Combining glass box and black box evaluations in the identification of heart disease risk factors and their temporal relations from clinical records

Cyril Grouin, Véronique Moriceau, Pierre Zweigenbaum

HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2016 August 08.

Published in final edited form as:

J Biomed Inform. 2015 December ; 58(Suppl): S133–S142. doi:10.1016/j.jbi.2015.06.014.

Combining glass box and black box evaluations in the identification of heart disease risk factors and their temporal relations from clinical records

Cyril Grouin[a,*], **Véronique Moriceau**[a,b], and **Pierre Zweigenbaum**[a]

[a]LIMSI–CNRS, Orsay, France

[b]Université Paris-Sud, Orsay, France

Abstract

Background—The determination of risk factors and their temporal relations in natural language patient records is a complex task which has been addressed in the i2b2/UTHealth 2014 shared task. In this context, in most systems it was broadly decomposed into two sub-tasks implemented by two components: entity detection, and temporal relation determination. Task-level ("black box") evaluation is relevant for the final clinical application, whereas component-level evaluation ("glass box") is important for system development and progress monitoring. Unfortunately, because of the interaction between entity representation and temporal relation representation, glass box and black box evaluation cannot be managed straightforwardly at the same time in the setting of the i2b2/UTHealth 2014 task, making it difficult to assess reliably the relative performance and contribution of the individual components to the overall task.

Objective—To identify obstacles and propose methods to cope with this difficulty, and illustrate them through experiments on the i2b2/UTHealth 2014 dataset.

Methods—We outline several solutions to this problem and examine their requirements in terms of adequacy for component-level and task-level evaluation and of changes to the task framework. We select the solution which requires the least modifications to the i2b2 evaluation framework and illustrate it with our system. This system identifies risk factor mentions with a CRF system complemented by hand-designed patterns, identifies and normalizes temporal expressions through a tailored version of the Heideltime tool, and determines temporal relations of each risk factor with a One Rule classifier.

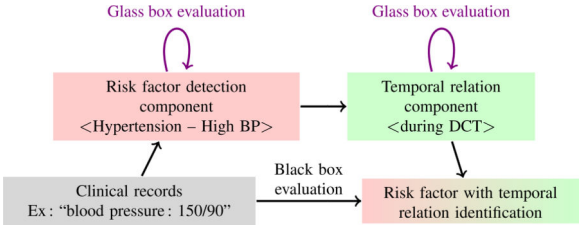
Results—Giving a fixed value to the temporal attribute in risk factor identification proved to be the simplest way to evaluate the risk factor detection component independently. This evaluation method enabled us to identify the risk factor detection component as most contributing to the false negatives and false positives of the global system. This led us to redirect further effort to this

[*]Corresponding author: cyril.grouin@limsi.fr (Cyril Grouin).

component, focusing on medication detection, with gains of 7 to 20 recall points and of 3 to 6 F-measure points depending on the corpus and evaluation.

Conclusion—We proposed a method to achieve a clearer glass box evaluation of risk factor detection and temporal relation detection in clinical texts, which can provide an example to help system development in similar tasks. This glass box evaluation was instrumental in refocusing our efforts and obtaining substantial improvements in risk factor detection.

Graphical abstract



Keywords

Natural Language Processing; Electronic Health Records; Risk Factors; Program Evaluation

1. Introduction

Medical records for diabetic patients contain information about heart disease risk factors. In electronic health records, this information is mainly given in the form of unstructured text. To improve patient care, automatic extraction of medically relevant information can provide clinicians with clues on diverse heart disease risk factors, and their progression over time. Tracking the progression over time of heart disease risk factors in diabetic patients was the topic of the i2b2/UTHealth 2014 challenge [1, 2]. The determination of risk factors from clinical texts requires to detect diseases *(diabetes, coronary artery disease)*, associated risk factors *(cholesterol and hyperlipidemia, hypertension, obesity, smoker status, family history)*, and clues thereof (*medications*); the other part of the task demands to find where in time most of these risk factors occurred on the patient's timeline.

This task description led us to split our system into two components: one for risk factor detection (possibly decomposed into as many sub-components as types of risk factors), and one for temporal relation determination. Combined in a pipeline, they enumerate the risk factors present in a patient record then compute their temporal relations to the current visit. Overall system performance is indeed the most important type of evaluation for the final clinical task. However, when this evaluation reveals a certain number of false positives or false negatives, it is also important to know which component most needs improving. Principled system development should therefore provide a way to evaluate each component independently of each other and of the full system, ideally in such a way as to predict their impact on overall system performance.

We show in this paper that this is not straightforward to obtain in the i2b2/UTHealth 2014 challenge risk factors task (Section 3), and explain why. We examine potential solutions to

this problem, find out that none is fully satisfactory, and implement the one which requires the least modification to the i2b2 evaluation framework (Section 4). We illustrate its application on our risk factor and temporal relation detection system (Sections 5 and 6) and use it to point more clearly at directions for its improvement. We follow the most promising of these directions and obtain substantial gains in system performance (Section 7), then conclude (Section 8).

2. Related work

Information extraction tasks often proposed dual evaluation scenarios in which both full-task (black box) evaluation and component (glass box) evaluation were organized. This is often non-trivial to achieve because of interrelationships between components. For example, the detection of relations generally depends on the former detection of entities which these relations link (note that joint methods are also proposed by some authors, but are not the subject of this paper).

Binary semantic relations such as those which hold between medical problems, tests, and treatments [3] rely on the detection of these concept types. Nevertheless, the 2010 i2b2/VA challenge defined and evaluated two separate sub-tasks through micro-averaged precision, recall and F-measure: concept extraction and relation classification. This provided a glass box evaluation of each sub-task. It did not propose an evaluation of end-to-end concept extraction and relation classification systems, but this (black-box) evaluation would have been easy to run based on the evaluation measures of the relation extraction sub-task.

Binary temporal relations (before, after, etc.) which link events and temporal expressions depend on the detection of these events and times. The 2012 i2b2 temporal relations challenge [4] defined sub-tasks for the identification of EVENTs, the identification of temporal expressions (TIMEX3s), and the detection of the temporal relations between them. This led participants to create three separate components and enabled them to evaluate each of those components through glass box evaluation. Non-trivial issues stemmed from the need to normalize various equivalent configurations of temporal relations. For this purpose their transitive closure was computed before computing their F-measure. Note that choosing the transitive closure instead of, e.g., a minimal underlying temporal graph [5], changes the number of relations that are evaluated.

Co-reference relations detect which mentions in a text refer to the same entities; therefore the determination of these relations also depends on the detection of entity mentions [6]. The 2011 i2b2/VA challenge [7] defined separate sub-tasks for mention detection and co-reference resolution, thus providing glass box evaluation for each sub-task. It also defined an end-to-end task where system mentions were used as input to the co-reference resolution step. Co-reference resolution was evaluated through the MUC, B^3, and CEAF metrics. However, Cai and Strube (2010) [6] showed that the original B^3 and CEAF measures have problems when applied to end-to-end systems, i.e., with concept mentions computed by a first component, and hence not always adequate; they proposed adaptations of these measures to alleviate these problems.

Sometimes relations are viewed instead as concept attributes: this was the case of the 2014 i2b2/UTHealth challenge [2], which defined a task where risk factors had to be detected, together with their temporal relation to the document creation time (DCT). This challenge relates to a large subset of the history of previous i2b2 challenge tasks as well as to the 2014 ShARe/CLEF eHealth T2 shared task [8]. This challenge, which we describe in more detail in Section 3, only defined a black-box evaluation of the end-to-end task, but did not provide a separate, glass-box evaluation of risk factor detection and temporal relation detection. To develop an optimal end-to-end system, we considered it important to obtain a separate evaluation for each of these components. We present in Section 4 the issues we encountered when trying to obtain such a glass-box evaluation and the solution we adopted. We illustrate this glass-box evaluation with the system we developed for the i2b2/UTHealth 2014 challenge (Sections 5 and 6) and discuss how it helped focus error analysis and system improvement (Section 7).

3. Definition of the i2b2/UTHealth 2014 composite task

3.1. Corpus

The corpus we used for the following experiments is the 2014 i2b2/UTHealth corpus, composed of 1,304 patient records from 3 cohorts of diabetic patients for a total of 296 patients. For each patient, about 3 to 5 records are provided per patient, referring to different times in the patient's timeline. The training corpus contained 790 records (178 patients) and the test corpus contained 514 records (118 patients). Patients were distinct between training and test corpora. Based on a random selection, we split the training corpus into our training sub-corpus (89 patients, 390 records) to develop our system, and our development sub-corpus (30 patients, 131 records) to tune the system. Our internal test sub-corpus is composed of 269 records (59 patients).

In our experiments, results on the internal test sub-corpus were obtained with systems trained on the training + development sub-corpora, and results on the official test corpus were obtained with systems trained on the full training corpus.

3.2. Task description: risk factor detection

The task consists in identifying risk factors for diabetic patients in clinical records [9] among 8 categories: diabetes mellitus (DM), coronary artery disease (CAD), hyperlipidemia (HLD), hypertension (HTN), medication (MED), obesity (OBE), family history of CAD (FAM), and smoker status (SMO). The first six categories are events which may take place before, during or after the current visit. Information on how risk factor events are expressed in the document must be specified: for instance, "HTN" or "hypertension" are explicit *mentions* of the risk factor, whereas a test result such as a blood pressure measurement over 140/90 mm/hg is categorized as a *high bp*. This defines sub-types of risk factor events, the full set of which is shown in Table 1. For instance, an expression such as "150/90" should be recorded as *HTN* with sub-type *high bp*.

The task is a document-level entity detection task: what must be determined is whether a risk factor of a given sub-type is present or not in a document, not its specific occurrences

and locations in the document. If multiple explicit mentions of hypertension are found in the document, only one *HTN–mention* record must be created; if the text includes one or more test results revealing hypertension, they must be reported independently as one *HTN–high bp* record.

3.3. Task description: temporal relation determination

Every event risk factor record must be temporally linked to the document creation time (DCT, assumed to represent the time of the visit) through one or more of the three relations BEFORE, DURING, AFTER. Multiple relations are represented by multiple records, each with the relevant temporal relation as an attribute. Therefore, if a text explicitly mentions that the patient has "hypertension" (chronically, thus before, during and after the visit) and also reports a test result performed during the visit revealing a high blood pressure, four records should be produced: *HTN–mention–before*; *HTN–mention–during*; *HTN–mention–after*; and *HTN–high bp–during*.

4. Black box and glass box evaluation of risk factors and their temporal relations

In this section, we summarize the way the i2b2/UTHealth 2014 risk factors task is evaluated. We point at difficulties this induces specifically on the glass box evaluation of risk factor detection and propose methods to cope with these difficulties.

4.1. The i2b2/UTHealth 2014 representations and evaluation

In the i2b2/UTHealth 2014 risk factors task, a risk factor is represented as an object of the relevant type *(CAD, HTN, MED, etc.)* with attributes. One of the attributes (*indicator* or *type*) represents what we described as its sub-type in Table 1. Another attribute (*time*) represents its temporal relation to the document creation time (DCT), i.e., the visit. The external representation of these objects in annotation files takes the form of XML elements, examples of which are shown in Figure 1.

As explained above, if a given information item is provided multiple times in the document, it is to be recorded only once. E.g., if blood pressure is measured twice during the visit and the two measurements correspond to a high blood pressure, only one *(HTN–high bp–during)* record is to be output (this is the case in Figure 1).

Conversely, if a risk factor is true in multiple time spans, one record must be output to represent each such temporal relation. For instance, an explicit mention of "hypertension" generally means that the patient has a chronic condition which spans the before, during, and after periods; in that case, three records are output, as in Figure 1.

The evaluation in the i2b2/UTHealth 2014 risk factors task measures the correction and completeness of these records to compute precision and recall.

4.2. Issues in glass box evaluation of risk factor detection

Glass box evaluation of a single component aims at evaluating its individual successes and mistakes. This is hopefully useful to assess its contribution to the results obtained by the full system when it is evaluated as a black box. As much as possible, it is therefore advisable, for consistency of interpretation, to use the same evaluation measures for both the full system and its individual components.

Glass box evaluation of the last component in a pipeline is simple if gold standard input is available for this component. One only needs to run the component on this gold standard input and to evaluate its output with the same evaluation measures as the full system. This is the case of the temporal relation determination component in the present task: gold standard risk factors are easily derived from the gold standard representations provided with the training corpus by ignoring the value of the temporal relation attribute.

Conversely, glass box evaluation of a non-final component is simple if gold standard input is available for this component, if the component's output has the same form as the full system's output, and if the parts of the output representation that are to be contributed by subsequent components can be ignored by the evaluation program or can be set to the gold standard values. We are close to this situation for the risk factor component in the present task: its output has the same form as the final output (a set of risk factor objects) and it can be fed directly to the evaluation program (the 2014 i2b2/UTHealth evaluation program can be downloaded from https://github.com/kotfic/i2b2_evaluation_scripts/tree/v1.2.1).

However, this evaluation program, despite its numerous and very useful features, has no option to ignore temporal attributes. To cope with this limitation, we might set the temporal attributes of the detected risk factors to those in the gold standard. However, false positive risk factors are by definition not in the gold standard, therefore no temporal attribute value can be drawn from the gold standard to complete them. In that situation, the evaluation procedure must choose which temporal relations to assign to each false positive risk factor. It can decide to assign between one and three relations among before, during, after. Because multiple relations for a risk factor duplicate this risk factor, this will lead to between one and three complete false positive risk factors being scored for one initial relation-less false positive risk factor. This will thus result in different levels of penalization of each false positive risk factor.

More generally, what happens here is that at the level of the risk factor component, the detection of a risk factor sub-type (e.g., *HTN–mention*) will count more or less in the final evaluation depending on the number of temporal relations it will bear eventually. This highlights a difference between a component-oriented glass box evaluation, which would count as one such a risk factor for true positives, false positives, and false negatives, and a system-oriented glass box evaluation, which should count its actual impact on the final output of the whole system. But that impact depends on the behavior of subsequent components, here the temporal relation component: the number of false positive records in the final output depends on the number of relations it would assign to each specific false positive risk factor. We assume this is also why the evaluation program cannot just ignore the temporal relation in its evaluation.

4.3. Methods for glass box evaluation of risk factor detection

The global task of the overall system we are considering in this paper consists in recognizing risk factors and assigning each risk factor a temporal relation. The task organizers provided a black-box evaluation procedure for such an overall system. However, since we divided the system into several components, a separate, glass-box evaluation of each individual component is desirable to trace the origin of remaining problems.

As discussed in the previous section, we encountered difficulty in scoring risk factor detection individually. This stems from the fact that a single risk factor gives rise to a variable number of final records in the task-oriented evaluation. To score risk factor detection individually, we consider the following two options, illustrated in Table 2 for the risk factors mentioned in Figure 1:

Fixed: evaluate the risk factor component according to its individual distribution of true positives, false positives, and false negatives, ignoring the temporal attribute, thus providing an unbiased evaluation of its individual worth; but this will not be directly comparable to the results of the full system;

Gold: evaluate the risk factor component in a way that is comparable with the full system, i.e., by giving gold-standard values to the temporal attribute of each detected risk factor; but for false positive risk factors, which are thus not in the gold standard, this requires to choose which number of temporal relations should be associated to the risk factor, in other words, how many times this false positive will be duplicated. This number will influence the distribution of the evaluated risk factors, whose evaluation will therefore no longer be an evaluation of the risk factor component individually.

As we can see, neither of these two methods is fully satisfactory.

Technically, the first can be implemented by replacing with a fixed value the temporal attribute values both in the gold standard and in the risk factor component output, and removing the potential duplicates this may cause.

The second requires copying the temporal attribute values of gold standard risk factors to the risk factor component output, and deciding what to do for false positives. One way would be to give each of them one dummy relation value, minimizing the final cost of false positives. Another way would be to give them a more realistic number of relations, such as the average number of relations per risk factor produced by the temporal relation component (actually, since an integer number of relations must be produced, one would have to simulate a distribution of relations whose expectation would equal that produced by the temporal relation component). Or yet another way would apply a similar method but with statistics drawn from the gold standard instead of the temporal relation component output. We might also discuss an intermediate way to score false negative risk factors: when assessing a missed risk factor, instead of counting all its gold standard relations, one might count only one, as in the first solution. This would lead to a mixed evaluation, where positives would count as in the second solution, whereas false negatives would count as in the first solution.

Table 3 summarizes the discussed solutions to fix the temporal relations issue in the risk factor glass box evaluation. As one can see, depending on the chosen solution, the numbers of true positive, false positive and false negative risk factors vary accordingly.

Because our first goal is to produce an individual evaluation of the risk factor detection component, which does not depend on the choice and performance of subsequent component run on its output, we opted for the first solution. Besides, it is much simpler to implement.

5. A system to identify risk factors and their temporal relations

We present our system for the detection of risk factors and their temporal relations to the current visit. This system will be used to illustrate the use of glass box evaluation on its two components: risk factor detection and temporal relation determination.

5.1. Risk factor detection component

5.1.1. Lexicon mapping—A baseline approach for risk factor detection consists in collecting all risk factor mentions in the training corpus and extracting these mentions from the test corpus. In the training corpus, we collected 2,225 mentions for medications, 661 for CAD, 512 for hypertension, 427 for smoker status, 348 for diabetes, 97 for hyperlipidemia, 38 for obesity and 12 for family history. Note that this collection only concerns explicit mentions, not test results.

5.1.2. CRF approaches—We designed experiments based on CRF [10] using the Wapiti implementation [11]. Since the task aims at identifying risk factors at the document level, we defined two CRF models: (*i*) a model based on a complex set of features in order to improve precision (Complex CRF model), and (*ii*) a model based on a simple set of features so as to improve recall (Simple CRF model). Because of its very low number of true positives, we did not address the Family History risk factor and gave it a constant value of “not present”. We sum up the experiments we made and the underlying hypotheses in Table 4.

Features: We used the following features to build our models: (*i*) lexical features: the token itself; (*ii*) typographical features: token length in characters, typographic case of the token, presence of punctuation marks, presence of digits; (*iii*) morpho-syntactic features: part-of-speech of the token provided by the Tree Tagger [12] POS tagger; (*iv*) semantic features: pharmacological class of drug names (among 17 classes: *aspirin, beta blocker, diuretic, insulin*, etc.), normalization with respect to the DCT (BEFORE, DURING, AFTER) of the temporal expression found on the same line as the token; and (*v*) document structure: section in which the token occurs among 21 sections we manually defined, based on structure properties found in corpus (*allergies, history of present illness, medications*, etc.). For some features (token, typographic case, POS), we also defined bigrams of features. All features are used in the complex CRF model while the simple CRF model only relies on the token and the part-of-speech tag. We did not perform any cross-validation to train our model. However, an automatic feature selection was performed through the $l1$ regularization.

Post-processing rules (PP): We designed post-processing rules to capture risk factors based on lab results:

- *DM:* A1C test over 6.5 or two fasting glucose measurements over 126;
- *HLD:* total cholesterol over 240 or LDL measurement over 100 mg/dL;
- *HTN* blood pressure measurements over 140/90 mm/hg.

Optimization against annotation imbalance (OAI): Since we considered the risk factor identification task as a token tagging task, the vast majority of tokens in a document are not tagged (97.1%): this creates an imbalance in the annotation, where this untagged category is overrepresented.[1] This is accentuated by the fact that not all mentions of risk factors needed to be marked by the human annotators, provided at least one mention was annotated for each relevant risk factor type. In order to reduce this issue, instead of using the whole document, we defined several window sizes within which unannotated tokens were kept before and after each annotated span. We hypothesize this should both reduce annotation imbalance and reduce the impact of the non-exhaustiveness of mention-level annotation. We achieved our best configuration using a window of 35 unannotated tokens before and after each annotated span.

5.2. Temporal expression extraction component

Since temporal expressions play an important role in this task, a first process detects temporal expressions, normalizes them, and makes them available to subsequent processes. We used the HeidelTime tool [13] to identify temporal expressions within clinical records. HeidelTime is a rule-based system composed of extraction rules and lexicons for normalization. Absolute and relative temporal expressions are extracted and normalized according to the tense of the verb used in the sentence. As a reference time, normalization uses the document creation time (*Record date* in the records). We added about 10 rules to normalize some specific temporal expression formats (e.g. M/DD or M/YY) and then computed the temporal relations (BEFORE, DURING, AFTER) by comparing the normalized value of each expression with the DCT. We thus used those normalizations as semantic clues to identify the risk factors. Figure 2 shows an example of HeidelTime outputs.

5.3. Temporal relation determination component

To compute temporal relations of risk factor events, we designed a supervised learning method inspired by the one we developed for the 2014 ShARe/CLEF eHealth Evaluation Lab [14]. We summarize below the choices we made to design features for this sub-task.

Risk factor type models—As explained earlier, six types of risk factors are events which may bear temporal relations (*CAD, DM, HLD, HTN, OBE, MED*). We observed in the training corpus that each of them has a distinct distribution of temporal relations. Therefore, we created distinct models for these six event types.

[1]On the training corpus, out of 603,985 tokens, 586,730 are not annotated (i.e., 97.1% of all tokens). Out of the 17.255 annotated tokens, 9.718 mark the beginning of an annotated span, for a mean length of 1.78 token per annotated span.

Sub-type features—Second, these event types have sub-types, which we presented in Table 1. Each sub-type in Tables 1a and 1b is used as a feature. In a few cases, a medication can have two sub-types, but as a simplifying approximation we ignored their second type.

Document structure features—Third, a study of the training corpus showed that document structuring into sections often goes together with specific distributions of temporal relations in each section. The section of an event was thus used as a feature. Sections were determined with the same method as was described in the Risk factor detection component section. Besides, BEFORE relations tend to occur closer to the beginning of a document, whereas AFTER relations tend to occur closer to the end. Therefore, as a back-off, we created features to represent this information: we modeled the position of an event in the text as a feature encoding its relative position by splitting the text into five equal-sized bins.

Sentence features—Finally, features were obtained from the sentence in which an event occurs: the bag of words of the sentence, specific time-related words and patterns found in the training corpus (*today*, *at home*, etc.), and the temporal relations inferred from the temporal expressions found in the sentence (see above the presentation of the Temporal expression extraction component). Note that these inferred temporal relations take into account linguistic information such as the verb tense. A simple feature selection scheme was applied to bag-of-word features: we defined thresholds on global token frequency (tf_{min}), number of different sentences in which the token occurred (df_{min}), and kept only the top n most frequent tokens. We empirically set $tf_{min} = 5$, $df_{min} = 3$, and $n = 100$.

As explained earlier, a given event may be assigned multiple relations among BEFORE, DURING, and AFTER. It is thus an instance of multi-label classification. One way to address this problem is to transform it according to the Label Powerset method into a single-label classification task with 2^3 classes: all combinations of these three labels. Another way consists in training a distinct, independent classifier for each of the three relations. We tested both methods (the MEKA[2] extension of the WEKA[3] machine learning suite was used for multi-label classification) and eventually kept the second. When all three classifiers return a negative result for a given risk factor, we assign it the BEFORE value, which we observed gave the best results on the training corpus.

We tested several supervised classification methods, among which we kept WEKA's One Rule (which can be seen as a decision tree with only one level) and SVM (Support Vector Machine with the Sequential Minimal Optimization algorithm) classifiers. As a baseline, we also examined the majority class decision rule (Zero Rule in WEKA).

6. System Evaluation

In this section we instantiate the black box evaluation on our full system and the glass box evaluations on our components. By giving a fixed value to the temporal attribute in the

[2] http://meka.sourceforge.net/
[3] http://www.cs.waikato.ac.nz/~ml/weka/

evaluation of risk factor detection, we obtain a clearer view of where it stands with respect to an ideal component and of its impact on the performance of the global system. This is obtained similarly, albeit much more easily, for the temporal relation component evaluation by running it on gold standard risk factors.

6.1. Black box evaluation of the full system

6.1.1. Results—The black box evaluation of the full system is obtained by running the full pipeline with the risk factor detection and temporal relation detection components, then the evaluation program.

Table 5 shows the global performance of the full pipeline on the internal test sub-corpus. The One Rule classifier obtains the best micro-average precision, recall, and F-measure on this internal test set, and is therefore the chosen classifier for the official test set.

Table 6 shows the global performance of the full pipeline on the official test corpus. For reference, we include the Majority and SVM classifier results in the table.

6.1.2. Discussion—The best F-measure obtained for the full system is at 0.857, which is close to the median position among the twenty participants of the i2b2/UTHealth 2014 challenge Task 2, based on the results published at the workshop.

The Majority decision method of the temporal relation component already obtains a high result, and has been used by several other participants of the challenge. Using an SVM classifier takes into account the collected features and improves this baseline for all risk factors except *MED* (the risk factor with the largest number of occurrences in the training corpus). Surprisingly, even a much simpler classifier such as One Rule performs as well as (even slightly better than) the SVM. We return to these points when discussing the temporal relation component (Section 6.3).

This short examination of the global system through its black box evaluation shows that it performs less well for CAD, Obesity, and Medication, whereas the other event risk factors (DM, HLD, HTN) obtain F-measures around 0.90: does that come from difficulties in the risk factor detection component or in the temporal relation component? This is what their glass box evaluation, to which we now turn our attention, should reveal.

6.2. Glass box evaluation of risk factors

6.2.1. Results—As explained above, we obtain the glass box evaluation of the risk factor component by giving a fixed value to the temporal attribute in its output. We do so by replacing all values in the gold standard and in the risk factor detection output by a generic ‘continuing’ value, which happens to be known by the standard i2b2/UTHealth 2014 evaluation program.

Table 7 displays the results of the mention mapping method on the official test set (micro-average recall, precision, and F-measure). The evaluation presented in this table only takes into account the categories allowing the “mention” value for the *indicator* attribute (see Table 1). Since there is no “mention” attribute value for categories *FAM, MED* and *SMO*,

results cannot be computed for those categories. Moreover, the global results only take into account the results we achieved on the five first categories. The very high recall obtained (at least 0.988) shows that explicit mentions of risk factors exhibit little variation. The moderately high precision (between 0.443 and 0.897) highlights a need for disambiguation.

Table 8 shows the global results we achieved on the risk factor identification sub-task on the internal test set (left pane). The baseline is the mention mapping method. The F-measures obtained by the <Complex+PP>, <Complex+PP+OAI>, <Simple+PP>, and <Simple+PP +OAI> configurations are very close to each other. However, <Simple+PP+OAI> achieves a better precision and recall balance and is thus chosen for our final risk factor detection component. The right pane of Table 8 shows the results obtained by this configuration on the official test set. For reference, it also includes the values obtained by the other configurations.

Table 9 displays the micro-averaged measures we achieve for each risk factor in this setting on the official test corpus.

6.2.2. Discussion—The detailed results on the internal test set, not displayed here for reasons of space, are similar to those on the official test set. They reveal that the class with the lowest results is *CAD* (F=0.697), while the rest of the event risk factors (*MED, OBE, HLD, DM*, and *HTN*) range from 0.864 to 0.905: the glass box evaluation thus identifies the risk factor component as a clear contributor to the lower performance of the full system on *CAD* (−0.30 F-measure compared to a perfect component). Among the next lowest F-measures are those of *MED* (−0.14) and *OBE* (−0.13), which is consistent with their lower scores in the full system too.

An error analysis for the <Simple+PP+OAI> model on the internal test corpus reveals that 124 CAD mentions are correctly detected but 187 are not extracted (no CAD with a *test* or *event* indicator was detected). 141 false positive mentions are also extracted (this is explained by the fact that almost all of these mentions contain the “coronary” token) and 4 CAD mentions are extracted as medications by the CRF, these medications being used for CAD treatment.

We also examined an example of risk factors which need post-processing rules to capture lab results and check their numerical values: the case of Diabetes Mellitus. All undetected DM mentions can be explained by the fact that their lab results are expressed by percentages, which are not taken into account by our post-processing rules.

6.3. Glass box evaluation of temporal relations

6.3.1. Results—Table 10 shows the individual performance of the temporal relation detection component: to obtain this glass box evaluation, this component is run on gold standard risk factors. It is evaluated with the standard i2b2/UTHealth 2014 evaluation program; but since *FAM* and *SMO* risk factors cannot bear a temporal relation, they are excluded from the computation of the *all* column through an option of this program.

6.3.2. Discussion—For most risk factors, a majority class decision produces perfect recall and high precision and F-measure on the internal test. This sets a very high baseline. Only *CAD* has a low recall and *HTN* a less high precision. This high baseline can be explained by the large majority of cases where a risk factor is true both before, during, and after the visit. Therefore, the majority predicts the systematic presence of all relations for all risk factors except CAD, where the majority only predicts the before relation (obtaining a high precision but a much lower recall).

The high performance of the One Rule classifier, already observed in the full system, can be partly explained by the way the problem has been modeled in our temporal relation component. By computing a decision separately for each type of risk factor and for each type of temporal relation, the problem is broken down into 6×3 = 18 sub-problems. Adding one additional level of discrimination is therefore close to building a decision tree with two or three levels. Besides, when the One Rule classifier selects an attribute, it creates a separate feature for each possible value of this attribute. For instance, since the document section attribute has 33 possible values, selecting this attribute (which is done for *MED*) gives One Rule an actual set of 33 features. The most frequent features used in these rules are the risk factor sub-type (see Table 1; in 9 cases), the document structure (section type in 3 cases and relative position in 2 cases), patterns (2 cases of starting and stopping a medication: a drug is taken after the visit if not stopped, and during the visit unless it is to be started; one case of mentioning the Emergency room, which excludes an *HTN* during the visit); and 1 sentence word discovered by the classifier (a mention of *abdomen* excludes the before relation for *OBE*). The fact that the SVM does not improve over One Rule probably means that our other features lack discriminatory power given the training set.

All F-measures are fairly high (0.930 to 0.988), but they are lower for *OBE, CAD*, and *MED* (both for One Rule and the SVM). The glass box evaluation of this component thus shows that it contributes to the lower F-measure of the full system for these three risk factors, aggravating the lower F-measure of the risk factor component for these three risk factors: −0.07 for *OBE* and *CAD* and −0.05 for *MED*.

7. Discussion

We first performed a black box evaluation of a system which identifies risk factors and their temporal relations for diabetic patients from clinical records. It obtained a global micro-averaged F-measure of 0.857 on the i2b2/UTHealth Task 2 test corpus. To obtain more detailed information on the system, we also designed and applied a principled glass box evaluation of its two components.

7.1. Lessons from the glass box evaluation

The first component identifies risk factors with a CRF system and a few post-processing rules based on numerical values over defined thresholds. The second component detects temporal relations with a One Rule classifier after a decomposition into 18 situations.

Overall, glass box evaluation measured an intrinsic F-measure of the risk factor component on the internal test corpus (Table 8, <Simple+PP+OAI>, Internal test, F=0.849), which has a

distance of 0.151 to the maximum of F=1. In contrast, on the same corpus, the temporal relation component (Table 10, One Rule, Internal test, F=0.960) has a much shorter distance of 0.040 to the maximum. This shows that in general, among the missing 0.159 F-measure above the score of 0.841 obtained by the full system with <Simple+PP+OAI+One Rule> (Table 5), the largest progress can come from improvements in the risk factor detection component.

Glass box evaluation also revealed more precisely that the lower scores obtained for the *CAD* risk factor are much more strongly caused by the risk factor component (Table 8, <Simple+PP+OAI>, Internal test, F=0.697) than by the relation detection component (Table 10, One Rule, Internal test, F=0.934) (1−F-measure = −0.303 *vs*. −0.066). An error analysis pointed at the main cause for this observation. Glass box evaluation also showed that the next lower scores, for *OBE* and *MED*, are similarly explained, although in a less extremely unbalanced way: −0.127 *vs*. −0.070 for *OBE*, and −0.136 *vs*. −0.049 for *MED*.

We therefore devoted our subsequent work to the further examination of the risk factor component. Since medications are by far the most numerous mentions in the training corpus (see Section 5.1.1), we focused our effort on them.

7.2. Additional experiments: medication detection

We analyzed the results of glass-box evaluation of medication detection on our internal test sub-corpus. Since recall was much lower than precision, we started with false negatives. Reasons for missed medications included issues in the handling of case and multi-word mentions. Some of the annotated multi-word drug mentions included drug modifiers (*enteric, coated, IV, low, regular, Sliding, sublingual*), aspects of prescription (*add, changes, discontinue, increased, restarting, start*), company names (*Bayer*), and other precisions (*bisulfate* in *clopidogrel bisulfate*, or *maleate* in *enalapril maleate*). This likely made it more difficult for the CRF to learn the more important words, i.e., the drug names (*losartan, nitrate, Novolin, ascorbic acid*) as well as their common abbreviations (*nitro, NTG* for *nitroglycerin*). A number of typos were also spotted (*Altase* for *Altace*, *atneolol* for *atenolol*).

We realized that using the CRF to detect medications was sub-optimal because it aimed to find precise boundaries for each medication mention, which were not needed by the task. We therefore decided to use a fully different method to detect medications: the application of a lexicon of core medication names to the texts, using longest exact match. By core medication names, we mean drug names and abbreviations as exemplified above. Each entry maps an input term to a medication category, with one or two medication sub-types (see Table 1). We initialized the lexicon with medication mentions collected from the training corpus. We reduced its size by only keeping core names for each medication, at the same time making it more able to recognize multi-word mentions built around the same core name. For instance, since *aspirin* was an entry in the lexicon, there was no point in keeping *baby aspirin* or *Bayer aspirin*.

To augment coverage, we added medication names found in main headings and entry terms of the MeSH thesaurus for a number of pharmacological classes (calcium channel blocker,

ACE inhibitor, thiazolidinedione, ARB, sulfonylureas). For this purpose, we followed the MeSH hierarchies and manually selected core names as explained above. Handling typos (other than by adding them into the lexicon) requires a strictly controlled method to avoid the spurious creation of false positive medications; given their relatively small proportion and the risk incurred in handling them, we decided to leave typo processing for future work.

Some false positives were filtered by including longer terms in the lexicon. For instance, while *insulin* by itself most often refers to insulin medication in the training corpus, it can also occur in *insulin dependent diabetes* and other such expressions. We included non-medication entries for these in the lexicon, longest match effectively blocking the recognition of the shorter *insulin* in this context. Specific cases of frequent false positives involved Lasix, a diuretic drug which however does not significantly lower blood pressure and should therefore not be annotated. The generic mention of *beta blockers* was not to be annotated either, and we thus removed this term and its abbreviations from our lexicon. A frequent cause of false positives is the mention of medications in the Allergy section. We therefore stipulated that medications detected in this section would not be sent to the system output. The largest cause of remaining false positives was the presence of negated and hypothetical medications, which we plan to address in future work.

This lexicon-based method obtained a glass box evaluation of P=0.920, R=0.980, F=0.949 for medication detection on the internal test. We used it to bypass CRF processing for medications in a revised risk factor component (based on the Simple+PP+OAI setting), which was evaluated at P=0.892, R=0.887, F=0.889 on the internal test. This increased F-measure by 4 points compared to the glass box evaluation of the initial risk factor component (see Table 8).

Table 11 shows the results of black box evaluation of the former and updated (+MEDLex) full system (with One Rule temporal relation detection). The new medication detection improved medication F-measure by 9 points on the internal test and by 6 points on the official test, thanks to a gain of 20 recall points on the internal test and of 15 points on the official test. It also improved the global F-measure by 4.5 points (internal test) and 3 points (official test).

These improvements brought the performance of medication detection closer to the top systems (3 F-measure points from the top reported medication detection result, P=0.901, R=0.959, F=0.929, [15]), with a recall close to the best (P=0.873, R=0.971, F=0.919, [16]) and a precision that is now the focus of further improvement. More broadly, for the detection of medications, the best systems in the i2b2/UTHealth 2014 challenge used a large lexicon and took into account cues for negation and other situations where a medication is mentioned but not taken by the patient, including sections such as Allergies. Two highly-performing systems [15, 17] used MedEx [18] to collect features and detect medications: MedEx uses a large lexicon of drug names obtained from RxNorm.

In conclusion, the significant improvement we obtained in medication detection was triggered and monitored by the glass box evaluation of the risk factor component. This glass box evaluation allowed us to measure medication recall values close to 0.99 on the internal

test set at some points in our development and to know precisely that we were reaching a ceiling (and should now aim at a better balance with precision). This would not have been possible with the black box evaluation, where because of limitations in the temporal relation component, we would have reached a lower ceiling without knowing whether the remaining false negatives were caused by the risk factor component or by the temporal relation component.

While our medication lexicon still deserves to be extended, the main effort needed in future work on our medication detection method will be the reduction of false positives through the processing of context cues (mostly negated and hypothetical medications). Further error analysis on the training corpus also showed that while the Allergy section rule significantly reduced false positives, wrong detection of Allergies sections also caused false negatives.

8. Conclusion

While developing a system to detect risk factors and their temporal relations, we felt the need for a more precise, glass box evaluation of each component. This was not directly available for risk factors in the i2b2/UTHealth 2014 task 2, and we proposed a method to do so, consisting in standardizing to a fixed value the temporal relation attribute of all risk factors.

We applied this method to perform glass box evaluations of our risk factor and relation detection components. It revealed that the individual performance of the relation detection component was quite high. This can be explained partly by the fact that after breaking down the problem into 18 sub-problems of the form (risk factor type + temporal relation), in each sub-problem a majority decision provides a high baseline method for temporal relation detection. In this situation, training a simple classifier such as One Rule on each sub-problem could obtain a high performance, using risk factor sub-type features in most cases. Temporal relation detection for *MED* and *OBE* has however still room for improvement. Individual F-measure was much lower for the risk factor component. A first error analysis pointed at directions for improvement in this component: fixing the post-processing rules which detect *CAD*, and investigating further the detection of *MED* and *OBE*.

This diagnosis led us to decide that further effort would be better invested in improving the risk factor component. We did so, focusing on the detection of medications, radically changed its method, and significantly improved its performance, with gains of 3 to 6 F-measure points and of 7 to 20 recall points depending on the corpus and evaluation.

Defining an independent evaluation of the risk factor detection component proved to be the delicate point in this process and was instrumental in reaching these conclusions and improvements.

Acknowledgments

This work was supported by the French National Agency for Medicines and Health Products Safety under grant Vigi4MED *(Vigilance dans les forums sur les Médicaments)* ANSM-2013-S-060 and by the National Research Agency, grant Accordys (Content and Knowledge Aggregation for Case-based Reasoning in the field of Fetal Dysmorphology) ANR-12-CORD-0007-03.

References

1. Stubbs A, Uzuner O, Kumar V, Shaw S. Annotation guidelines: risk factors for heart disease in diabetic patients. i2b2/UTHealth NLP Challenge. 2014
2. Stubbs A, Kotfila C, Xu H, Uzuner O. Practical applications for NLP in clinical research: the 2014 i2b2/UTHealth shared tasks. Proc of i2b2/UTHealth NLP Challenge. 2014
3. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011; 18:552–556. [PubMed: 21685143]
4. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc. 2013; 20(5):806–813. [PubMed: 23564629]
5. Tannier X, Muller P. Evaluating temporal graphs built from texts via transitive reduction. Journal of Artificial Intelligence Research. 2011; 40:375–413.
6. Cai, J.; Strube, M. Evaluation metrics for end-to-end coreference resolution systems. Proceedings of the SIGDIAL 2010 Conference; Tokyo, Japan. Association for Computational Linguistics; 2010. p. 28-36.
7. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc. 2012; 19:786–791. [PubMed: 22366294]
8. Kelly, L.; Goeuriot, L.; Leroy, G.; Suominen, H.; Schreck, T.; Mowery, DL.; Velupillai, S.; Chapman, WW.; Zuccon, G.; Palotti, J. Proceedings of the ShARe/CLEF eHealth Evaluation Lab. Springer-Verlag; 2014. Overview of the ShARe/CLEF eHealth evaluation lab 2014.
9. Stubbs A, Kotfila C, Xu H, Uzuner O. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. J Biomed Inform.
10. Lafferty, JD.; McCallum, A.; Pereira, FCN. Proc of ICML. Williamstown, MA: 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data; p. 282-289.
11. Lavergne, T.; Cappé, O.; Yvon, F. Proc of ACL, Uppsala. Sweden: 2010. Practical very large scale CRFs; p. 504-13.
12. Schmid, H. Probabilistic part-of-speech tagging using decision trees; Proc of International Conference on New Methods in Language Processing; Manchester, UK. 1994. p. 44-49.
13. Strötgen J, Gertz M. Multilingual and cross-domain temporal tagging. Language Resources and Evaluation. 2013; 47(2):269–298.
14. Hamon T, Grouin C, Zweigenbaum P. Disease and disorder template filling using rule-based and statistical approaches. Working notes of the ShARe/CLEF eHealth Evaluation Lab. 2014
15. Chen Q, Li H, Tang B, Liu X, Liu Z, Liu S, Wang W. Identifying risk factors for heart disease over time—HITSZ's system for track 2 of the 2014 i2b2 NLP challenge. Proc of i2b2/UTHealth NLP Challenge. 2014
16. Yang H, Garibaldi J. Automatic extraction of risk factors for heart disease in clinical texts. Proc of i2b2/UTHealth NLP Challenge. 2014
17. Torii M, wei Fan J, li Yang W, Lee T, Wiley MT, Zisook D, Huang Y. De-identification and risk factor detection in medical records. Proc of i2b2/UTHealth NLP Challenge. 2014
18. Xu H, Stenner S, Doan S, Johnson K, Waitman L, Denny J. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010; 17(1):19–24. [PubMed: 20064797]

Author Manuscript

Author Manuscript

Author Manuscript

Highlights

- A system to detect risk factors and temporal relations in clinical records
- Black box and glass box evaluations are both needed to develop complex systems
- Proposed a way to limit the impact of a component when evaluating the other
- Glass box evaluation of individual components shows directions for improvement

```
<CAD id="DOC24" indicator="mention" time="before DCT" />
<CAD id="DOC16" indicator="mention" time="during DCT" />
<CAD id="DOC3" indicator="mention" time="after DCT" />
<HYPERTENSION id="DOC21" indicator="mention" time="before DCT" />
<HYPERTENSION id="DOC2" indicator="mention" time="during DCT" />
<HYPERTENSION id="DOC10" indicator="mention" time="after DCT" />
<HYPERTENSION id="DOC7" indicator="high bp" time="during DCT" />
```

Figure 1.
Example risk factor representations in external (XML) format

```
Record date: 2015-06-09
(...)
Date of Visit: <TIMEX3 tid="t3" type="DATE" value="2015-06-09"
    rel="during">06/09/2015</TIMEX3>
Kirill was seen <TIMEX3 tid="t12" type="DATE" value="2015-06-09"
    rel="during">today</TIMEX3> by myself and Dr. Veronica Morris.
HISTORY OF THE PRESENT ILLNESS: Kirill is a 53-year-old man who suffered a
    ''silent heart attack'' in <TIMEX3 tid="t15" type="DATE"
    value="2013-11" rel="before">November 2013</TIMEX3>.
```

Figure 2.
HeidelTime outputs: normalization of absolute/relative dates and temporal relations with respect to the record date

Table 1

Sub-types of medications (b) and other risk factor events (a)

	event type	sub-type
(a)	diabetes mellitus (DM)	mention, A1C, glucose
	coronary artery disease (CAD)	mention, event, symptom, test
	hyperlipidemia (HLD)	mention, high LDL, high chol.
	hypertension (HTN)	mention, high bp
	obesity (OBE)	mention, BMI

	event type	sub-type
(b)	medication (MED)	insulin, metformin, calcium channel blocker, statin, aspirin, ACE inhibitor, beta blocker, nitrate, diuretic, ezetimibe, ARB, sulfonylureas, fibrate, thienopyridine, niacin, thiazolidinedione, DPP4 inhibitors

Table 2

Methods to handle temporal relations when scoring the three example risk factors of Fig 1. HTN = Hypertension. See text for the meaning of Fixed and Gold

Method	Risk factor 1	Risk factor 2	Risk factor 3
Source	CAD mention	HTN mention	HTN high bp
Fixed	CAD mention *dummy*	HTN mention *dummy*	HTN high bp *dummy*
Gold	CAD mention before	HTN mention before	
Gold	CAD mention during	HTN mention during	HTN high bp during
Gold	CAD mention after	HTN mention after	

Table 3

Number of risk factor true positives (TP), false positives (FP) and false negatives (FN) depending on the chosen solution for glass box evaluation of risk factor detection

Temporal relation solution	TP	FP	FN
#1 Fixed value	One relation	One relation	One relation
#2 Copy gold standard	All relations (up to 3)	Dummy relations (1, or up to 3)	All relations (up to 3, or only 1)

Table 4

Summary of experiments performed to identify risk factors. PP=post-processing rules, OAI=optimization against annotation imbalance (*n*=number of tokens before/after annotated tokens)

CRF model	PP	OAI	Tested hypothesis
Complex	No	No	A CRF with complex features identifies more risk factors than a lexicon projection
Complex	Yes	No	Post-processing rules identify risk factors represented as numerical values higher than defined threshold
Simple	Yes	No	A CRF with simple features (the token and its part-of-speech tag) identifies already known risk factors
Simple	Yes	Yes (*n* = 35)	The reduction of unannotated tokens occurring before and after annotated tokens counters annotation imbalance and improves results

Table 5

Micro measures (P=Precision, R=Recall, F=F-measure) for the full prediction of risk factors and temporal relations computed on the internal test set. Detailed categories: DM=Diabetes mellitus, HLD=Hyperlipidemia, HTN=Hypertension, OBE=Obese, MED=Medication, FAM=Family history of CAD, SMO=Smoker. The Simple+PP+OAI setting of the risk factor component is used (see Table 8). Majority, One Rule, and SVM are the three classifiers of the temporal relation determination component. They are applied to the event risk factors, i.e., all but FAM and SMO. Bold face shows the highest value across all three classifiers, with ties if difference ≤ 0.002

Classifier		Internal test: full system								
		CAD	**DM**	**HLD**	**HTN**	**OBE**	**MED**	**FAM**	**SMO**	**all**
Majority	P	.718	.839	.922	.698	.703	.851	.952	.752	.818
	R	.311	**.909**	**.831**	**.902**	**.872**	**.787**	.952	.792	.789
	F	.433	.872	.874	.787	.779	.817	.952	.772	.803
One Rule	P	**.783**	**.933**	**.948**	**.947**	.723	**.864**	.952	.752	**.875**
	R	**.670**	.895	.831	.888	.865	.774	.952	.792	**.809**
	F	**.722**	**.914**	**.886**	**.916**	.788	**.817**	.952	.772	**.841**
SVM	P	.743	.898	.938	.932	**.763**	.859	.952	.752	.864
	R	**.670**	.901	**.831**	.878	**.872**	.775	.952	.792	**.809**
	F	.704	.899	.881	.904	**.814**	**.815**	.952	.772	.836

Table 6

Micro measures (P=Precision, R=Recall, F=F-measure) for the full prediction of risk factors and temporal relations computed on the official test set. The Simple+PP+OAI setting of the risk factor component is used (see Table 8). Bold face shows the highest value across all three classifiers, with ties if difference ≤ 0.002

Classifier		Official test: full system								
		CAD	**DM**	**HLD**	**HTN**	**OBE**	**MED**	**FAM**	**SMO**	**all**
Majority	P	.752	.841	.928	.744	.790	.854	.963	.771	.837
	R	.318	**.887**	**.852**	**.922**	**.878**	**.819**	.963	.787	.811
	F	.447	.864	.888	.824	.832	**.836**	.963	.779	.824
One Rule	P	**.797**	**.942**	**.964**	**.938**	**.853**	.860	.963	.771	**.881**
	R	**.691**	.880	**.850**	.913	.863	.814	.963	.787	**.833**
	F	**.740**	**.910**	**.903**	**.925**	**.858**	**.836**	.963	.779	**.857**
SVM	P	.757	.917	.955	.931	.828	**.864**	.963	.771	.876
	R	**.693**	.875	**.850**	.913	.847	.808	.963	.787	.830
	F	.724	.896	.899	.922	.838	**.835**	.963	.779	.852

J Biomed Inform. Author manuscript; available in PMC 2016 August 08.

Table 7

Micro measures for risk factor mention mapping on the official test set, evaluated against gold standard risk factors of sub-type mention. Detailed categories: DM=Diabetes mellitus, HLD=Hyperlipidemia, HTN=Hypertension, OBE=Obese, MED=Medication, FAM=Family history of CAD, SMO=Smoker

Baseline	Official test								
	CAD	DM	HLD	HTN	OBE	MED	FAM	SMO	all
Precision	.443	.803	.698	.897	.747	.000	.000	.000	.718
Recall	.996	1.00	.996	.995	.988	.000	.000	.000	.996
F-measure	.613	.891	.821	.943	.851	.000	.000	.000	.835

Table 8

Micro measures for risk factor identification on both internal and official test sets. Bold font highlights improvement of results with respect to the preceding experiment

Experiment	Internal test			Official test		
	P	**R**	**F**	**P**	**R**	**F**
Baseline	.580	.795	.671	.591	.807	.682
Complex	**.914**	.753	**.826**	**.915**	.749	**.824**
Complex+PP	**.917**	**.791**	**.849**	**.915**	**.781**	**.843**
Complex+PP+OAI	.912	**.792**	.848	.909	**.797**	**.850**
Simple+PP	.900	**.796**	.845	.900	**.806**	**.851**
Simple+PP+OAI	.893	**.809**	.849	.898	**.823**	**.859**

Table 9

Micro measures (P=Precision, R=Recall, F=F-measure) for risk factor identification on the official test set. Bold font highlights improvement of results with respect to the preceding experiment. Detailed categories: DM=Diabetes mellitus, HLD=Hyperlipidemia, HTN=Hypertension, OBE=Obese, MED=Medication, FAM=Family history of CAD, SMO=Smoker

Experiment		Official test:risk factor detection component								
		CAD	DM	HLD	HTN	OBE	MED	FAM	SMO	all
Complex	P	.897	.967	.989	.943	.940	.935	.963	.725	.915
	R	.533	.692	.670	.690	.630	.790	.963	.732	.749
	F	.669	.807	.799	.797	.755	.857	.963	.729	.824
Complex+PP	P	.897	.962	.980	.938	.937	.935	.963	.725	.915
	R	.533	**.772**	**.707**	**.857**	**.740**	**.793**	.963	.732	**.781**
	F	.669	**.857**	**.821**	**.896**	**.827**	**.859**	.963	.729	**.843**
Complex+PP+OAI	P	.846	.956	.976	.923	.930	.930	.963	**.738**	.909
	R	.531	**.800**	**.732**	**.864**	**.800**	**.814**	.963	**.746**	**.797**
	F	.652	**.871**	**.836**	.892	**.860**	**.868**	.963	**.742**	**.850**
Simple+PP	P	.799	.926	.961	.931	.919	.920	.963	**.751**	.900
	R	**.583**	**.810**	**.797**	**.875**	.790	.806	.963	**.768**	**.806**
	F	**.674**	.864	**.871**	**.902**	.850	.859	.963	**.759**	**.851**
Simple+PP+OAI	P	.795	.930	.961	.920	.887	.915	.963	**.770**	.898
	R	**.620**	**.834**	**.801**	**.891**	**.860**	**.819**	.963	**.791**	**.823**
	F	**.697**	**.880**	**.874**	**.905**	**.873**	**.864**	.963	**.780**	**.859**

Table 10

Micro measures for temporal relations computed for gold risk factors on the internal test set. Bold face shows the highest value across all three classifiers, with ties if difference ≤ 0.002

Classifier		Internal test: temporal relation component						
		CAD	**DM**	**HLD**	**HTN**	**OBE**	**MED**	**all**
Majority	P	.915	.899	.937	.730	.837	.907	.875
	R	.499	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	.954
	F	.646	.947	.967	.844	.911	**.951**	.913
One Rule	P	**.950**	**.989**	**.988**	**.986**	.874	**.918**	**.941**
	R	.919	.981	.988	.982	.993	.987	**.979**
	F	**.934**	**.985**	**.988**	**.984**	.930	.951	**.960**
SVM	P	.936	.965	.970	.977	**.911**	**.918**	.936
	R	**.927**	.981	.997	.963	1.00	.985	**.978**
	F	.931	.973	.983	.970	**.953**	**.950**	.956

J Biomed Inform. Author manuscript; available in PMC 2016 August 08.

Table 11

Micro measures for black box evaluation of the full system on both internal and official test sets. The CRF part of risk factor identification uses the Simple+PP+OAI setting. MED = evaluation restricted to medications only. MEDLex = with new medical lexicon method. Bold font highlights the best results in two compared experiments

Experiment	Evaluation	Internal test			Official test		
		P	R	F	P	R	F
	Black box, MED	**.864**	.774	.817	**.860**	.814	.836
+MEDLex	Black box, MED	.849	**.972**	**.907**	.836	**.965**	**.896**
	Black box, all	**.875**	.809	.841	**.881**	.833	.857
+MEDLex	Black box, all	.865	**.907**	**.886**	.864	**.909**	**.886**