{tgt}, \texttt{sol}(\texttt{tgt}))$ is substituted by $(\texttt{tgt}', \texttt{sol}(\texttt{tgt}'))$, where $\texttt{tgt}'$ is more general than $\texttt{tgt}$. $(\texttt{tgt}', \texttt{sol}(\texttt{tgt}'))$ is a generalized case that has a larger coverage than $(\texttt{tgt}, \texttt{sol}(\texttt{tgt}))$ [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 [9]. All the answers are reviewed by experts. The prototype presents itself as a single page application built using Angular[9] with a backing REST API built with Go[9] and the Gin framework.[10] The data is stored in a triple store Apache Jena and exposed as a SPARQL endpoint using Apache Fuseki.[11] Figure 3 and 4 show screenshots of the prototype.

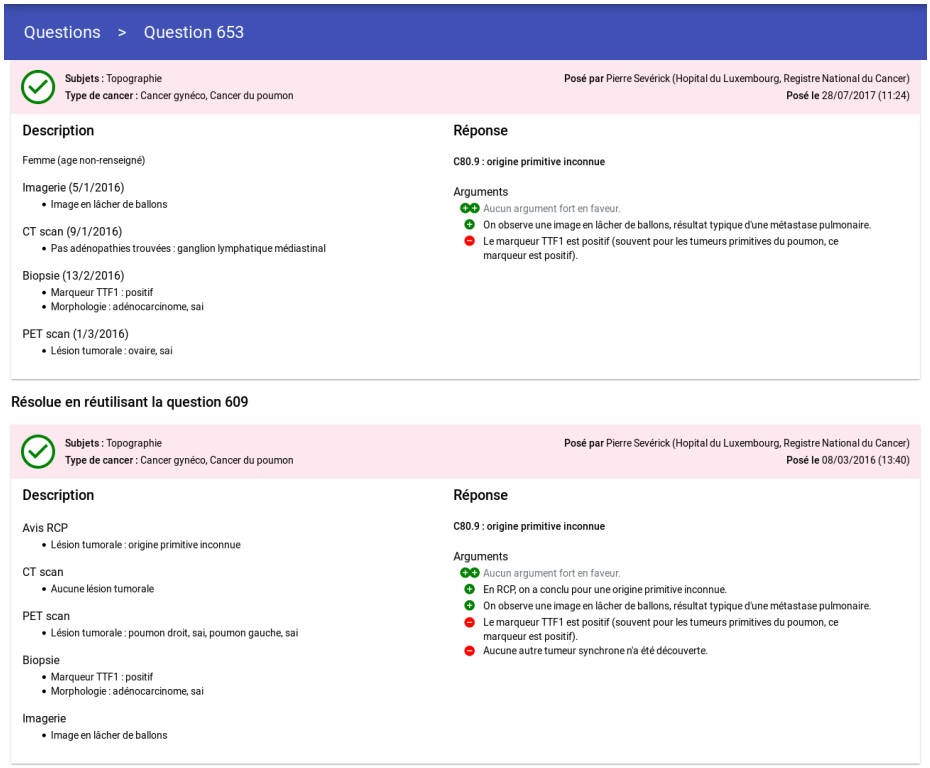


Fig. 3. Example of a solved case. The top displays the new question asked and the provided solution. The bottom displays the source case used to solve the new question.

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

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

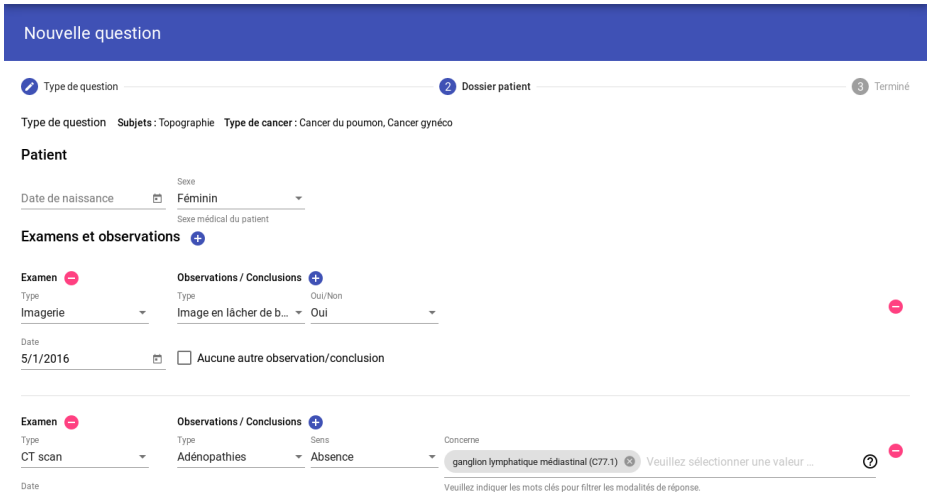


Fig. 4. Form used to describe coding questions and patient records. The French labels for body parts and morphologies are taken from the SNMIFRE (a French translation of SNOMED, http://bioportal.lirmm.fr/ontologies/SNMIFRE).

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

To the best of our knowledge, few other reseach attempts to use arguments in the context of the retrieval process. The closest method found is a work by McSherry [7]. The proposed approach creates explanations afterwards, using the closest source case to provide the conclusion and the closest source case with the opposite conclusion to compute which attributes favor the conclusion and which attributes do not. Unlike our approach, each argument is linked to a single attribute. Thus they cannot show how the combination of attributes might influence a given outcome.

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 authors wish to thank the anonymous reviewers of the Joint Workshop on Artificial Intelligence in Health for their remarks which have helped in improving the quality of the paper. The first author would also 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. McSherry, D.: Explaining the Pros and Cons of Conclusions in CBR. In: European Conference on Case-Based Reasoning. pp. 317–330. Springer (2004)
8. Richter, M.M., Weber, R.O.: Case-based reasoning: a textbook. Springer Science & Business Media (2013)
9. 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)