7

TMR-BASED KNOWLEDGE-DRIVEN PAPER RETRIEVAL
- APPLIED TO UPDATING OF CLINICAL GUIDELINES

Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world.

Albert Einstein

Clinical Guidelines are important knowledge resources for medical decision making. They provide clinical recommendations based on a collection of research findings with respect to a specific disease. Since, new findings are regularly published, CGs are also expected to be regularly updated. However, selecting and analysing medical publications require a huge human efforts, even when these publications are mostly regrouped into repositories (e.g., MEDLINE database) and accessible via a search engine (e.g PubMed). Automatically detecting those research findings from a medical search engine such as PubMed supports the guideline updating process. A simple search method is to select the medical terms that appear in the conclusions of the guideline to generate a query to search for new evidences. However, some challenges rise in this method: how to select the important terms, besides how to consider background knowledge that may be missing or not explicitly stated in those conclusions. In this paper we apply a *knowledge model* that formally describes elements such as actions and their effects to investigate (i) if it favors selecting the medical terms to compose queries and (ii) if/how a search enhanced with background knowledge can provide better result than other methods. This work explores a knowledge-driven approach for detecting new evidences relevant for the clinical guideline update process. Based on the outcomes of two experiments, we found that this approach can improve the recall by retrieving more relevant evidences than previous methods.

This chapter is based on *Zamborlini, V.; Hu, Q.; Huang, Z.; da Silveira, M.; Pruski, C.; ten Teije, A.; van Harmelen, F. "Knowledge-driven Paper Retrieval to support updating of Clinical Guidelines: A use case on PubMed", Lecture Notes in Artificial Intelligence LNAI 10096, 2017 (to appear).*

7.1 INTRODUCTION

Clinical guidelines (CGs) are a collection of best practices, selected based on the latest research findings. Thus, they are expected to be updated regularly and frequently to accommodate new research findings (also known as "evidences") in medicine. However, the quantity of new medical findings published almost every day increases the workload and the complexity of CGs updating tasks. For instance, almost 20.000 new papers (or almost 55 new papers per day) about *breast cancer* were added to PubMed in 2015. Reviewing all these papers to extract relevant information for CGs updating becomes a laborious work. Therefore, automatically detecting relevant papers (i.e. supported with a computer tool) and highlighting the new findings, e.g. taking as reference the PubMed dataset, is considered a relevant approach for supporting the guideline updating process.

In order to find new evidences, the experts need to define queries to be posed against a dataset, e.g. PubMed. Selecting appropriate keywords (e.g., terms from Mesh vocabulary) for building PubMed queries has been done for long time by hand. Previous work [73] had shown that PubMed queries can be constructed based on the conclusion text to find relevant papers. Guideline conclusions are the summaries of a number of medical evidences on which the guideline recommendations are based (e.g. *After a radiation boost the risk of local recurrence of breast cancer is lower*). However, simply considering all medical terms from the conclusions may decrease the performance of the tools (their disjunction generates a huge quantity of papers to review, or their conjunction may exclude relevant papers from the results). In [40, 41], we propose an approach for detecting new and relevant evidences for clinical guideline update, by using a semantic-distance measure for ranking the medical terms extracted from guideline conclusions. The ranked terms are selected to generate PubMed queries used to find the evidences. That semantics-distance-based approach can provide a better result for the search, compared with previous approach [73]. However, we found that some guideline conclusions are weakly linked to their targeted evidences, i.e. the terms used in the conclusions are not enough to retrieve all the relevant evidences. How to obtain the medical terms in order to optimise the performance of the search system remains an open issue.

The goal of this work is on the one hand to investigate if a knowledge-driven approach can favor addressing the aforementioned open issue and, on the other hand, to investigate the applicability of the TMR (Transition-based medical Recommendation) model for supporting the guideline update task. This model is meant for representing knowledge underlying clinical guidelines, and has been applied to address the multimorbidity issue [113]. Therefore, we propose a method for automatically constructing PubMed queries from formal representation of the CGs conclusions based on the TMR model. The method relies on its causation structure, namely actions and effects, for both selecting medical terms and guiding 4 possible logical patterns for bulding the queries. It allows for enriching the original terms provided in the conclusions with alternative descriptions.

We have conducted two experiments on the Dutch Clinical Guidelines of Breast Cancer. We formalize some conclusions from the older version, submit the automatically generated queries to PubMed and verify if we find at least the list of publications referenced by the corresponding conclusion in the new version of the CG. This list of publications references is named in this paper as *goal evidences*. We analyse the results and compare to our previous approach.

This paper is organized as follows. The fully automatic approach previously mentioned, the semantic-distance method, is presented in Sect. 7.2. Sections 7.3 to 7.4 describe the TMR model applied for guideline update and the method for automatic query construction based on this model. Section 7.5 reports the experiments and results. Discussion and conclusion are presented in Sec. 7.6 and 7.7.

7.2 SEMANTIC-DISTANCE METHOD FOR GUIDELINE UPDATE

Evidence-based clinical guidelines rely on published scientific research findings. Those publications are usually available in Web-based biomedical databases, like PubMed. In [40, 41], we proposed a semantic distance method for automatic detecting new evidences for guideline update. The main challenge consists in how to select relevant medical terms to compose a query for retrieving new medical evidences from datasets such as PubMed. Indeed, simply using all the terms as conjunction or disjunction has been shown not to be effective. The semantic distance measure is based on the (widely shared) assumption that the more frequently co-occurring terms exist the more semantically related they are. Based on this measure the terms are ranked and selected via an heuristic function developed based on three criteria: term coverage, evidence coverage and bounding number. We have reported several experiments to evaluate this approach. We selected the Dutch breast cancer guidelines version 1.0, 2004 [63] and version 2.0, 2012 [64] as the test data.

Figure 7.1 depicts a workflow of the general idea applied in the aforementioned experiments. Firstly, the original guideline [63] is the input for automatic extraction of a list of relevant medical terms. To do so a NLP tool is applied to the text of each guideline conclusion and the heading of the sections or subsections, producing a list of medical terms according to UMLS vocabulary. Then, the semantic distance is calculated for each term by checking its occurrence in PubMed articles. The terms are ranked and the most important ones are selected to compose the term-queries to be submitted to PubMed. From the retrieved publications, it is verified if the expected goals per conclusion are achieved (based on [64]), and the respective recall and precision are calculated.

From these experiments, we concluded that improvements are still needed in order to reduce the size of the obtained results and to find more goal evidence for more guideline items. We observed that some guideline conclusions are weakly linked to their targeted evidences, i.e. in some cases the exact terms used in the conclusions are not enough to retrieve all the relevant evidences.

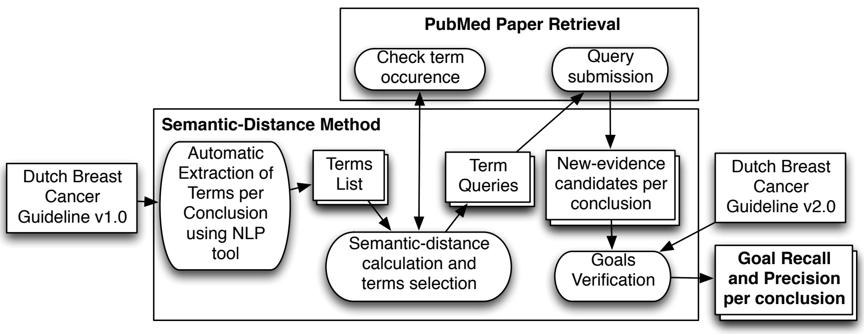

Figure 7.1: Experiments' workflow for appling Semantic Distance approach for the Dutch Breast Cancer Guidelines.

7.3 TMR MODEL APPLIED FOR GUIDELINE UPDATE

The TMR model has been developed for representing knowledge underlying CGs, aiming at supporting different guideline-related tasks. In our previous work [112, 113] we addressed the task of combining several guidelines (multimorbidity) by automatically detecting interactions among recommendations. In this paper we will: (i) check the applicability of the TMR model on supporting the update task and (ii) identify improvements required to better address this task.

Figure 7.2 presents an excerpt of the UML class diagram of the TMR model, including at the bottom an example of instantiation. The model addresses the guideline as source of medical conclusions, which are beliefs about a causation relation[1]. In the example, "*After a **boost** the **risk of local recurrence** is lower*" is a conclusion/belief (*CB#1*) from the Dutch Breast Cancer Guideline about the causation relation between two event types: the action *boost (of radiation dose)* often causes the transition (*Tr#1*) of decreasing the property *risk of local recurrence (of Breast Cancer)*[2].

Therefore, *causation beliefs* refer to the causation relation between two *event types*; particularly between an *action type* and a *transition type*. A belief has a *source* and a (causation) *frequency*. The belief can be simple as in the example, or it can be composed, having other beliefs as its *part* (*hasPart*). The relationship *causes* is derived from a causation belief (or exists only in its context) and therefore is depicted as a dotted line. A transition type *affects* a property, referred as *trope type* in the model, according to its *derivative* (increase, de-

1 The relation to the evidences will be addressed in future work.
2 This is our interpretation, as the conclusion itself is not very precise.

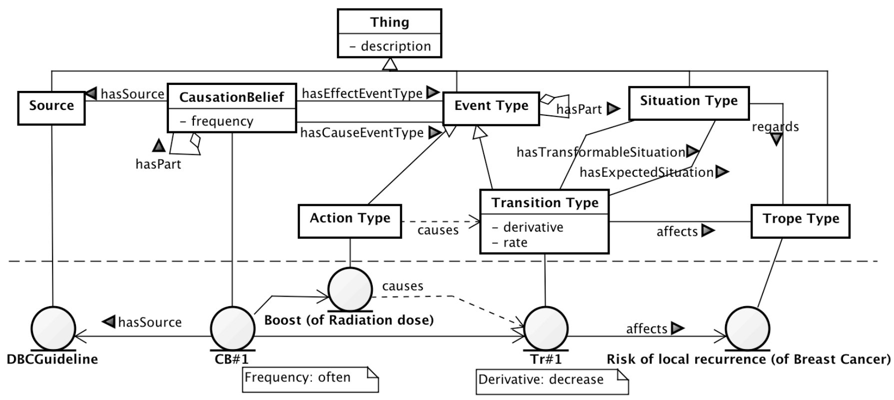

Figure 7.2: UML class diagram for an excerpt of the TMR Model.

crease, maintain). Another way to define a transition is by providing the *transformable* and *expected Situation Types*. An event type can also be composed, having other ones as its *part* (*hasPart*). All the elements can have *description*(s). In particular, the composition of either beliefs or event types are added in this work, as it was not required in previous work for applying the TMR model[3].

We advocate that this excerpt of the TMR model contains essential information for searching for new evidences, namely **action and effect** (sometimes the expected situation has a description, like *fever*, sometimes it is the property affected by the transition, like *temperature*). It allows to offer four options for how to query for papers by varying the logic composition of actions and effects, which we call query patterns. The first two patterns are simple consultation for publications regarding either actions or effects, and the last two are conjunction and disjunction combinations of the previous ones.

query patterns:

QP1 - **Action**: this pattern allows for retrieving any publication referring to the action in a conclusion. For example, to retrieve papers about new/unknown effects for that action (e.g. *Silicon Implant*);

QP2 - **Effect**: this pattern allows for retrieving any publication referring to the effect in a conclusion. For example, to retrieve papers about new/unknown actions that produce or interfere in that effect (e.g. *Systemic Syndrome*);

QP3 - **Action AND Effect**: this is a more strict pattern that allows for retrieving publications referring simultaneously to both action and effect (e.g. *(Silicon Implant)+AND+(Systemic Syndrome)*);

QP4 - **Action OR Effect**: this pattern accumulates the results of 1 and 2 (e.g. *(Silicon Implant)+OR+(Systemic Syndrome)*).

The conclusions might refer to more than one action or effect, as well as multiple descriptions for the same element can be found in the guideline itself or in external vocabularies. In the sequel we present and justify other logic compositions for each of these cases:

other logic compositions:

LC1: If a composed action is described in a conclusion, it is addressed as a conjunction of actions (*(Mastectomy)+AND+(Breast Reconstruction)*);

3 For sake of readability, hereafter we omit mentioning "type" for the model elements.

LC2: If more than one effect is described we consider the result relevant if it contains any of the effects, thus it is addressed as a disjunction (*(Aesthetic Result)+OR+(Psychological Well-being)*).

LC3: If multiple descriptions are provided to the same element, they are addressed as disjunction (*(Breast Reconstruction)+OR+(Mamaplasty)*);

Finally, we acknowledge that not all the knowledge within guideline conclusions are currently covered by the TMR model. Taking examples extracted from the Dutch Breast Cancer Guideline, (i) the conclusion might contain comparison of the outcome of a certain intervention with other ones or with no intervention (e.g. in Table 7.1 conclusion 5_1); or (ii) it might describe risk factors (e.g. "*young age ($\leqslant$ 40 years) is an independent (negative) risk factor for the development of local recurrence after BCT (Breast-Conserving Therapy).*"). Those topics and their role on searching new evidences will be addressed in future work.

7.4 EXTRACTING PUBMED-QUERIES FROM TMR MODEL

We propose a method for investigating the applicability of the TMR model and in sequence we discuss the implementation and the design choices required to this end. Finally, SPARQL queries are provided for automatically extracting the descriptions that are relevant for constructing a search query[4]. For the experiment we chose to apply query-pattern QP3 (action AND effect) to have the results comparable to [40, 41].

7.4.1 *Experiment Method*

In the previous approach [40, 41] the main challenge was to select and combine from the guideline conclusions the important terms to construct the PubMed queries (see Sect. 7.2). In order to understand the improvements that can be achieved by using the TMR model, we propose an experiment in two parts:

TMR-STRATEGY 1 - **Without intervention on conclusion content**: only the original description text is used for querying. It relies on synonyms provided by PubMed according to the Mesh terminology.

4 Codes are available at `https://github.com/veruskacz/KR4HC2016`.

TMR-STRATEGY 2 - **With human intervention on conclusion content**: new (alternative) descriptions are added. It assumes that: (i) relevant information is often made implicit in the conclusion; (ii) Mesh-based strategy embedded in PubMed query service provides limited options of synonyms; (iii) sometimes other related terms rather than synonyms are important (e.g. more general or more specific terms). The acquisition of such descriptions is further discussed in this paper.

Firstly, by applying TMR-STRATEGY 1 (or simply TMR-1), we investigate how the TMR knowledge-structure, particularly actions and effects, contributes for selecting important terms from a guideline conclusion. Then, given the obtained results, we analyse the missing goals and understand the reasons why they cannot be retrieved. Finally, by applying the TMR-STRATEGY 2 (or simply TMR-2), we investigate if/how the TMR model allows for the enrichment required in order to retrieve the missing goals. Observe that our claim for strategy 2 is *NOT* that now we can retrieve *ALL* the goals, since by knowing the goals we can fine tune the queries until we get what we want. The claim is that a knowledge-driven approach such as the one proposed here provides means to perform the enrichment required to improve the results in this but also in other experiments.

Figure 7.3 depicts at the top the workflow for applying the TMR-Strategy 1 to the Dutch Breast Cancer Guideline, and at the bottom the workflow for TMR-Strategy 2 (the experiment is reported in Sect. 7.5). In the former, the gray-shaded components are the same as in the Semantic Distance workflow (Fig. 7.1). In other words, the main change with respect to the first strategy regards the way the original guideline is processed into term-queries. The original guideline (version 1.0) is the input for the manual process of modeling the guideline conclusions according to the TMR model. The resultant structured guideline is then the input for the automatic construction of the term-queries by applying the SPARQL queries. The final output is a list of goals' recall and precision per conclusion, together with the missing goals per conclusion. The latter will be input for the next strategy, as well as the structured guideline, both marked with a star (*).

At the bottom part of Fig. 7.3, the missing goals are the input for the manual process of analyzing the terms that were missing per conclusion in order to retrieve the missing goals, according to the MESH terms used to annotate them in PubMed or the papers' titles. For

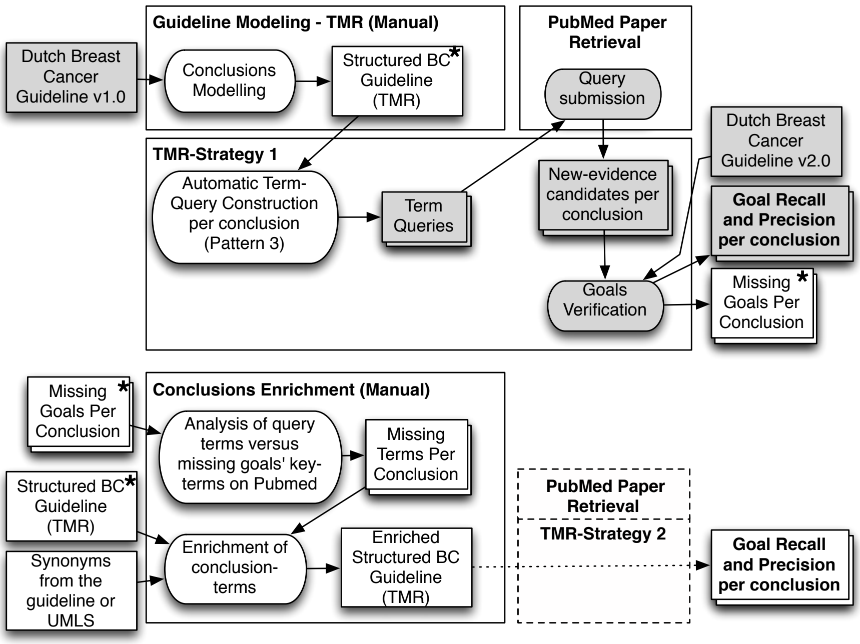

Figure 7.3: Experiments' workflows for applying TRM-strategies 1 and 2.

the experiment, we also manually consulted UMLS vocabulary for standard descriptions as alternative for the original terms, although not everything could be (trivially) found. Moreover, the guideline itself provides alternative descriptions for a same element in several conclusions. Finally, the missing terms, together with the structured guideline and a list of synonyms (from UMLS and the guideline), is input for the process of enriching the descriptions of the conclusions terms. The enriched structured guideline is then input for applying the TMR-Strategy 2. Details are omitted for both 'dotted boxes' PubMed Paper Retrieval and TMR-Strategy 2, since they repeat the corresponding boxes in the top part. In other words, the main change in the workflow for the second experiment is the process of enriching the input for the TMR-Strategy. The final output is a new list of goals' recall and precision per conclusion.

7.4.2 *Formalisation*

This is an additional section, not included in the original paper, meant to provide a technology-independent solution, while the technology-specific version is provided in the next Section. It presents algorithms as semi-formal descriptions of how to compose a term-based query according to the proposed approach.

First, Algorithm 1 presents the selection of the composition-process based on the 4 query patterns (Sect. 7.3). Next, Algorithm 2 presents the composition of 'Effects' described in a conclusion, according to the Logic Composition LC2. In its turn, Algorithm 3 describes how alternative descriptions are included in the query, according to LC3. Finally, the algorithms for the composition regarding the 'Action' are available in Appendix B.

Algorithm 1 has as input a set C of guideline conclusions, a query pattern qp and a list altList of relations to be used for alternative descriptions. The output is a result set R of tuples with conclusion's id and the term-based query to be submitted to the PubMed API.

We will illustrate the application of the algorithm taking a simplified excerpt of the experiment further described using conclusions from the Dutch Breast Cancer Guideline (Sect. 7.5). So the query pattern chosen is $qp = 3$. i.e. a conjunction of action and effects' descriptions and the set of conclusion is $C = \{5\_1\}$, from Table 7.5.1:

Algorithm 1 : Algorithm for composing term-based queries provided guideline conclusions according to the TMR model, a query pattern reference and a list of relations used for acquiring alternative descriptions.

```
input  : a set C of TMR-conclusions
         a query pattern reference qp ∈ [1..4]
         a list altList of relations used for alternative descriptions
output : a result set R of tuples with conclusion's id and term-based query
begin
    R ← ∅                                          /* intialize the result set */
    C′ ← all identifiable conclusions c ∈ C or subconclusions c_i related to c via
      hasPart
    for c′ ∈ C′ do                                 /* for each conclusion of set C′ */
        /* an identifiable conclusion can regard only one action type (can be
        a composed action), but serveral effects caused by that action    */
        switch qp do
            case 1 do Q ← composeAction(c′,altList)
            case 2 do Q ← composeEffects(c′,altList)
            case 3 do Q ← composeAction(c′,altList) + 'AND' +
              composeEffects(c′,altList)
            case 4 do Q ← '(' + composeAction(c′,altList) + ') OR (' +
              composeEffects(c′,altList) + ')'
        R ← R + [(c′.id, Q)]
    return R
```

> **5_1**: A descriptive study found that women who undergo **breast reconstruction immediately following the mastectomy** are more satisfied with the **aesthetic result** and experience greater **psychosocial wellbeing**.

Each conclusion (causation belief) in a guideline should be atomic with respect to the action type, i.e. it should regard only one action (atomic or composed) that causes one or more described effects. We provide these conclusions with an identifier, thus they are said identifiable conclusions. In the case of conclusion 5_1, since it comprises several effects for one action (e.g. *Action A causes effects E1 and E2*), it is divided into several sub-conclusions/beliefs (without an identifier) about that action causing each of the referred effects.

Finally, we will consider two possible values for altList: *empty set* means only the original descriptions will be used, and altList = { *relatedTo* } means a very general relation among terms will be used for acquiring alternative descriptions from external knowledge sources. Therefore, for applying the algorithms we will consider the following data available: conclusion 5_1, divided in two sub-conclusions, and two statements from external sources. We allow the data to be simplified, as a more detailed example is provided later on this section.

> **5_1**:
> (*) **breast reconstruction immediately following the mastectomy** *(:Action Type)* affects **aesthetic result** *(:Trope Type)*
> (*) **breast reconstruction immediately following the mastectomy** *(:Action Type)* affects **psychosocial wellbeing** *(:Trope Type)*.
>
> (1) **aesthetic result** *(:Trope Type)* is related to **body image** *(:Trope Type)*.
>
> (2) **aesthetic result** *(:Trope Type)* is related to **breast appearance** *(:Trope Type)*.

The outputs for each option provided for altList are:

> **OUTPUT1**: for altList = ∅
> (5_1, *'(breast reconstruction immediately following the mastectomy) + AND + ((aesthetic result) + OR + (psychosocial wellbeing))'*)
>
> **OUTPUT2**: for altList = { *relatedTo* }
>
> (5_1, *'(breast reconstruction immediately following the mastectomy) + AND + (((aesthetic result) + OR + (body image) + OR + (breast appearance)) + OR + (psychosocial wellbeing))'*)

Notice that in practice some conclusions do regard different actions causing an effect (e.g. *Action A or action B causes effect E*). In this case we split it into one identifiable conclusion per action (e.g. the conclusion 6_1.1 and 6_1.2 in Table 7.5.1 are the division of one conclusion into two identifiable ones).

In its turn, Algorithm 2 describes the function *composeEffects* that has two inputs: a conclusion c (for which a query is being composed) and a list altList. The output is a string containing the descriptions provided to the effects in a certain conclusion c or its sub-conclusions. As previously discussed in Sect. 7.3, besides using the description provided to the transition itself (e.g. *fainting*), if any, the conclusion may describe instead the affected trope type, i.e. the property (*body temperature*), or the situation type to be achieved/avoided (e.g. *fever*). To address this issue, the corresponding relations linking a property and situation to a transition (*affects* and *hasExpectedSituation*) are added to altList, which means the descriptions provided to those elements will also be included, together with any other alternative relation provided.

Considering the input is conclusion 5_1, for each of the options for altList the following output is obtained:

OUTPUT1: for altList $= \emptyset$
'((aesthetic result) + OR + (psychosocial wellbeing))'

OUTPUT2: for altList $=$ { *relatedTo* }
'(((aesthetic result)+OR+(body image)+OR+(breast appearance))+OR+(psychosocial wellbeing))'

Finally, Algorithm 3 describes the function *composeDescriptions* that has two inputs: an event type e (action or transition) and a list altList.

Algorithm 2 : *composeEffects* - Algorithm for composing the descriptions provided to the effects in a conclusion, according to the proposed Logic Composition LC2

```
input  : a conclusion c
         a list altList of relations used for alternative descriptions
output : query Q for descriptions of effects regarded by c
begin
    Q ← ∅
    E ← all EventTypes related via hasEffectEventType to c or its (non-identifiable)
      subconclusions
    altList ← {affects, hasExpectedSituation} + altList        /* Add the relations
      regarding the affected property and post-situation */
    for e ∈ E do                                    /* LC2:'OR' for several effects */
        if Q ≠ ∅ then
            Q ← Q + ' OR '
        Q ← Q + '(' + composeDescriptions(e,altList) + ')'
    return Q
```

The output is a string containing the description of the provided event type, if any, concatenated (as disjunction) to the descriptions of the elements associated to the event type via the relations provided in `altList`.

Considering the input is one of the (unnamed) transition types from conclusion 5_1, which affects the trope type *aesthetic result*, and that for each of the options for `altList` it was added {*affects*, *hasExpectedSituation*}, the following output is obtained:

> **OUTPUT1**: for `altList` = {*affects*, *hasExpectedSituation*}
> *'(aesthetic result)'*
>
> **OUTPUT2**: for `altList` = {*affects*, *hasExpectedSituation*, *relatedTo*}
> *'(aesthetic result) + OR + (body image) + OR + (breast appearance)'*

The Listing 7.1 in Sect. 7.4.4 presents a single SPARQL code that implements the three algorithms aforementioned. For each of the identifiable conclusions it concatenates as disjunction (LC2) the descriptions provided for their effects/transitions, which in turn are a concatenation as disjunction (LC3) of their alternative descriptions. However, instead of a list of relations for alternative descriptions, the code uses a fixed set of relations. It also does not contain the selection according to the query pattern, as it is specific for effects and it assumes the query pattern is applied by the application.

Algorithm 3 : *composeDescriptions* - Algorithm for composing descriptions provided to a certain event type, including alternative descriptions, according to the proposed Logic Composition LC3

```
input  : an event type e
         a list altList of relations used for alternative descriptions
output : query Q with descriptions related to e, assuming there will be always at least
         one description
begin
    if e.description ≠ ∅ then
        Q ← '('+e.description + ')'                    /* event's description */
    else
        Q ← ∅
    A ← all elements associated to e via relations in altList
    for alt ∈ A do                                 /* LC3:'OR' for alt. descriptions */
        if alt.description ≠ ∅ then
            if Q ≠ ∅ then
                Q ← Q + ' OR '
            Q ← Q + '(' + alt.description + ')'
    return Q
```

```
SELECT ?id (GROUP_CONCAT(?transitionLbl; separator="+OR+") AS ?term)
WHERE {
?conclusion     rdf:type            tmr:CausationBelief .
?conclusion     tmr:conclusionID    ?id
{ GRAPH ?conclusion {    #effect is referred by the main conclusion
[]              tmr:causes          ?transition }
} UNION { ?conclusion   tmr:hasPart  ?subConclusion.
GRAPH ?subConclusion {   #effect is referred by the sub-conclusion
[]              tmr:causes          ?transition } .
{ SELECT ?transition (CONCAT("(",GROUP_CONCAT(CONCAT("(",?label,")");
separator="+OR+"), ")") AS ?transitionLbl)
WHERE {
{?transition tmr:affects ?resource}
UNION {?transition tmr:hasExpectedSituation ?resource}.
 Opt1  ?resource rdfs:label  ?label.
 Opt2  {?resource tmr:interpretedAs ?altResource}
 Opt2   UNION {?resource tmr:relatedTo ?altResource}.
 Opt2  ?altResource rdfs:label ?label
} GROUP BY ?transition }
} GROUP BY ?id
```

Listing 7.1: SPARQL code for selecting the original text (Option 1) or alternative text (Option 2) for **effects** within each conclusion, grouped by the conclusion identifier.

7.4.3 *Implementation and Design choices*

The TMR model is implemented using semantic web technologies (for more details see [113]). In order to allow for distinguishing original and alternative descriptions of guideline conclusions, some design choices are made: (i) the elements originally extracted from the conclusions are represented as RDF resources having as description the string used in the original text, via the predicate rdfs:label (e.g. Sect. 7.5 Listing 7.2 lines 18-20 and 26-28); (ii) then, the human-supervised alternative descriptions are represented as separated resources, which may contain several descriptions via the predicate rdfs:label and to which the original resources are linked via either tmr:relatedTo or tmr:interpretedAs predicates (e.g. Sect. 7.5 Listing 7.3). The latter is a special case of the former for addressing synonyms. The tmr:relatedTo predicate is intentionally broad, and other special cases, such as hierarchy, will be investigated in future work. Moreover, the same resource can be reused as alternative descriptions for several original ones.

7.4.4 *SPARQL queries*

The Listing 7.1 presents a SPARQL query developed for retrieving and composing the terms corresponding to effect for querying PubMed. It offers two options for the TMR-strategies: **(Opt 1)** uses the *label* describing the *resource* (affected property or expected situation) originally represented as part of the conclusion (or sub-conclusions such as for conclusion 6_1 in Sect. 7.5.1); **(Opt 2)** uses the *labels* of alternative *resources* linked to the original ones via relations *interpretedAs* or *relatedTo*. The codes, available online, use one or the other according to the experiment. When more than one label is available, they are concatenated as disjunction (OR) within the inner 'select-clause' grouped by transition (effect), corresponding to the logic composition LC3, i.e. more than one description per element is composed as disjunction (the same holds for the code that composes the actions). The labels also need to be enclosed between parenthesis to ensure the associative properties (e.g. '(local tumor) OR (survival benefit)'). Finally the outer 'select-clause' concatenates the results of the inner 'select-clause', in this case as a disjunction (OR) grouped by conclusion ID, i.e. if more than one effect is described within an identified conclusion they are combined as disjunction according to the logic composition LC2 (for actions they would be composed as conjunction according to LC1). More examples ares in Sect. 7.5.1. The SPARQL query for actions, available online, is slightly more complex to handle the of composition of actions.

7.5 EXPERIMENTS

This section reports on the experiments performed to evaluate the use of TMR model on supporting the guideline update task. We perform the retrieval of medical papers from PubMed that are relevant for updating (part of) the Dutch Breast Cancer Guideline of 2004 [63]. We evaluate the results according to a set of goal evidences used in the updated version of that guideline in 2012 [64].

7.5.1 *Breast Cancer Guideline*

Table 7.1 presents a set of conclusions extracted from the Dutch Breast Cancer guideline of 2004. The original conclusions are manually en-

Table 7.1: Set of conclusions from the Dutch Breast Cancer guideline of 2004, with highlighted terms corresponding to actions and effect.

ID	Conclusion Text	Action	Effect
1_1	Addition of **radiotherapy** following **local excision of DCIS** results in a significantly lower **risk of local recurrence** (this is valid for all subgroups).	radiotherapy **AND** local excision of DCIS	risk of local recurrence
1_3	Adjuvant **therapy with tamoxifen** in **breast-conserving treatment of DCIS**, results in limited improvement of **local tumour** control and no **survival benefit**.	therapy with tamoxifen **AND** breast-conserving treatment of DCIS	local tumour **OR** survival benefit
3_1	**Breast-conserving therapy** including **irradiation** is safe because the **survival rate** is comparable to that seen after modified radical mastectomy.	breast-conserving therapy **AND** irradiation	survival rate
3_2	An excellent **cosmetic result** can be achieved in at least 70% of patients after **breast-conserving therapy**.	breast-conserving therapy	cosmetic result
3_3	After a **boost** the **risk of local recurrence** is lower.	boost	risk of local recurrence
4_1	**Postoperative locoregional radiotherapy** reduces the **risk of locoregional recurrence** by two-thirds, and results in a better **chance of survival**.	postoperative locoregional radiotherapy	risk of locoregional recurrence **OR** chance of survival
5_1	A descriptive study found that women who undergo **breast reconstruction** immediately following the **mastectomy** are more satisfied with the **aesthetic result** and experience greater **psychosocial wellbeing**	breast reconstruction **AND** mastectomy	aesthetic result **OR** psychosocial wellbeing
6_1.1	There are no signs that either **primary** or **secondary breast reconstruction** results in a higher **risk of recurrent breast cancer**.	primary breast reconstruction	risk of recurrent breast cancer
6_1.2		secondary breast reconstruction	risk of recurrent breast cancer
6_2	There are no indications to suggest that a **skin-sparing mastectomy** followed by **immediate reconstruction** leads to a higher **risk of local** or **systemic recurrence of breast cancer**.	skin-sparing mastectomy **AND** immediate reconstruction	risk of local recurrence of breast cancer **OR** risk of systemic recurrence of breast cancer
7_1	There is no causal relationship between **silicone implants** and the occurrence of **systemic syndromes**.	silicone implants	systemic syndromes
8_1	**Radiotherapy** is associated with significantly more **complications** in the presence of a **breast reconstruction**.	radiotherapy **AND** breast reconstruction	complications

coded in RDF according to the TMR model[5], with the following restrictions (see Sect. 7.3): (i) from the 16 conclusions updated in the

5 RDF Data is available at https://github.com/veruskacz/KR4HC2016

new version of the guideline, 11 conclusions were selected that convey actions and their effects (rather than risk factor); and (ii) only the parts of the text regarding the main causation structure are represented. Other parts, as previously mentioned, will be addressed in future work. Another observation is that conclusion 6_1 is divided into two conclusions according to the TMR model, to which we refer as 6_1.1 and 6_1.2. This is because the two actions mentioned in the text are not a composed action that causes the referred effect, but as two alternative actions causing the same effect. Moreover, even when effects (or affected properties) are mentioned as conjunction, such as in conclusion 1_3 (*local tumor control AND no survival benefit*), saying that both effects are expected, we are interested to retrieve papers that says something about any of them. Actually, these conclusions can be the summary of two (or more) different papers talking about one or another effect. Therefore we combine them as disjunction (OR) according to LC2.

The Listing 7.2 presents the RDF representation of (part of) the conclusion 5_1 according to the TMR model. Based on this representation, for each conclusion, the SPARQL queries for strategy 1 are applied and combined to obtain the query pattern QP3 (conjunction of action and effect). The following is obtained for conclusion 5_1[6]:

(breast reconstruction + mastectomy) +AND+ ((aesthetic result) +OR+ (psychosocial wellbeing))

However, alternative descriptions can be provided to the original terms. For example, the original composed action, described as "*Breast reconstruction immediately after mastectomy*", can be (re)interpreted as one action named "*primary breast reconstruction*" according to the vocabulary used in other conclusions of the same guideline.

The Listing 7.3 presents the RDF representation of the aforementioned alternative description according to the TMR model: the resource corresponding to the composed action extracted from the guidelines, namely data:ActBC2004-BreastReconstructionImediatellyAfterMastectomy is linked to another resource data:ActPrimaryBreastReconstruction via predicate tmr:interpretedAs. The latter resource, in turn, has the alternative description defined via rdfs:label predicate. For each conclusion, some alternative descriptions are provided based on: similar descriptions in other conclusions, descriptions provided by UMLS vocabu-

6 When the logic operator is omitted (*breast reconstruction + mastectomy*) it is equivalent to a conjunction (*breast reconstruction + AND + mastectomy*).

```
#Conclusion 5_1 - divided in two sub-conclusions
data:CB-PrimaryBreastReconstruction-SatisfactionAppearance-WellBeing{
 data:CB-PrimaryBreastReconstruction-SatisfactionAppearance-WellBeing
    rdf:type          tmr:CausationBelief ;
    tmr:conclusionID  "5_1";
    tmr:hasPart        data:CB-PrimaryBC-SatAppearance,
                       data:CB-PrimaryBC-WellBeing.}
# Sub-conclusions about causing higher satisfaction with appearance
data:CB-PrimaryBC-SatAppearance {
 data:CB-PrimaryBC-SatAppearance
    rdf:type   tmr:CausationBelief.
 data:ActBC2004-BreastReconstructionImediatellyAfterMastectomy
    tmr:causes data:TrBC2004-HigherSatisfactionWithBreastAppearance.}
# Composed extracted action
data:ActBC2004-BreastReconstructionImediatellyAfterMastectomy
    rdf:type     tmr:ActionType, tmr:ComposedEvent;
    tmr:hasPart  data:ActBC2004-BreastReconstruction,
                 data:ActBC2004-Mastectomy.
# One of the sub-actions
data:ActBC2004-BreastReconstruction
    rdf:type     tmr:ActionType ;
    rdfs:label   "breast reconstruction"@en .
# One of the transtions (effect)
data:TrBC2004-HigherSatisfactionWithBreastAppearance
    rdf:type     tmr:TransitionType ;
    tmr:affects  data:TropeBC2004-AestheticResult.
# Property affected by the transition
data:TropeBC2004-AestheticResult
    rdf:type     tmr:TropeType;
    rdfs:label   "aesthetic result"@en.
```

Listing 7.2: RDF code for conclusion 5_1 from Dutch Breast Cancer Guideline.

```
# Link for original action resource with the reinterpreted one
data:ActBC2004-BreastReconstructionImediatellyAfterMastectomy
    tmr:interpretedAs  data:ActPrimaryBreastReconstruction.
# Reinterpretation for the aforementioned extracted action
data:ActPrimaryBreastReconstruction
    rdf:type     tmr:TropeType;
    rdfs:label   "Primary Breast Reconstruction"@en .
```

Listing 7.3: RDF code for enriching terms from conclusion 5_1.

lary, and finally some were based on the Mesh description obtained in the analysis.

Again, by applying the SPARQL queries for strategy 2 according to query pattern QP3, and given a number of alternative descriptions, the following query is obtained for conclusion 5_1:

(((Breast Reconstruction + Mastectomy) + OR + (Primary Breast Reconstruction) + OR + (Immediate Breast Reconstruction))) + AND + ((Depression) + OR + (Psychosocial Wellbeing) + OR + (Esthetics) + OR + (Body Image) + OR + (Cosmetic Result of Breast) + OR + (Breast Aesthetics) + OR + (Breast Appearance) + OR + (Quality Breast Appearance))

The results are presented and discussed in the next section.

7.5.2 *Results & Analysis*

The results for both TMR-strategies 1 and 2 are presented in Table 7.2, together with the results obtained in previous text-based experiments [41]. For each conclusion we present the recall based on the goal evidence list (2/5 means 2 out of 5). The best recall results compared to the strategy right before are highlighted in the table (as Semantic Distance is the first, we highlight the better results compared to the TMR-1). In particular, some of the conclusions for which no results were obtained by using the Semantic Distance approaches are highlighted in bold (namely, 3_1, 3_3, 5_1, 7_1). Precision is not calculated since the goal list does not comprise all the possibly relevant papers but the ones indeed used in the updated version of the guideline. Calculating the precision requires an expert to evaluate the relevance of all retrieved papers. Instead, we present the total number of papers retrieved (#Papers), and for the purpose of helping comparing the approaches, we compute a 'cost' regarding number of retrieved papers per hitted goal (when the number of hitted goals is zero, we multiply the number of retrieved papers per two as a simplification, considering a division per zero would be infinite cost). The average recall increases from 32.6% (Sem.Dist.) to 35.8% (TMR-1) and to 80.9%

Table 7.2: Results obtained for experiment according to the method here proposed (strategies 1 and 2) and the previous Topic-Centric method described in Sect. 7.2.

Concl.	**Semantic Distance**			**TMR-Strategy 1**			**TMR-Strategy 2**		
	Recall	#Papers	Cost	Recall	#Papers	Cost	Recall	#Papers	Cost
1_1	2/5	60	30	1/5	110	110	3/5	2153	718
1_3	1/4	36	36	2/4	36	18	4/4	582	145
3_1	**0/14**	49	98	0/14	555	1110	**8/14**	1957	245
3_2	1/2	28	28	1/2	126	126	2/2	688	344
3_3	**0/2**	33	66	**1/2**	424	424	2/2	1418	709
4_1	3/5	1628	543	2/5	276	138	2/5	155	77
5_1	**0/3**	82	164	0/3	231	462	2/3	1130	565
6_1.1	5/5	9911	1982	0/5	12	24	5/5	540	108
6_1.2				0/5	8	16	2/5	298	149
6_2	1/3	72	72	1/3	53	53	2/3	427	213
7_1	**0/2**	372	744	**1/2**	48	48	2/2	97	48
8_1	1/2	324	324	2/2	535	267	2/2	295	147
AVG.	**32.6%**	**1145.18**	**372**	**35.8%**	**221.18**	**254**	**80.9%**	**885.45**	**297**

Figure 7.4: Left hand side shows the recall for each approach grouped per conclusion. Right hand side depicts the same results in ascending order.

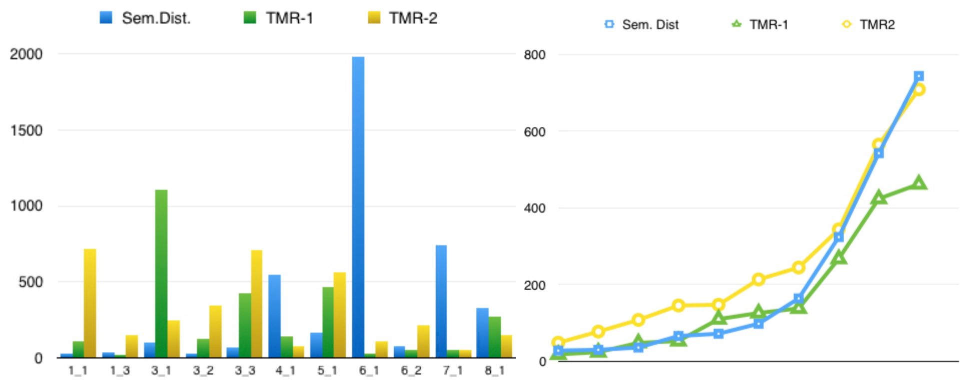

Figure 7.5: Left hand side shows the cost for each approach grouped per conclusion. Right hand side depicts the results in ascending order, excluding the highest values.

(TMR-2), while the average cost reduce from 372 (Sem.Dist.) to 254 (TMR-1) and then increase a bit to 297 (TMR-2).

Figure 7.4 shows on the left hand side the recall for each approach grouped per conclusion, while on the right hand side a complementary vision depicts the recall results in ascending order. This means that the x-axis does not identify the same conclusions anymore. For example, the first four results for Semantic Distance corresponds to conclusions 3_1, 3_3, 5_1 and 7_1 where recall is zero, while the first three results for TMR-1 corresponds to conclusions 3_1, 6_1 and 7_1 where recall is zero, and the first result for TMR-2 correspond to 4_1 where recall is 2/5. Similar graphics are presented for cost values in Fig. 7.5. On the right hand side, the highest values were excluded in order to better visualise the other ones in the graphic. It means that the last results in the graphic correspond to the second last result of each strategy, namely conclusions 7_1 (Sem.Dist.), 5_1 (TMR-1) and 3_3 (TMR-2). By combining both recall and cost graphics, one can observe that the TMR-1 performs slightly better than Semantic Distance by providing slightly better recall at a smaller cost. In its turn, the TMR-2 provides much better recall at a slightly higher cost.

The improvements, however, come with the price of manually instantiating the TMR model for each conclusion, in the first place, and providing the alternative descriptions in the second place. The number of retrieved papers is also an important feature that is not addressed in this work, but which we are currently investigating, e.g. selecting the high quality publications. Ranking and filtering strategies are particularly important when more alternative terms are provided, specially using more flexible mechanisms such as related terms.

Another interesting result is the analysis of the missing goals (goal papers not retrieved by TMR-strategy 1) aimed at understanding the limitations of text-based querying PubMed and the possibilities of overcoming them by exploring the TMR model. Table 7.3 illustrates the analysis by presenting for three conclusions the original query terms, the PubMed-ID of some missing goals, some related Mesh terms and Title provided by PubMed for each goal, and finally a description of possible reasons why that goal was not retrieved and the type(s) of divergence. The divergences, annotated in the last column of the table, are due to: (S) low coverage of **S**ynonyms, particularly when a composed action has itself a description; (R) some terms are described more generically, more specifically or by **R**elated concepts; (O) some terms are **O**mitted; (P) some **P**arts of terms are omitted (e.g. *risk of*). By manually addressing some of the divergences, TMR-2 allows for providing better recall with respect to TMR-1, even though the results are biased.

7.6 DISCUSSION

The results show that TMR-1 performed at least as good as Semantic Distance and TMR-2 performed much better at a not much higher 'reading cost' but much higher 'human-intervention cost'. We conclude that a knowledge-driven approach based on the TMR model allows for improving the recall with respect to previous experiments regarding the Dutch Breast Cancer guidelines, as the causation structure suggests important terms to be queried, namely action and effect. Through the analysis of the missing goal-evidences after an initial ex-

Table 7.3: Part of the analysis of missing goals based on results of TMR-strategy 1. Divergences are **S**ynonyms, **R**elated concepts, **O**mitted terms, omitted **P**arts of term.

Concl./Terms	**Goal**	**Mesh terms / Title**	**Reason/Divergence (S,R,O,P)**	
3_1 - breast-conserving therapy - irradiation - survival rate	1627428	- Mastectomy, Segmental - Breast Neoplasms/radiotherapy - Survival Rate	*Breast conserving therapy* can be **interpreted as** *Mastectomy, Segmental*.	(S)
	12812844	- Mastectomy, Segmental - Breast Neoplasms/radiotherapy - Survival Analysis	*Survival Rate* **is related to** *Survival Analysis* (besides the same as line 1)	(R)
	15894097	- Survival Rate - Breast Neoplasms/therapy	The actions are **not mentioned**, or are mentioned as a **broader category**, namely *Breast Neoplasm Therapy*.	(O) or (R)
	11355595	- Mastectomy, Segmental - Breast Neoplasms/radiotherapy	The effect is **not mentioned**.	(O)
5_1 - aesthetic result - psychosocial wellbeing - breast reconstruction - mastectomy	10718173	- Body Image - Depression - Mammaplasty/methods - Mastectomy/psychology Title: The psychological impact of immediate rather than delayed breast reconstruction.	(1) *aesthetic result* **is related to** *Body Image* (2) *psychosocial wellbeing* **is related to** *Depression* (3) The composition *breast reconstruction immediately after mastectomy* can be **interpreted as** immediate breast reconstruction, as in the title of this publication.	(S) and (R)
6_1.1 - primary breast reconstruction - risk of recurrent breast cancer	2545180	- Surgery, Plastic - Neoplasm Recurrence, Local Title: Oncological aspects of immediate breast reconstruction following mastectomy for malignancy.	(1) *Primary Breast Reconstruction* **is interpreted as** *Immediate Breast Reconstruction* (2) *risk of recurrent breast cancer* **is related to** *Neoplasm Recurrence*, because they are actually synonyms by omitting 'risk of'	(S) and (P)

periment, we show that richer ways to describe the medical terms (e.g. richer synonym coverage) besides more flexible search strategies (e.g. only actions or broader categories) are important features for searching relevant new evidences. In future work we will investigate if this holds for guidelines in general by performing experiments on other guidelines.

Although the alternative descriptions provided in our experiment can be considered biased towards the pre-defined goal evidences, we can conclude that the approach is expressive and flexible enough to allow for the adaptations required to retrieve the missing goals. The challenge is then what adaptations to do since the real problem does not provide a gold standard. We will investigate a semi-automated strategy for providing alternative descriptions based on controlled vocabularies such as UMLS or the conclusions' evidences. A text-based related work [45] do explore the CG evidences for retrieving new evidences from PubMed.

Two trade-offs are observed in this work and/or related ones [40, 41, 45, 73]: fully automatic (text-based) *versus* manual (knowledge-driven) approach, as well as recall *versus* precision importance. In particular, the main difference between our previous approaches [40, 41] and the current one is the selection of terms for constructing the term-queries (it can be seen from Figures 7.1 and 7.3). The fully automatic semantic distance method is replaced by a knowledge-driven human intervention method. As expected, although the results are improved, the cost related to human intervention definitely increases. As we believe a middle term can be a suitable solution, we plan to pursue both (i) a semi-automated strategy for instantiating the TMR-model by using NLP and (ii) an interactive strategy that allows the experts to narrow down the results to the more relevant papers. Finally we will improve the model adding information potentially relevant for this task, such as link to evidences, risk factors and intervention comparison.

According to [92], current methodological books for guideline update do not provide formal explicit procedures for assessing the need for update. The authors refer to the use of terms 'dynamic updating' and 'living guideline' to suggest that guidelines are updated promptly and are always up-to-date, such as [78]. This regards more methodological than technological improvement, which would allow the responsible committee to update parts of the guideline more often, re-publishing it online and raising awareness of eventual updates. From a more computer science perspective, [86] already pointed to the need for a change in paradigm on the guideline authoring by adopting a modular structure so that parts of the guideline could be updated independently but also computer tools should support the living aspect of guidelines. Indeed, [10] claims that partial update of guidelines make more sense than updating the whole guideline at once and [91] does implement new paradigm for digital modular guideline authoring. However, to the best of our knowledge, systems to support 'living guideline' are being investigated, but still do not exist. Since our approach is also modular and the new evidence is retrieved per conclusion, which in turn is related to a recommendation, it would be a natural extension to point which part of the guideline can be updated considering the new evidence retrieved.

A follow-up challenge is to suggest what exactly needs to be updated, why and where in the guideline given what has changed in the new evidence with respect to the current evidence. This is in line with the idea of computer-supported 'dynamic updating' and 'living guideline', but it requires the new evidence to be provided in a structured way, rather than in natural language. Some reasons for update are summarized in [89]: "(i) changes in available interventions; (ii) changes in evidence on the benefits and harms of existing interventions; (iii) changes in outcomes that are considered to be important; (iv) changes in evidence that current practice is optimal; (v) changes in values placed on outcomes; and (vi) changes in resources available for health care". Moreover, the existence of higher quality evidence can be identified by using approaches such as [44] that calculates the evidence quality by analyzing the meta-data provided by PubMed.

Despite of the lack of precise guidance aforementioned, [89] advocates that "several reputable guideline producers base the need to update on systematic literature searches that focus on some or all of the PICO questions from the original guideline". The PICO model means: **P**opulation, **I**ntervention, **C**omparison and **O**utcome. The authors of [15, 121] advocate that automatically identifying the PICO elements for a query is very hard. However, given a PICO query (i.e. a query in which the elements are annotated according the PICO model), the obtained papers can be ranked based on a score attributed to the PICO elements and their matching against semi-automated detection of PICO elements in paper's abstract. Similar, broader or narrower elements are considered to 'expand' the original ones. Moreover, [15] highlights that the PICO model specifies the different roles of the elements in a query, and that this should be considered to thoroughly balance elements in the ranking function.

This is in line with various positions advocated in this paper: (i) the need for a semi-automated extraction of the TMR elements, as for the PICO elements, given the difficulties to guarantee correctness in automatic extraction; (ii) that terms have different roles in a conclusion (actions and effects) and therefore are composed differently into the query (see Sect. 7.3), such as for the PICO-based ranking; (iii) the need to 'expand' the original terms with related terms (TMR Strategy-2); and (iv) that precision measurement can be reconsidered by interactive ranking and filtering strategies. Finally, the PICO model partially aligns with the TMR model (actions $\cong$ intervention and effect $\cong$ outcome). Therefore we advocate our approach can be extended for supporting the elaboration of queries based on the PICO model. Further investigation is needed to extend the approach to consider **P**opulation and **C**omparison when composing the query.

7.7 CONCLUSION

This work analyses the information retrieval problem for supporting Clinical Guidelines update tasks. Part of the TMR-model is presented for structuring Clinical Guidelines' conclusions. The structured information is then used to design a new knowledge-driven approach for retrieving relevant scientific publication from the PubMed repository. The contribution is on the method to automatically generate PubMed queries using as input the clinical guidelines conclusions represented according to the TMR model. The implemented experiments evaluate the performance of the knowledge-driven approach for supporting the updating task of the Dutch Breast Cancer Guideline. First, an older version of the guideline is used to extract the conclusions and to formalise it according to the TMR model. Then, the list of scientific papers used to update the guideline, named goal evidences, is obtained from its latest version. Finally, PubMed queries are automatically generated per conclusion according to the proposed method, and are further submitted to PubMed API to retrieve a list of publications for each conclusion in the guideline. The resultant list is checked against the goal evidences. The performance of the proposed approach is compared with previous experiment, showing that the performance can be improved when using knowledge-driven strategies. Although the good performance observed, the proposed approach has a cost: human intervention is required to formalise the free-text guideline-conclusions according to TMR model and to enrich it with alternative descriptions. The weakness of the model is the limited type of information that it can represent, e.g. risk factor is not yet covered. We are working on improvements in the model, in the query elaboration and in the evaluation process to reach better performance. We also aim at reducing human intervention.