Case-Based Interpretation of Best Medical Coding Practices - Application to Data Collection for Cancer Registries

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

Case-Based Interpretation of Best Medical Coding Practices — Application to Data Collection for Cancer Registries

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

[1] Epidemiology and Public Health Research Unit, 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 dévaluation et dinformation médicales, Centre Hospitalier Régional Universitaire de Nancy, Nancy, France, n.jay@chru-nancy.fr

Abstract. Cancer registries are important tools in the fight against cancer. At the heart of these registries is the data collection and coding process. Ruled by complex international standards and numerous best practices, operators are easily overwhelmed. In this paper, a system is presented to assist operators in the interpretation of best medical coding practices. By leveraging the arguments used by the coding experts to determine the best coding option, the proposed system is designed to answer the coding questions from operators and provide an answer associated with 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

The Luxembourg National Cancer Registry (NCR) is a systematic, continuous, exhaustive and non redundant collection of data about cancers diagnosed and/or treated in Luxembourg. For every case matching the inclusion criteria of the NCR, data about the patient, the tumor, the treatment and the follow up are collected. The main objectives of the NCR are cancer monitoring (incidence rates, survival rates, comparisons on an international level, ...) and the evaluation of cancer case management (diagnosis, treatment, ...) in Luxembourg.

There are numerous cancer registries around the world (over 700 according to the Union for International Cancer Control[4]), with varying means and goals.

[4] http://www.uicc.org/sites/main/files/private/UICC Cancer Registries- why what how.pdf

In order for the collected data to be comparable, it is necessary to have a common definition of the collected data and the coding practices. This lead to the creation of various international coding standards, providing both common terminologies (e.g. the International Classification of Diseases (ICD)) and coding best practices [9]. It is essential to follow these standards in order to obtain standardized and reliable data. However, the broadness and complexity of the standards can make the work of the operators difficult. The operators are the people in charge of collecting and coding cancer cases. It takes months of time to attain excellence. Time and practice are essential. Complex cases add an extra level of difficulty.

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 male patient from the NCR. In 2013, he suffered from lasting pains in his side and a sudden loss of appetite. On January 12^{th}, 2014, a CT scan of his left kidney revealed nothing out of the ordinary. As the patient's condition continued to deteriorate, a second scan was made on February 15^{th}, 2014. This time, two suspicious neoplasms were found and the clinicians suspected cancer. Another CT scan made on March 10^{th}, 2014 showed signs of multiple renal adenopathy, which reinforced the cancer suspicion. On June 2^{nd}, 2014, a renal biopsy was carried out and the following histological findings pointed to a renal cell carcinoma. The operator, after reading the complete file and carefully selecting the important facts, determined that this type of cancer meets the inclusion criteria of the NCR and has to be coded into the database according to international standards. The most important values collected by the registry for this tumor are the incidence date (February 15^{th}, 2014), the topography (C64.9 – Kidney) and the morphology (M-8312/3 – renal cell carcinoma). The majority of questions concerns these values and, thus they are primary focus of this research project.

This example was rather easy to code. For the operator, the task is more complex as the data are contained within the various letters and free text reports that constitute the medical record. These documents have to be evaluated and summarized. It is possible for two reports to provide conflicting data. Here, the first CT scan showed nothing, unlike the following ones. Sometimes, important data are simply missing from the patient record. This can be the case if the patient has continued his treatment abroad or in a different hospital, if the patient died from an unrelated cause (e.g. car accident) or if the patient refused further treatment. Another possible explanation for the missing information is the difference in objectives between treatment and coding. Some aspects are assumed implicitly by the clinicians. In the case of breast cancer, no mention of a palpation usually means that no tumor is palpable, though a palpation was actually performed. However, in the case of the NCR, both exam and result must be explicitly documented. As such, aspects deemed unimportant by the clinicians might actually be very important for the registry and vice-versa. Furthermore, most medical reports do not structure their data beyond simple sectioning or identifying information (type of report, clinician, patient, ...). The important

information (e.g. the description and the conclusion) is found in the free text sections. In addition, this text can be very ambiguous (vague conclusions, inconsistencies between factual description and medical conclusion).

In order to solve these conflicts, which require not only a deep knowledge of the coding standards, but also a solid medical background, coding experts are consulted. The coding experts need to determine the coding practices which should be applied to the problematic patient record. However, as consistency is a key requirement for cancer registries (needed for temporal analysis and to track tendencies), experts have to ensure that two identical cases receive the same coding. If the standards clearly state how to solve such an instance, it is only a matter of finding the proper practices and interpreting them accordingly. This is not always possible, as the coding standards do not (and cannot) cover all possible aspects of a cancer patient. Should such a situation occur, a new practice is designed to complement existing ones. For any future identical patient, this new practice should then be applied (in order to guarantee the consistency of the registry). It is therefore crucial to remember these particular coding questions and how they were solved (e.g. what practices were eventually used).

Context and motivation are discussed in Section 2. Section 3 introduces some definitions and notations. Section 4 describes an approach to assist the data collection process for cancer registries and how case-based reasoning (CBR [1]) was applied. In Section 5, a prototype of the proposed method is described. The proposed method is discussed and compared with related work in Section 6. Section 7 presents a conclusion and points out what further efforts need to be undertaken in the future.

2 Context and motivation

For the Luxembourg National Cancer Registry, the operators can ask questions at any time using a ticketing system. The operator provides a free text description of their question with the minimum amount of required data about the patient, the tumor and, if relevant, the treatment. However, for the most part, the operator chooses what is worth providing. Of course, should anything important be missing, the experts will ask additional data or provide a tentative answer taking into account the missing data (e.g. if the missing value is A, then solution B, else solution C).

While providing a very individualized response, this approach complicates the sharing process. As the operators can only see the questions asked by other operators from the same hospital, a common question will be asked and answered several times. This repetition can lead to inconsistent answers for the same question. As consistency is an important quality measure for cancer registries, this issue needs to be addressed. As of today, this issue is remedied partially with continuous training sessions for the operators, during which the most important questions are discussed with coding experts.

Answering all the problems encountered by the operators is very time consuming for coding experts. The aim of this research is to decrease this workload.

To achieve that goal, a shared tool for both operators and coding experts is implemented. It allows the operators to ask questions, tries to answer them as best as possible and provides experts with an interface to answer the remaining questions.

Given the similarity between the working process of the experts and case-based reasoning, we have chosen to base our approach on CBR. Nevertheless, other reasoning or optimization algorithms were also considered for this task. Very popular methods are black box learning algorithms (like neural networks). Indeed, given enough representative data, this approach would yield good results. Some papers explored this in a related domain, automatic data collection or annotation. This research area focuses on the creation of solutions for the annotation and coding of electronic medical patient records. In a workshop of the 2007 BioNLP conference, a shared task focused on the assignment of ICD-9-CM codes to radiology reports [15]. Several methods were proposed with very interesting results (see [8]). However, those good results are due to two factors specific to radiology. The classification only used around 40 diagnosis codes from ICD-9-CM (out of over 14 000) and a representative data set (with proportionate representations for every possible code) was provided. While there are considerably fewer codes for cancer registries, there is no comprehensive data set available for the learning and evaluation process of any of the proposed methods. This is probably one of the major problems for this kind of method. Another weakness is the explanation. By contrast to automatic coding, for which explaining the reason why the system has chosen to code a patient record in a given way may be slightly less important, it is essential for a decision support system.

3 Preliminaries

Case-based reasoning. In a given application domain, a *case* is the representation of a problem-solving episode frequently represented by a pair $(\texttt{pb}, \texttt{sol}(\texttt{pb}))$ where `pb` is a problem related to the application domain and $\texttt{sol}(\texttt{pb})$ is a solution of `pb`. Given a new problem `tgt`—the target problem—, case-based reasoning aims at solving `tgt` by reusing a case base. A *source case* is an element of the case base. A classical way to do so consists in selecting a source case judged similar to `tgt` (retrieval step) and to reuse it to solve `tgt`.

RDFS is a knowledge representation language of the semantic web [5]. An RDFS formula is a triple `(s p o)` that can be understood as a sentence in which `s` is the subject, `p` (the predicate) is a verbal group and `o` is an object. Thus `(romeo loves juliet)` is a triple stating that mister Montague has strong feelings for miss Capulet. An RDFS base is a set of triples and is generally assimilated to an RDFS graph where nodes are subjects and objects, and where edges are labeled by properties. E.g., the graph

$$\text{romeo} \underset{\text{loves}}{\overset{\text{loves}}{\rightleftarrows}} \text{juliet} \xrightarrow{\text{age}} 13$$

states that Romeo and Juliet love each other and that Juliet is 13.

Some properties are associated with semantics, in particular `rdf:type`, abbreviated as `a` and meaning "is an instance of", and `rfds:subClassOf`, abbreviated as `subc` and meaning "is a subclass of". For example, from

$$\mathtt{G} = \text{ romeo} \xrightarrow{\text{a}} \text{Man} \xrightarrow{\text{subc}} \text{Human}$$

it can be inferred that romeo $\xrightarrow{\text{a}}$ Human.

SPARQL is a query language for RDFS. In this paper, the only type of SPARQL query used is `ASK`. This query tests the existence of a subgraph in a given graph, using variables. In SPARQL, variable names start with `?`, e.g., `?x`, `?tumor`. For example, the following query tests if someone (`?x`) (in the queried graph) loves a human (`?human`): `ASK {?x loves ?human . ?human a Human}`.

RDFS was chosen for its status as a recognized knowledge representation language, with numerous available tools. It also provides access to the Linked Open Data, which are open knowledge bases. This enables the usage of previously coded medical knowledge for the resoning tool presented in this paper.

4 Case-base interpretation of best practices

This section describes the proposed approach to assist operators in their coding task. This research project has been elaborated after discussing actual coding problems with operators and experts from the Luxembourg National Cancer Registry. First, the running example is introduced, followed by an overview of the global architecture of the system. Finally, the representation of the cases and the steps of the proposed approach are detailed.

4.1 Introduction of the running example

For the following sections, the same example will be used to explain and demonstrate the proposed approach. In the descriptions below, important patient features are in ***bold italics***.

Target problem. (`tgt`) The question concerns the nature (primary, metastasis, ...) of a lung tumor. This is a recurring question, as the lung is an organ that very easily develops metastases. As the coding of the tumor varies heavily based on its nature, it is an important question for the operator. The nature of the tumor depends on its localization and where the cancer initially developed. There are essentially two possibilities: primary or secondary. The tumor at the initial localization is the primary tumor. From that tumor, cells may detach themselves and, traveling through the body using the cardiovascular system, develop new tumors in other body parts. These new tumors are called metastases and are of secondary nature.

The target problem concerns a woman, born on December 5^{th}, 1950. In ***2006***, ***breast cancer*** was diagnosed and treated. In ***2016***, a ***lung tumor*** was discovered within the right lower lobe. A CT scan indicated ***no mediastinal adenopathy***[5]. A histological analysis of a sample identified the morphology[6] of the cancer as ***adenocarcinoma***. The ***TTF1*** marker test was ***negative***[7]. After further testing, ***no other tumor site*** was found. In the patient record, it was noted that the ***oncologist considered*** the lung tumor to be of ***primary nature***.

For our example, three source cases are described hereafter. A case is a representation of a coding episode based on best coding practices. For the sake of simplicity, all the source cases concern the same subject, i.e. the nature of a lung tumor. For each case, the patient record is described, followed by the answer and a description of the arguments in favor of and against the proposed answer.

Source 1 ($\texttt{srce}_1$) concerns a woman, born on July 23^{rd}, 1946. In ***2012***, she was diagnosed with ***breast cancer*** (***adenocarcinoma***) and treated. In ***2015***, a ***lung tumor*** was discovered within the middle left lobe. A histological analysis identified the morphology of the cancer as ***small-cell carcinoma***.

In this case, the answer to the question of the nature of the lung tumor was primary tumor. As for the argumentation, there was ***one strong argument in favor***, namely the morphology of the tumor. Indeed, small-cell carcinoma most commonly arise within the lung.

Source 2 ($\texttt{srce}_2$) concerns a woman, born on March 14^{th}, 1930. In ***2011***, she was diagnosed with ***colorectal cancer*** and treated. In ***2013***, a ***lung tumor*** was discovered within the left middle lobe. A CT scan indicated ***no mediastinal adenopathy*** and showed ***multiple pulmonary opacities*** indicative of a lung metastasis. The patient was already very fragile, thus no further tests were performed. The ***oncologist concludes*** that the lung tumor was a ***metastasis*** of the previous cancer.

In this case, the answer to the question of the nature of the lung tumor was a metastasis (of the colorectal cancer). There were ***four weak arguments in favor*** in this case: the close antecedent, the absence of mediastinal adenopathy, the oncologist's opinion and the multiple pulmonary opacities.

Source 3 ($\texttt{srce}_3$) concerns a man, born on August 14^{th}, 1953. In ***2000***, ***prostate cancer was*** diagnosed and treated. In ***2014***, a ***lung tumor*** is discovered within the upper left lobe. A CT scan indicated ***no mediastinal adenopathy***. A histological analysis of a sample identified the morphology of the cancer as ***adenocarcinoma***. After further testing, ***no other tumor site*** is found. In the patient record, it is noted that the ***oncologist considered*** the tumor to be of ***primary nature***.

[5] An adenopathy is an enlargement of lymph nodes, likely due to the cancer.
[6] The morphology describes the type and behavior of the cells that compose the tumor.
[7] For primary lung adenocarcinoma, TTF1 marker is usually positive.

In this case, the tumor was primary. There were ***three weak arguments in favor***: the oncologist's opinion, the fact that no other synchronous tumor was found and the long time span between the previous cancer and the current one (a shorter time span would have been in favor of a metastasis). There was also ***one weak argument against***, namely the absence of mediastinal adenopathy.

4.2 Global architecture

Figure 1 summarizes the main process for our approach. It uses a 4-R cycle adapted from [1] and four knowledge containers [16]. To solve a new problem, first a description must be provided. That description is then used with the domain knowledge (DK) and the retrieval knowledge (RK) to find a suitable source case $\texttt{srce}$ and its solution $\texttt{sol}(\texttt{srce})$ within the case base. Then, in the reuse step, the solution $\texttt{sol}(\texttt{tgt})$ for $\texttt{tgt}$ is produced from $(\texttt{srce}, \texttt{sol}(\texttt{srce}))$ together with the domain knowledge and adaptation knowledge (AK). The pair $(\texttt{tgt}, \texttt{sol}(\texttt{tgt}))$ may then be revised by an expert, leading to the pair $(\texttt{tgt}', \texttt{sol}(\texttt{tgt}'))$. Finally, this pair may be retained by adding it into the case base.

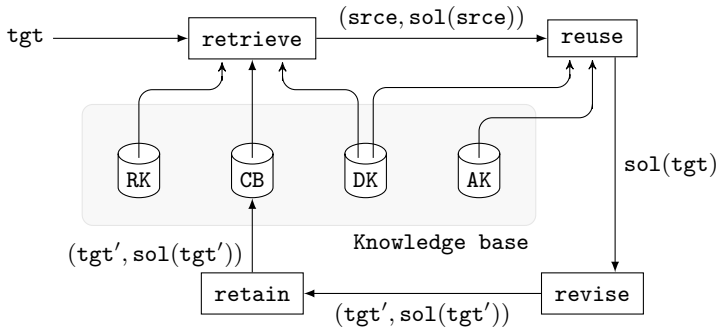

Fig. 1. Adapted 4-R cycle and knowledge containers for the proposed approach.

4.3 Case-based interpretation of best practices

A case $(\texttt{srce}, \texttt{sol}(\texttt{srce}))$ is composed of three parts: the question, the patient record and the solution.

The question part indicates the subject (incidence date, topography, tumor nature, ...) as well as the focused entity from the patient record. In the running example, the question is about the tumor's nature and focuses on the lung tumor.

The patient record represents the data from the hospital patient record (patient features, tumors, exams, treatments, ...) needed to answer the question. The relevant data depends on the subject and is defined by coding experts. For

the source cases, only the required information is provided. For the target problems, this assumption cannot be made. The operator may simply not know what is needed and thus is encouraged to provide as much information as possible.

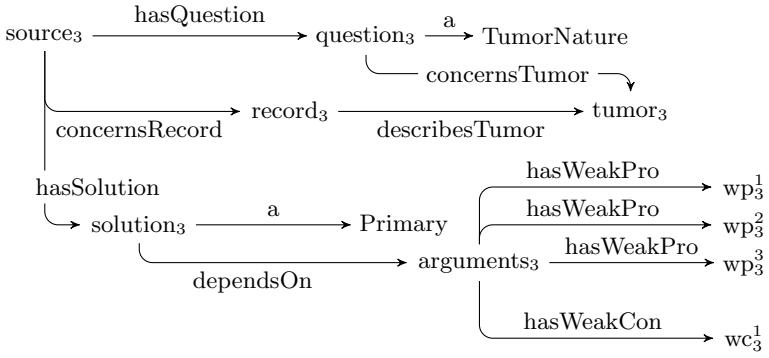

Fig. 2. Partial RDFS graph for source case `srce3` (patient record details in figure 3).

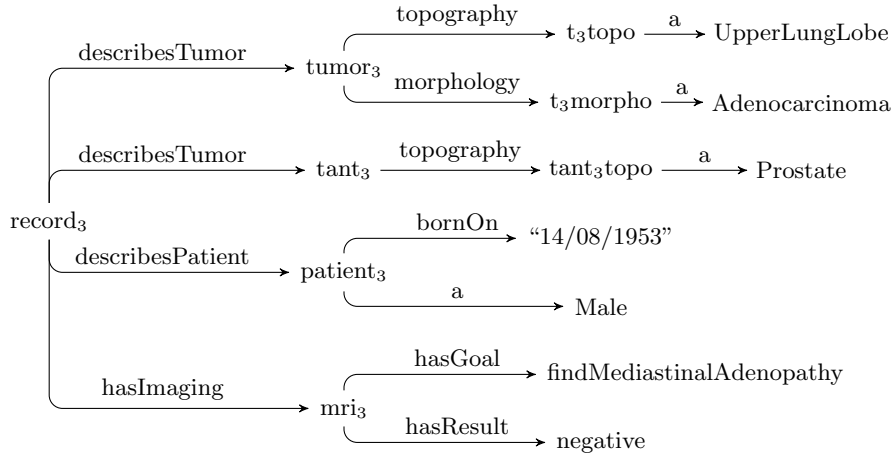

Fig. 3. Partial RDFS graph for the patient record of `srce3`.

The solution contains the answer to the question, an optional textual explanation and the most important arguments in favor of (**pros**) and against (**cons**) the given answer. The optional explanation is provided by the coding experts, and may point out key features or best practices for operators. The arguments have two uses. They will help explain the answer to operators and serve as a reminder for coding experts. They will also be used by the algorithm during the

retrieval step to match the target case with solution cases. In the proposed approach, 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 $\mathtt{a}$ that associates a Boolean to a case and is stored as a SPARQL ASK query. The argument type (i.e., strong or weak and pro or con) is defined by the coding experts, in accordance with the coding standards and best practices.

A partial RDFS graph for $\mathtt{srce}_3$ is shown in figures 2 and 3. One of the pros is that no other synchronous tumor is found. This argument $\mathtt{wp}_3^1$ is formalized as follows:

$\mathtt{wp}_3^1(\mathtt{case}) =$

```
ASK {
  case hasQuestion ?question .
  case concernsRecord ?record .
  ?question concernsTumor ?tumor .
  FILTER NOT EXISTS {
    ?record describesTumor ?other_tumor .
    ?tumor != ?other_tumor .
    ?tumor isSynchronousWith ?other_tumor
  }
}
```

$\mathtt{wp}_3^1$ applies to $\mathtt{srce}_3$ means that $\mathtt{wp}_3^1(\mathtt{srce}_3) = \mathtt{true}$.

4.4 Retrieve

The retrieval of source cases is limited to cases concerning the same subject as the target problem. For the running example, this means that only source cases concerning the nature of the tumor will be taken into account.

To find the most appropriate source case among the selected cases, the arguments will be considered. The arguments are part of the reasoning process which leads the coding experts to the final solution. As such, they can be used to identify similar cases.

Knowing the target problem $\mathtt{tgt}$, retrieval knowledge consists in preferring one source case to another, the preferred source case being the retrieved one. This preference relation is denoted by the preorder $\preccurlyeq_{\mathtt{tgt}}$.

For a given source case $\mathtt{srce}$, let $\mathtt{sp}(\mathtt{srce})$ be the set of its strong pros, $\mathtt{wp}(\mathtt{srce})$ the set of its weak pros and $\mathtt{wc}(\mathtt{srce})$ the set of its weak cons. For $\mathtt{srce}_3$, $\mathtt{sp}(\mathtt{srce}_3) = \emptyset$, $\mathtt{wp}(\mathtt{srce}_3) = \{\mathtt{wp}_3^1, \mathtt{wp}_3^2, \mathtt{wp}_3^3\}$ and $\mathtt{wc}(\mathtt{srce}_3) = \{\mathtt{wc}_3^1\}$.

Let $\mathtt{args} \in \{\mathtt{sp}, \mathtt{wp}, \mathtt{wc}\}$ be an argument type, $\mathcal{N}^{\mathtt{args}}(\mathtt{srce}, \mathtt{tgt})$ denotes the number of arguments of type $\mathtt{args}$ of a the source case $\mathtt{srce}$ which are valid for a case $\mathtt{tgt}$.

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

Table 1 presents the different values of $\mathcal{N}^{\texttt{args}}(\texttt{srce}_\texttt{i}, \texttt{tgt})$ for $\texttt{tgt}$ and the possible source cases $\texttt{srce}_1$, $\texttt{srce}_2$ and $\texttt{srce}_3$. For example, out of the four weak pros of $\texttt{srce}_2$, only one can be applied to $\texttt{tgt}$, thus $\mathcal{N}^{\texttt{wp}}(\texttt{srce}_2, \texttt{tgt}) = 1$.

Table 1. Number of valid source case arguments for the running example ($\mathcal{N}^{\texttt{args}}(\texttt{srce}_\texttt{i}, \texttt{tgt})$) and their distance to $\texttt{tgt}$.

args	sp	wp	wc	$\texttt{dist}(\texttt{srce}_\texttt{i}, \texttt{tgt})$
$\texttt{srce}_1$	0	0	0	3.45
$\texttt{srce}_2$	0	1	0	5.37
$\texttt{srce}_3$	0	3	1	2.51

To compare two source cases $\texttt{srce}_\texttt{i}$ and $\texttt{srce}_\texttt{j}$, three criteria are combined. The first criterion concerns the strong pros and consists in computing:

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

The second criterion concerns the weak pros and cons and consists in computing:

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

where $\lambda_\texttt{p}$ and $\lambda_\texttt{c}$ are two non-negative coefficients that are currently fixed to $\lambda_\texttt{p} = 2$ and $\lambda_\texttt{c} = 1$. When more data are available, these parameter values will be reevaluated. The third criterion concerns the patient record similarity and consists in computing:

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

where $\texttt{dist}$ is a distance function between patient records. $\texttt{dist}$ has been implemented using an edit distance between graphs [6].

These criteria are considered lexicographically, first $\Delta^{\texttt{s}}_{\texttt{i,j}}$, then $\Delta^{\texttt{w}}_{\texttt{i,j}}$ and finally $\Delta^{\texttt{dist}}_{\texttt{i,j}}$, that is $\texttt{srce}_\texttt{i} \preccurlyeq_{\texttt{tgt}} \texttt{srce}_\texttt{j}$ if

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

This means that, for our approach, the criterion based on the strong pros outweighs the one based on the weak pros and cons, which in turn outweighs the criteria based on the patient record similarities. This order has been chosen to match the coding experts' reasoning. For the implemented prototype, several source cases are retrieved, ordered by $\preccurlyeq_{\texttt{tgt}}$ and according to a threshold (which remains to be accurately fixed).

Table 2 shows the values of the various helpers for the running example. None of the strong arguments of the source cases are valid for $\texttt{tgt}$, thus the weak arguments are considered. The comparison shows that $\texttt{srce}_3$ is preferred to $\texttt{srce}_2$ and that both are preferred to $\texttt{srce}_1$.

Table 2. Comparing source cases with respect to the target problem of the running example case (with $\lambda_p = 2$ and $\lambda_c = 1$).

i	j	$\Delta^s_{i,j}$	$\Delta^w_{i,j}$	$\Delta^{dist}_{i,j}$	$\preccurlyeq_{tgt}$
1	2	0	-2	1.88	$\texttt{srce}_2 \preccurlyeq_{\texttt{tgt}} \texttt{srce}_1$
1	3	0	-5	-0.94	$\texttt{srce}_3 \preccurlyeq_{\texttt{tgt}} \texttt{srce}_1$
2	3	0	-3	-2.86	$\texttt{srce}_3 \preccurlyeq_{\texttt{tgt}} \texttt{srce}_2$

4.5 Reuse

Once an appropriate source case has been found, the solution associated to the source case is copied and then the arguments that do not apply to the target problem, if any, are simply removed. This step is repeated for every retrieved source case. For the running example, the most appropriate source case is $\texttt{srce}_3$. The answer for $\texttt{srce}_3$ is to consider the tumor to be of primary nature and thus, for tgt, the answer to the question is also a primary tumor. All the arguments of $(\texttt{srce}_3, \texttt{sol}(\texttt{srce}_3))$ apply, therefore $\texttt{sp}(\texttt{tgt}) = \texttt{sp}(\texttt{srce}_3)$, $\texttt{wp}(\texttt{tgt}) = \texttt{wp}(\texttt{srce}_3)$ and $\texttt{wc}(\texttt{tgt}) = \texttt{wc}(\texttt{srce}_3)$.

4.6 Revise and retain

Currently, the retrieve and reuse steps have been implemented in a prototype described in Section 5. This section presents first thoughts about the revise and retain steps.

Let $(\texttt{tgt}, \texttt{sol}(\texttt{tgt}))$ be the reused case. It may be revised by a coding expert, to modify the answer and/or add, remove or modify some arguments. The expert may also want to remove information from tgt as to keep only the relevant information with regard to the problem-solving process (i.e., the reuse of the arguments). In such a situation, $(\texttt{tgt}, \texttt{sol}(\texttt{tgt}))$ is substituted by $(\texttt{tgt}', \texttt{sol}(\texttt{tgt}'))$, where $\texttt{tgt}'$ is more general than tgt. $(\texttt{tgt}', \texttt{sol}(\texttt{tgt}'))$ is a generalized case that has a larger coverage than $(\texttt{tgt}, \texttt{sol}(\texttt{tgt}))$ [12].

When the system will be in use, the revise step is going to be triggered systematically, at least for the very beginning. Nevertheless, this should unburden the experts, since, hopefully, revising a case will be less time-consuming than solving it.

For now, it is planned to retain all the revised cases. Currently, between 100 and 200 cases per year require expert help. If the case base happens to be too large, a case base management process may be considered [17].

It may occur that the retrieve step fails, if some thresholds are chosen for the retrieval step. For example, it can be considered that for the source case to be retrieved, at least one of its pros has to be applicable to the target problem. In such a situation, the target problem is solved by the coding experts, and thus, the revise and retain steps enrich the case base. This constitutes a case authoring process.

5 A prototype

The prototype is a web application, allowing an operator to ask questions. It tries to solve these questions, using the approach previously described (sections 4.4 and 4.5). The web application is composed of two parts: a form for the description of the target problem (see figure 4) and a presentation the proposed solutions accompanied by a summary of the target problem (see figure 5). In this first implementation, the number of items that can be provided by the user are fixed (e.g., there can only be one tumor, one antecedent and one synchronous tumor) and only a single question subject is possible, namely the tumor's nature.

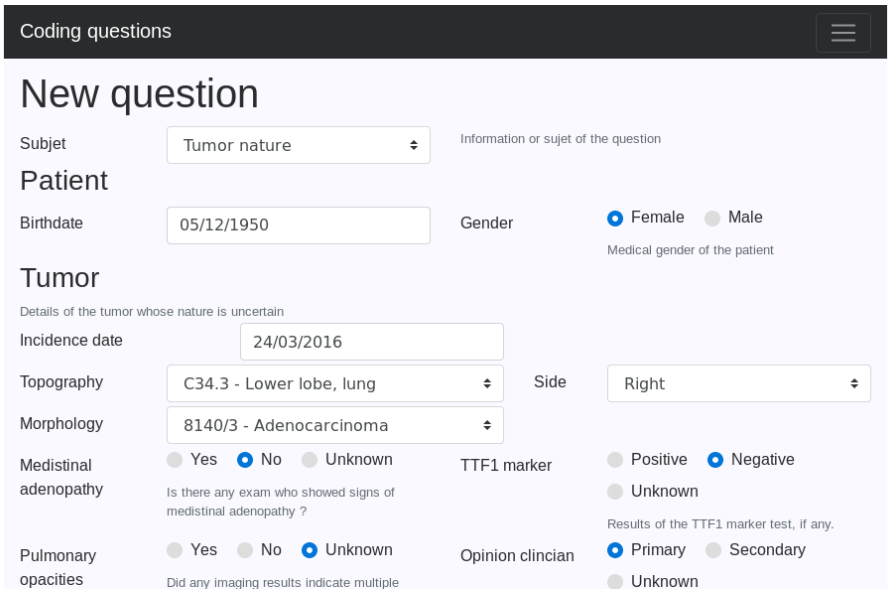

Fig. 4. Form used to describe the target problem of the running example.

6 Discussion and related work

The system described in this paper can be seen as an example of interpretative case-based reasoning. Other approaches in this area include Murdock et al. [14]. Their approach focuses on assisting intelligence analysts in evaluating hypotheses of hostile activities such as take over attempts by criminal groups. The hypothesis (target problem) is matched to a model (source case), which represents a general sequence of events for the given hypothesized event. Then, their system compares the facts from their target hypothesis with those from the model. If a successful match is found, their system relies on this match to generate arguments to justify

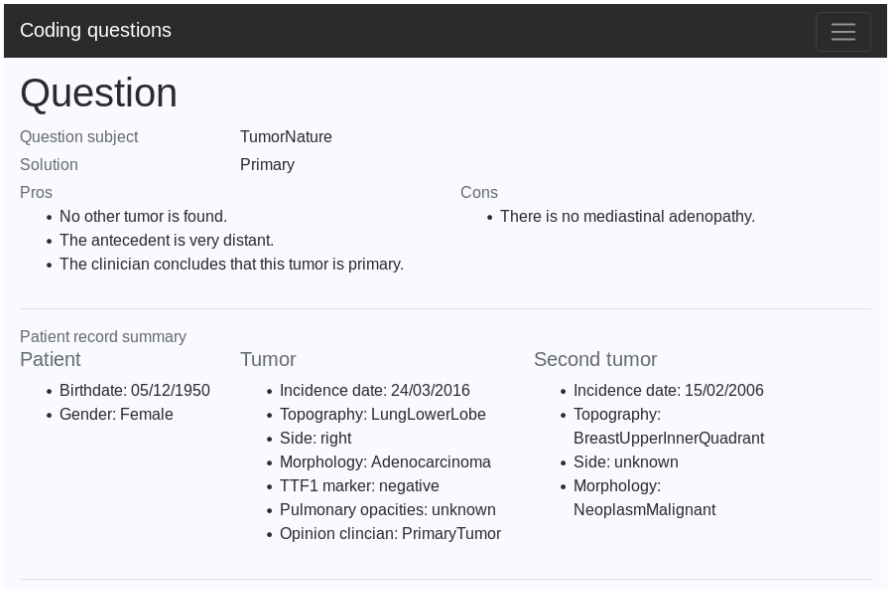

Fig. 5. Summary of the described target problem and the proposed solution. The most appropriate source cases are shown similarly to the target problem (not visible in this screenshot).

or discredit the hypothesis. It is left to the user to decide whether or not the target hypothesis is valid. Contrarily to our approach, the arguments are used solely to explain the proposed solution.

Case-based reasoning has been used a lot in the legal domain (HYPO [3], CATO [2]). Here, source cases are old court cases. The argumentation focuses on the reuse of these precedents, on how similarities can be highlighted and differences downplayed, in order to justify the desired outcome for the target court case. This marks a difference with the approach described in this paper, where arguments are described per case and implicitly linked to the source case.

Particularly in the context of assisting users, explanations are essential, as they provide a measure of understanding for the user and promote the trustworthiness of the system. Similarly to this research, pros and cons are considered by McSherry in [13]. He describes a system for binary classification which uses the closest source case to provide the conclusion and the closest source case with the opposite conclusion to compute which attributes favor the conclusion (pros) and which attributes do not (cons). 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.

In health sciences, case-based reasoning is not the only area that is currently very actively researched [4]. Automatic annotation of medical documents is another such area [15]. While our approach focuses on assisting operators in their

tasks, these approaches seek to replace the need for operators in their current capacity. They focus on analyzing and annotating medical reports [10, 11].

7 Conclusion and future work

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.

This approach has modeled the reasoning processes of the coding experts that were observed. A first prototype has been developed for this purpose and has to be deployed and evaluated (does the system decrease the experts' workload while maintaining the coding quality?).

Currently, the arguments used by our approach remain very simple. As such, they only cover a part of the problem-solving process and resemble hints or highlights. To better represent the solving process, more complex arguments are required. Complex arguments could be combined from simpler arguments using a few operators (e.g. and, or). This should allow for the inclusion of other arguments which, by themselves, do not favor or disfavor a given outcome, but might do so when combined. Furthermore, arguments of a source case are presently reused as such for the target problem. It is planned to examine how these arguments could be adapted to take into account the differences between the source case and the target problem.

Another crucial aspect for the cancer registries is the evolution of the coding practices. Any change in the coding practices will provoke changes in the case base and the associated knowledge containers. It might be interesting to consider methods to detect the needed changes and to help maintain the represented knowledge [7].

When the system is tested, validated, improved and routinely used by operators and experts, a second version of it 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 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. Aleven, V., Ashley, K.D.: Teaching case-based argumentation through a model and examples empirical evaluation of an intelligent learning environment. In: Artificial intelligence in education. vol. 39, pp. 87–94 (1997)
3. Ashley, K.D.: Modeling legal arguments: Reasoning with cases and hypotheticals. MIT press (1991)
4. Bichindaritz, I., Marling, C., Montani, S.: Case-based Reasoning in the Health Sciences. In: Workshop Proceedings of ICCBR (2015)
5. Brickley, D., Guha, R.V.: RDF Schema 1.1, https://www.w3.org/TR/rdf-schema/, W3C recommendation, last consultation: March 2017 (2014)
6. Bunke, H., Messmer, B.T.: Similarity measures for structured representations. In: European Workshop on Case-Based Reasoning. pp. 106–118. Springer (1993)
7. Cardoso, S.D., Pruski, C., Da Silveira, M., Lin, Y.C., Groß, A., Rahm, E., Reynaud-Delaître, C.: Leveraging the Impact of Ontology Evolution on Semantic Annotations, pp. 68–82. Springer International Publishing, Cham (2016)
8. Crammer, K., Dredze, M., Ganchev, K., Talukdar, P.P., Carroll, S.: Automatic code assignment to medical text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. pp. 129–136. Association for Computational Linguistics (2007)
9. 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)
10. Kavuluru, R., Han, S., Harris, D.: Unsupervised extraction of diagnosis codes from emrs using knowledge-based and extractive text summarization techniques. In: Canadian Conference on Artificial Intelligence. pp. 77–88. Springer Berlin Heidelberg (2013)
11. Kavuluru, R., Hands, I., Durbin, E.B., Witt, L.: Automatic Extraction of ICD-O-3 Primary Sites from Cancer Pathology Reports (2013), http://www.ncbi.nlm.nih.gov/pmc/papers/PMC3845766/
12. Maximini, K., Maximini, R., Bergmann, R.: An investigation of generalized cases, pp. 261–275. Springer (2003)
13. McSherry, D.: Explaining the Pros and Cons of Conclusions in CBR. In: European Conference on Case-Based Reasoning. pp. 317–330. Springer (2004)
14. Murdock, J.W., Aha, D.W., Breslow, L.A.: Assessing elaborated hypotheses: An interpretive case-based reasoning approach. In: International Conference on Case-Based Reasoning. pp. 332–346. Springer (2003)
15. Pestian, J.P., Brew, C., Matykiewicz, P., Hovermale, D.J., Johnson, N., Cohen, K.B., Duch, W.: A shared task involving multi-label classification of clinical free text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. pp. 97–104. Association for Computational Linguistics, 1572411 (2007)
16. Richter, M.M., Weber, R.O.: Case-based reasoning: a textbook. Springer Science & Business Media (2013)
17. Smyth, B., Keane, M.T.: Remembering to forget. In: Proceedings of the 14th international joint conference on Artificial intelligence. pp. 377–382. Citeseer (1995)