Identification de facteurs de risque pour des patients diabétiques à partir de comptes-rendus cliniques par des approches hybrides

Cyril Grouin[1] Véronique Moriceau[1, 2] Sophie Rosset[1] Pierre Zweigenbaum[1]
(1) LIMSI–CNRS, UPR 3251, rue John von Neumann, 91400 Orsay
(2) Université Paris-Sud, Campus universitaire d'Orsay, 91400 Orsay
{prenom.nom}@limsi.fr

Résumé. Dans cet article, nous présentons les méthodes que nous avons développées pour analyser des comptes-rendus hospitaliers rédigés en anglais. L'objectif de cette étude consiste à identifier les facteurs de risque de décès pour des patients diabétiques et à positionner les événements médicaux décrits par rapport à la date de création de chaque document. Notre approche repose sur (i) HeidelTime pour identifier les expressions temporelles, (ii) des CRF complétés par des règles de post-traitement pour identifier les traitements, les maladies et facteurs de risque, et (iii) des règles pour positionner temporellement chaque événement médical. Sur un corpus de 514 documents, nous obtenons une F-mesure globale de 0,8451. Nous observons que l'identification des informations directement mentionnées dans les documents se révèle plus performante que l'inférence d'informations à partir de résultats de laboratoire.

Abstract.
Risk factor identification for diabetic patients from clinical records using hybrid approaches
In this paper, we present the methods we designed to process clinical records written in English. The aim of this study consists in identifying risk factors for diabetic patients and to define the temporal relation of those medical events wrt. the document creation time. Our approach relies (i) on HeidelTime to identify temporal expressions, (ii) on CRF and post-processing rules to identify treatments, diseases and risk factors, and (iii) on rules to determine the temporal relation of each medical event. On a corpus of 514 documents, we achieved a 0.8451 global F-measure. We observe we performed best on the identification of information mentionned in the text than information inference from lab results.

Mots-clés : Comptes-rendus hospitaliers, extraction d'information, apprentissage statistique.

Keywords: Electronic Health Records, Information Extraction, Machine Learning.

1 Introduction

Les documents cliniques contiennent des informations personnelles (description de l'environnement familial et social) et cliniques (examens, maladies, traitements) qui sont structurées et redondantes. Les comptes-rendus hospitaliers, notamment américains, sont structurés selon le modèle SOAP [1] (informations subjectives, informations objectives, résultats, conclusion/conduite à tenir) afin d'assurer l'interopérabilité entre comptes-rendus hospitaliers. Les informations les plus utiles pour établir le diagnostic d'un patient sont généralement répétées dans les différents documents qui constituent le dossier médical personnel d'un patient. La redondance et la dispersion de ces informations entre plusieurs documents nécessitent la mise au point d'outils informatiques permettant d'analyser le contenu de ces documents pour en extraire les informations pertinentes pour l'aide au diagnostic. Le résultat de ces traitements automatiques vise à découvrir de nouvelles informations (interactions médicamenteuses, effets secondaires, facteurs de risque, etc.) qui échappent à l'humain en raison de leur nombre.

Le diabète (terme générique couramment employé pour désigner le "diabète sucré") est une maladie qui se traduit par la perturbation de la régulation des sucres de l'organisme par l'insuline [2]. Cette maladie se traduit par une augmentation du taux de sucre dans le sang. L'organisation mondiale de la santé (OMS) rapporte que le diabète de type 2 constitue la

1. L'acronyme renvoie aux quatre principales sections présentes dans un compte-rendu clinique : *Subjective, Objective, Assessment and Plan*.
2. `http://www.chu-rouen.fr/page/diabete`

forme de diabète la plus répandue dans le monde (90% des diabètes rencontrés [3]). Qualifié de "diabète de la maturité", il survient chez les adultes âgés et augmente de 50% le risque de décès par une maladie cardio-vasculaire [4]. Les facteurs de risque qui augmentent les risques de décès sont connus et documentés (hypercholestérolémie, hypertension, obésité, tabagisme).

L'objectif du travail décrit dans cet article consiste à repérer automatiquement les facteurs de risques de développement de maladies cardio-vasculaires par des patients diabétiques, depuis les documents cliniques au format textuel. Les méthodes que nous avons développées ainsi que les données utilisées s'inscrivent dans le cadre de la campagne d'évaluation internationale i2b2 dont l'édition 2014 portait notamment sur cette problématique (Stubbs *et al.*, 2014a), à partir de documents cliniques rédigés en anglais.

2 État de l'art

Les informations pertinentes pour établir des diagnostics cliniques sont exprimées de deux manières différentes dans les comptes-rendus cliniques. Soit de manière explicite dans le cas où l'information est directement mentionnée (*le patient est connu pour une histoire de diabète*), soit de manière implicite par le biais d'événements médicaux (mode de vie : *tabagisme actif, boit occasionnellement de l'alcool*, résultats de laboratoire : *pression artérielle de 146/88*) qu'il importe d'analyser pour en inférer des informations (*146/88 mm/hg* $\rightarrow$ hypertension). Si la majorité des informations cliniques concernent le patient, certains événements peuvent se rapporter à la famille du patient (*son père avait du diabète, était fumeur, et est décédé d'un infarctus du myocarde à 65 ans*). Il importe alors de déterminer à qui se rapportent les informations mentionnées. D'autre part, l'existence continue et répétée dans le temps de certains événements médicaux influe sur la nature du diagnostic. Il est donc essentiel de tenir compte du positionnement temporel de ces événements par rapport à la date de consultation mentionnée en début de compte-rendu clinique. Plusieurs éditions récentes des campagnes d'évaluation internationales organisées par l'institut i2b2 [5] ont portés sur ces aspects. La comparaison des méthodes employées nous permet de mettre en évidence les méthodes les plus pertinentes au vu des objectifs poursuivis et des résultats obtenus.

2.1 Détection d'événements médicaux

La détection des événements médicaux (examens, maladies, modes de vie, traitements) depuis des comptes-rendus clinique est possible, soit par des approches à bases de règles et de projection de lexiques, soit par des approches par apprentissage statistique. Le choix de l'approche dépend, d'une part de la disponibilité de corpus annotés (pré-requis indispensable pour l'apprentissage statistique), et d'autre part du type d'informations à traiter (repérage d'entités nommées réalisable au moyen des deux approches par opposition à l'inférence d'informations uniquement possible avec des règles).

Afin de détecter dans des comptes-rendus hospitaliers les occurrences de traitements médicaux (noms de médicaments) et les informations associées (dosage, mode d'administration, durée, fréquence, etc.), (Doan *et al.*, 2010) ont produit un système qui découpe les documents en sections et en phrases, puis qui applique des règles d'étiquetage sémantique. Le découpage permet de repérer les passages porteurs d'informations, mais également de permettre le calcul de rattachements d'informations dans le cas de reprises pronominales. Sur un corpus de 251 documents issu de la campagne i2b2 2009 (Uzuner *et al.*, 2010), pour une évaluation au niveau des tokens [6], les auteurs rapportent une F-mesure globale de 0,821 avec une précision (0,840) supérieure au rappel (0,803). En l'absence de corpus annoté, tous les participants de cette campagne ont produit des systèmes à base de règles.

En matière de détection du tabagisme chez des patients *(non fumeur, fumeur, ancien fumeur, statut inconnu)* depuis des comptes-rendus cliniques, (Clark *et al.*, 2008) ont mis en place un système en deux étapes de manières à (i) classer les documents selon qu'ils contiennent des indices sur le tabagisme, et (ii) pour les documents contenant de tels indices, effectuer une analyse linguistique du contenu pour associer ces indices à des expressions temporelles. Pour réaliser cette deuxième étape, les auteurs ont utilisé une approche à base de SVM. Sur le corpus i2b2 2006 composé de 104 documents (Uzuner *et al.*, 2008), et pour une évaluation au niveau du document [7], les auteurs rapportent une exactitude de 93,6% pour la détermination du statut du tabagisme et une exactitude de 100% concernant la première étape de filtrage des documents.

3. http://www.who.int/mediacentre/factsheets/fs312/fr/
4. http://www.who.int/mediacentre/factsheets/fs138/fr/
5. Integrating Informatics and Biology to the Bedside, https://www.i2b2.org/NLP/
6. Une évaluation au niveau des tokens prend en compte toutes les occurrences d'une même forme : pour un nom d'un médicament répété plusieurs fois dans un document, l'évaluation prendra en compte chaque apparition.
7. Dans le cas d'une évaluation au niveau du document, une seule occurrence de l'information traitée est attendue.

Enfin, de manière à identifier les facteurs de comorbidité *(asthme, attaque cardiaque, dépression, diabète, hypercholestérolémie, hypertension, hypertriglycéridémie, maladie cardio-vasculaires, obésité, etc.)*, (Childs *et al.*, 2008) ont produit un système composé de 281 règles et 9 étapes dans le but de reproduire les "signaux" médicaux qu'un expert humain considérerait comme pertinents pour décider de l'existence (présent, possible, absent, inconnu) de chaque facteur dans un compte-rendu clinique. Parmi les étapes appliquées figurent notamment l'application de NegEx (Chapman *et al.*, 2001) pour marquer la négation et l'incertitude. Sur le corpus i2b2 2008 contenant 8044 facteurs à détecter (Uzuner, 2009), les auteurs ont obtenu une micro F-mesure de 0,9773.

2.2 Positionnement des événements dans la chronologie des patients

Le positionnement d'événements médicaux dans la chronologie du patient constitue un champs de recherche récent, dans lequel les approches à base d'apprentissage sont largement employées, parfois complétées par des règles (Sun *et al.*, 2013). L'identification des expressions temporelles et la normalisation de ces expressions selon un format pivot est généralement réalisée au moyen d'outils de repérage de ce type d'expressions, tels que GUTime (Verhagen *et al.*, 2005), HeidelTime (Strötgen & Gertz, 2010), ou SUTime (Chang & Manning, 2012).

A partir de documents déjà annotés en expressions temporelles et événements médicaux, (Cherry *et al.*, 2013) ont combiné des approches à base d'apprentissage (entropie maximale et SVM) et de règles et lexiques pour repérer les relations temporelles (avant, pendant, après) qui existent, soit entre deux événements médicaux (maladie, examen, traitement, etc.), soit entre un événement médical et une expression temporelle (date, heure). Afin de traiter les particularités des relations temporelles, les auteurs ont défini quatre systèmes permettant de repérer : (i) les relations temporelles locales, (ii) les relations temporelles qui existent entre sections d'un document, (iii) les relations distantes dans le document entre événements se produisant au même moment, et (iv) les relations distantes qui entrent dans le cadre d'un lien de causalité. Sur un corpus de 120 documents, les auteurs rapportent une F-mesure globale de 0,6837 avec une précision (0,7537) supérieure au rappel (0,6449).

2.3 Détection des événements médicaux et positionnement temporel

Sur la campagne d'évaluation i2b2 2014 dans laquelle nous inscrivons ce travail, plusieurs participants ont utilisés des annotations complémentaires, soit en annotant manuellement des données issues de leur organisme (Roberts *et al.*, 2014), soit en réutilisant les données de l'édition 2006 sur le tabagisme (Cormack *et al.*, 2014). Ces annotations permettent de s'assurer de la cohérence globale des annotations et de garantir leur pertinence pour la méthode utilisée. Les participants ayant obtenu les meilleurs résultats sont ceux qui, en plus de l'utilisation d'annotations complémentaires, ont conçu plusieurs classifieurs selon les types d'information à traiter, un classifieur global et un deuxième dédié au tabagisme (Torii *et al.*, 2014, F=0,9209), ou en enchaînant plusieurs étapes (Roberts *et al.*, 2014, F=0,9277) incluant identification des concepts au moyen de lexiques, filtrage de ces concepts et positionnement temporel au moyen d'approches par apprentissage telles que les modèles de langue.

3 Objectifs

3.1 Présentation

L'objectif global que nous poursuivons consiste à repérer les facteurs de risque de développement de maladies cardiaques par des patients diabétiques parmi huit catégories (Stubbs *et al.*, 2014b,a). Le moment où apparaissent ces différents facteurs de risque dans la chronologie de la consultation constitue également une information capitale et permet d'étudier la progression des maladies cardiaques dans le temps. Il existe trois types d'information pertinentes pour répondre à cette problématique :

– les maladies connues : *diabète, maladie coronaro-artérielle (CAD)*,
– les facteurs de risque associés : *cholestérol et hyperlipidémie, hypertension, obésité, tabagisme, histoire familiale de maladies coronaro-artérielles*,
– et des indices annexes : *médicaments*.

Pour chacune de ces huit catégories existent des informations associées, dont le tableau 1 renseigne des différentes valeurs possibles en anglais (langue utilisée dans les documents de notre corpus, voir section 4) : pour les médicaments, la classe

pharmacologique ; pour les maladies, un indicateur de la manière dont l'information est présentée dans le document : soit l'information est directement mentionnée (mention), soit elle doit être inférée à partir de résultats de laboratoire (A1C, high LDL, etc.) ; et pour le tabagisme, le statut.

(a)

Élément	Classes pharmacologiques
Médicament	insulin, metformin, calcium channel blocker, statin, aspirin, ACE inhibitor, beta blocker, nitrate, diuretic, ezetimibe, ARB, sulfonylureas, fibrate, thienopyridine, niacin, thiazolidinedione, DPP4 inhibitors

(b)

Élément	Indicateurs
Diabète	mention, A1C, glucose
CAD	mention, event, symptom, test
Hyperlipidémie	mention, high LDL, high chol.
Hypertension	mention, high bp
Obésité	mention, BMI

(c)

Élément	Statut
Tabagisme	current, past, ever, never, unknown

TABLE 1 – Classes pharmacologiques des médicaments (a), indicateurs associés aux maladies (b), statut du tabagisme (c)

Si les maladies ne sont pas directement mentionnées dans le document (par exemple, les occurrences "hypertension" et "HTN" constituent des mentions et doivent être notées comme telles en tant qu'indicateur), elles doivent être inférées depuis des résultats de laboratoire, uniquement si les valeurs de ces résultats dépassent des seuils prédéfinis [8] :

- Diabète :
  - dosage de l'hémoglobine A1c supérieur à 6,5 mmol/L,
  - ou deux valeurs successives de glycémie à jeun supérieures à 126 mg/dL ;
- Hyperlipidémie :
  - taux de cholestérol total supérieur à 240 mg/dL,
  - ou taux de cholestérol LDL (également appelé "mauvais cholestérol") supérieur à 100 mg/dL ;
- Hypertension : pression sanguine supérieure à 140/90 mm/hg.

3.2 Réalisation

Nous poursuivons donc l'objectif global de : (i) repérage des éléments médicaux parmi les huit catégories précédemment listées, (ii) spécification de la manière dont l'information est représentée dans chaque catégorie (classe pharmacologique, indicateur, statut), et (iii) de détermination du positionnement temporel de ces facteurs de risque par rapport à une date de référence (dans le cas présent, la date de création du document (DCT) a été retenue) parmi trois valeurs possibles (avant, pendant, après la DCT).

L'identification de ces différents éléments se fait au niveau du document. Ainsi, il importe uniquement de connaître les médicaments, les maladies et les facteurs de risque d'un patient, quel que soit le nombre d'occurrences de chacun de ces éléments. Cependant, parce que chaque occurrence d'un élément peut renvoyer à différents moments de la chronologie de la consultation, plusieurs positionnements temporels peuvent être affectés à un même facteur de risque. Un patient pourra, par exemple, avoir pris un traitement médical avant, pendant et après la consultation à laquelle renvoie le document clinique. Cette particularité renvoie donc à une tâche de classification multi-labels.

4 Corpus

Le corpus que nous avons utilisé provient de l'édition 2014 du challenge i2b2. Il se compose de documents cliniques rédigés en anglais, issus de la base de données MIMIC-II (Saeed *et al.*, 2011). Ils conservent la forme d'origine du document papier, en particulier une largeur de colonne fixe, la présence de lignes blanches entre deux lignes de texte pour reproduire un double espacement, le positionnement exact des éléments sur la page (tabulation, espaces multiples), la présence de symboles particuliers pour représenter les séparateurs de colonnes de tableau (accent circonflexe, barre verticale), etc. Nous observons que seuls certains documents ont fait l'objet d'un pré-traitement pour rétablir chaque

8. Les seuils des valeurs numériques correspondent aux seuils communément admis dans la communauté médicale. Ces seuils figurent dans le guide d'annotation qui a été utilisé par les annotateurs et fourni par les organisateurs aux participants.

phrase sur une même ligne, en supprimant les sauts de ligne à l'intérieur d'une phrase. Aucune tokénisation n'a été réalisée dans le corpus.

Le corpus comprend 1 304 documents cliniques relatifs à 296 patients distincts. Les dossiers de chaque patient intègrent entre 3 et 5 documents, renvoyant à différentes consultations dans le temps. Trois cohortes de patients constituent ce corpus : (i) les patients qui ont déjà une maladie coronaro-artérielle, (ii) les patients qui n'ont pas de maladie coronaro-artérielle, et (iii) les patients qui développent une maladie coronaro-artérielle pendant la période couverte par leur dossier. Les documents cliniques sont de différents types : document d'admission, document de transfert, lettre de sortie et lettres de correspondance entre chirurgien et médecin, sans que le type de document soit clairement spécifié dans chaque fichier. Nous donnons en figure 1 un extrait de document issu du corpus, en mettant en évidence les informations pertinentes pour notre double problématique de détection de facteurs de risque et de positionnement temporel de ces événements.

Record date : 2086-11-07
(...)
I just had the pleasure of seeing Ms. Benitez for a follow up cardiovascular examination. She has been reasonably stable since I saw her last **in July**. She has had several episodes of chest discomfort characterized as a tightness which has been quite transient. The frequency has been less than one time per month. The last occurred **yesterday** while she was walking in a shopping mall.
(...)
She has had no symptom of cardiovascular ischemia, denying transient hemiparesis, hemiparesthesia, or amaurosis fugax.
(...)
Her current medical regimen includes **losartan** 75 mg p.o. b.i.d., **nifedipine** XL 60 mg p.o. q.d., **aspirin** 325 mg p.o. q.d., **atorvastatin** 40 mg p.o. q.d., **metoprolol** 12.5 mg p.o. b.i.d., (...)
On examination she appears well. Weight is 142 pounds. **Blood pressure is 140/60**. Heart rate is 50 and regular.
(...)
Her electrocardiogram shows sinus bradycardia with first degree AV block and findings suggestive of a **prior anterolateral myocardial infarction**.

FIGURE 1 – Extrait de dossier patient. Les éléments en gras constituent des informations essentielles pour déterminer les facteurs de risque de développement de maladies cardio-vasculaires pour des patients diabétiques

Le corpus d'entraînement contient 790 documents (178 patients) tandis que le corpus de test se compose de 514 documents (118 patients). Les patients sont différents entre les deux corpus. Nous avons segmenté aléatoirement le corpus d'entraînement en sous-corpus d'apprentissage (390 documents pour 89 patients) pour la mise au point de nos méthodes, sous-corpus de développement (131 documents pour 30 patients) pour configurer le système, et sous-corpus de test interne (269 documents pour 59 patients) pour évaluer le résultat de nos méthodes. Nous représentons sur la figure 2 la répartition initiale des documents entre corpus d'entraînement et corpus de test officiel, fournis par les organisateurs, et la segmentation que nous avons faite du corpus d'entraînement en sous-corpus d'apprentissage, de développement et de test interne.

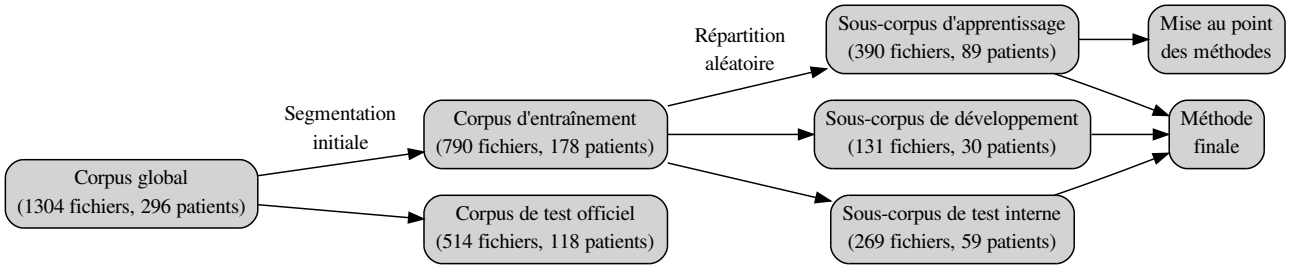

FIGURE 2 – Répartition des documents dans les différents corpus (entraînement et test officiel), et segmentation du corpus d'entraînement en sous-corpus d'apprentissage, de développement et de test interne

Nous donnons dans le tableau 2 des exemples d'événements médicaux à extraire pour chacune des huit catégories d'information, ainsi que les valeurs des informations associées (voir tableau 1 pour les différents types et valeurs possibles d'informations associées propre à chaque catégorie).

Phrase	Catégorie	Information associée	Temporalité
He was admitted to the hospital for *BG's of 400's*	Diabetes	*high glucose* (indicateur)	before
The patient is noted to have a history of mixed systemic conditions including *diabetes*, *coronary artery disease*, depressive disorder...	Diabetes	*mention* (indicateur)	before, during, after
	CAD	*mention* (indicateur)	before, during, after
pt had dissection and thus *2cd stent was placed*	CAD	*event* (indicateur)	during
She has occasional episodes of *angina*	CAD	*symptom* (indicateur)	before
Father: extensive CAD, with first MI in 50 s	Family History	*present* (indicateur)	—
Her HCL is still 36 and *LDL 118*	Hyperlipidemia	*high LDL* (indicateur)	before
The patient demonstrates a blood pressure of *146/88*	Hypertension	*high bp* (indicateur)	during
Medications on admission: *ASA*, *Lipitor* 20, *Lopressor* 50 bid	Medication	*aspirin* (classe)	before, during, after
	Medication	*statin* (classe)	before, during, after
	Medication	*beta blocker* (classe)	before, during, after
Vital signs: weight 241 lb, *BMI 37.8*	Obese	*BMI* (indicateur)	before, during, after
The patient *denies active tobacco* or alcoholic usage	Smoker	*never* (statut)	—

TABLE 2 – Exemples d'événements médicaux et informations associées (indicateur, classe pharmacologique, statut, temporalité) à extraire. Les passages en italiques désignent les indices permettant d'identifier ou d'inférer les événements

5 Méthodes

L'approche que nous avons retenue repose sur les étapes suivantes :
– Puisque les expressions temporelles constituent une information importante dans la tâche, nous commençons par identifier les expressions temporelles, nous les normalisons, et les réutilisons dans les étapes suivantes ;
– L'identification des maladies, des facteurs de risque et des médicaments se fait en trois étapes : (i) un pré-traitement du texte afin de réaliser une représentation vectorielle des caractéristiques de chaque token du document, (ii) une classification supervisée pour détecter les facteurs de risques directement mentionnés, ainsi que les noms de traitements médicaux, et (iii) un post-traitement à base de règles pour identifier certains facteurs de risque supplémentaires, tels que les résultats de laboratoire dont les valeurs supérieures à certains seuils déclenchent l'identification d'un facteur (par exemple, une pression sanguine supérieure à 140/90 mm/hg est signe d'hypertension et doit être identifiée).
– Enfin, l'identification du positionnement temporel est réalisé au moyen de règles produites manuellement.
Le schéma 3 décrit la succession de ces différentes étapes.

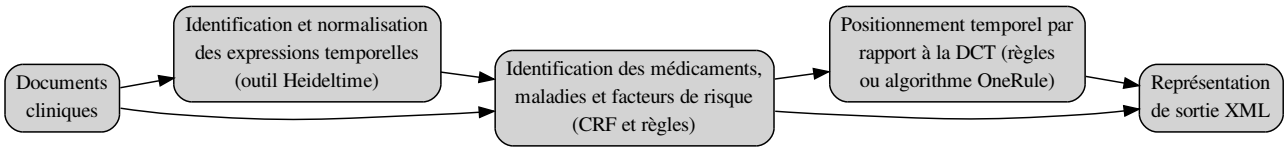

FIGURE 3 – Enchaînement des étapes suivies pour identifier les médicaments, maladies, facteurs de risque et déterminer le positionnement temporel de ces éléments par rapport à la date de création du document (DCT)

5.1 Normalisation des expressions temporelles

Nous avons utilisé l'outil à base de règles HeidelTime (Strötgen & Gertz, 2013) pour identifier les expressions temporelles absolues et relatives contenues dans les documents cliniques. La normalisation des expressions temporelles repose à la fois sur des lexiques et sur le temps des verbes présents dans la phrase de l'expression temporelle. Cette normalisation se fait en référence à la date de création du document (DCT). Dans les documents cliniques du corpus, nous avons pris pour référence la date introduite par la mention *Record date* présente dans chaque document.

Nous avons adapté l'outil aux caractéristiques du corpus (Moriceau & Tannier, 2014), d'une part en ajoutant une dizaine

de règles, notamment pour gérer certains formats particuliers de dates (M/JJ ou M/AA), et d'autre part en définissant des déclencheurs *(past, last, next, ago)* qui vont permettre de situer temporellement des durées qui ne peuvent être normalisées au format JJ-MM-AAAA. Par exemple, l'expression "2 weeks ago" ("il y a 2 semaines") est normalisée par HeidelTime en *P2W* et ce format de normalisation, qui n'indique pas si la durée est passée ou future, ne permet pas de situer directement cette expression par rapport à la DCT : la présence de tels déclencheurs permet alors de le faire.

Une fois réalisée la normalisation de toutes les expressions temporelles, nous avons calculé les relations temporelles de chaque expression (avant, pendant, après) en comparant la valeur normalisée de ces expressions avec la DCT.

5.2 Identification des maladies, médicaments et facteurs de risque

Afin d'identifier les maladies, médicaments et facteurs de risques, nous avons construit une chaîne de traitements reposant principalement sur une approche par apprentissage, complétée par des règles de post-traitements. Nous avons ainsi utilisé l'outil Wapiti (Lavergne *et al.*, 2010) fondé sur le formalisme des champs aléatoires conditionnels (CRF) (Sutton & McCallum, 2006). De manière à comparer les résultats obtenus par cette approche, nous avons constitué une approche basique (*baseline*) reposant sur la projection, sur les corpus de test, des observations effectuées sur les corpus d'entraînement.

5.2.1 Approche par apprentissage statistique

Nous avons construit nos modèles CRF en nous fondant sur les caractéristiques suivantes :
- **Caractéristiques lexicales :** le token ;
- **Caractéristiques typographiques :**
  - longueur du token en nombre de caractères,
  - casse typographique du token,
  - présence de signes de ponctuation dans le token,
  - présence de chiffres dans le token ;
- **Caractéristiques morpho-syntaxiques :** l'étiquette en partie du discours du token telle que fournie par l'outil Tree Tagger (Schmid, 1994) ;
- **Caractéristiques sémantiques :**
  - si le token est un nom de médicament, la classe pharmacologique du token d'après une liste constituée à partir des annotations présentes dans le corpus d'entraînement (voir tableau 1a pour les 17 classes pharmacologiques utilisées) ;
  - la normalisation des expressions temporelles identifiées dans la même phrase que le token, telle que fournie par l'outil HeidelTime ;
- **Caractéristique de structure :** la section dans laquelle apparaît le token, parmi 21 sections manuellement définies d'après les structures les plus fréquemment observées en corpus (*allergies, assessment and plan, chief complaint, family history, history of present illness, medications, physical exam, review of system, vital signs, social history,* etc.).

Pour certaines caractéristiques (token, casse typographique, partie du discours), nous avons également défini des bigrammes de caractéristiques. Nous n'avons réalisé aucune validation croisée pour construire notre modèle. Nous avons cependant configuré la pénalité laplacienne $l1$ implémentée dans l'outil Wapiti de manière à réduire le sur-apprentissage des catégories les plus représentées dans le corpus.

Nous avons défini deux modèles CRF. Le premier prend en compte la normalisation des expressions temporelles identifiées par HeidelTime dans la même phrase que le token (modèle CRF complet), tandis que le deuxième ne tient pas compte de cette normalisation (modèle CRF simplifié). Les modèles que nous avons appliqués sur le sous-corpus de test interne ont été constitués à partir du sous-corpus d'apprentissage (390 documents) lors de l'étape de mise au point des méthodes, alors que les modèles appliqués sur le corpus de test officiel ont été constitués sur l'ensemble du corpus d'entraînement (790 documents, regroupant sous-corpus d'apprentissage, de développement et de test interne) une fois la méthode finalisée (voir figure 2).

5.2.2 Règles de post-traitement

Puisque le système CRF permet principalement d'identifier les facteurs de risque directement mentionnés dans le texte, nous avons complété notre approche par une douzaine de règles de post-traitement de manière à identifier les facteurs de risque représentés sous la forme de résultats de laboratoire. Dans ce dernier cas, seules les valeurs supérieures à des seuils

prédéfinis par les médecins constituent effectivement un facteur de risque et doivent être identifiées comme tel (les valeurs inférieures à ces seuils sont considérées comme normales et ne constituent donc pas des facteurs de risque, voir section 3).

En ce qui concerne la présence de maladies coronaro-artérielles dans la famille du patient, une étude du corpus nous a permis de constater un trop faible nombre de cas pour que nous puissions les traiter de manière efficace. Nous avons donc choisi de systématiquement considérer qu'il n'existe pas d'histoire familiale de ce type de maladie.

5.3 Calcul du positionnement temporel par rapport à la DCT

5.3.1 Approche à base de règles

Afin de calculer le positionnement temporel de chaque élément médical précédemment identifié par rapport à la date de création du document, nous avons réalisé une étude statistique du corpus d'entraînement. Nous avons pu observer que le positionnement temporel dépend à la fois de la catégorie et de l'indicateur associé au facteur de risque. Sur cette base, nous avons défini cinq règles générales qui nous permettent de traiter rapidement le positionnement temporel des événements médicaux : (i) pour *Medication*, les trois valeurs ("before", "during", "after") sont systématiquement associées ; (ii) pour *CAD, Diabetes, Hyperlipidemia, Hypertension, Obese*, la valeur "before" est sélectionnée dans tous les cas ; (iii) pour les cinq mêmes types d'événements, la valeur "during" est sélectionnée dans tous les cas sauf pour la catégorie *Hyperlipidemia* si "indicator" a pour valeur *high chol.* ; (iv) pour ces cinq événements encore, la valeur "after" est associée pour certaines valeurs seulement de l'attribut "indicator" [9] ; et (v) pour la catégorie *Smoker*, si le statut n'a pas été prédit par le CRF, nous associons la valeur "unknown" par défaut.

5.3.2 Approche par apprentissage statistique

Nous avons également défini une deuxième approche, inspirée de celle élaborée lors de notre participation à la campagne d'évaluation ShARe/CLEF eHealth 2014 (Hamon *et al.*, 2014) pour affecter des valeurs d'attribut temporel aux concepts médicaux présents dans les documents cliniques. Cette méthode repose sur un apprentissage supervisé, fondé sur la distribution des relations dans les différentes sections du documents, complétée par l'utilisation de caractéristiques déterminantes décrites plus bas. Pour traiter la classification multi-labels, nous prenons comme classes cibles la concaténation des valeurs temporelles présentes pour un événement (par exemple, "before+during+after" si un événement est associé en même temps à ces trois valeurs). Nous avons testé plusieurs algorithmes (arbres de décision, Naïve Bayes, OneRule) implémentés dans l'outil Weka (Hall *et al.*, 2009) et avons retenu l'algorithme OneRule en raison des bons résultats produits sur le corpus d'entraînement.

Puisque la tâche définit six facteurs de risque pour lesquels une relation temporelle doit être identifiée *(CAD, Diabetes, Hyperlipidemia, Hypertension, Medications, Obese)*, nous avons créé un modèle distinct pour chacun de ces six facteurs. Les catégories *Smoker* et *Family history* n'impliquant pas un positionnement temporel des événements par rapport à la date de création du document [10] (voir tableau 2), nous ne les traitons pas ici. Les six modèles créés reposent sur les caractéristiques suivantes :

Information associée aux événements médicaux : classe pharmacologique ou valeur de l'indicateur ;

Informations de structure :

- position relative d'un événement dans le texte découpé en cinq blocs de taille égale (*relative_position*$= 0 \ldots 4$) ;
- section dans laquelle l'événement est identifié (*section_type*, selon les 21 sections manuellement définies).

La structure d'un document est modélisée au travers de sections et de positions relatives. Une étude du corpus d'entraînement nous a permis de mettre en évidence l'existence d'une corrélation entre les sections et la distribution des positionnements temporels. D'autre part, le positionnement "before" intervient souvent au début du document tandis que le positionnement "after" apparaît davantage vers la fin des documents. Nous avons donc utilisé ces informations de structure lors de la construction de nos modèles.

Les règles apprises par le classifieur OneRule sont indiquées dans le tableau 3 : pour chaque événement, le tableau présente l'attribut sélectionné pour déterminer le positionnement temporel de ce type d'événement, puis pour chaque valeur possible de cet attribut, la décision de positionnement retenue.

9. Pour *CAD* : event, mention, symptom ; *Diabetes* : mention ; *Hyperlipidemia* : mention ; *Hypertension* : mention ; et pour *Obese* : BMI et mention.

10. Dans le cadre de la campagne i2b2 2014, l'information de tabagisme (*Smoker*) n'est pas positionnée par rapport à la date de création du document. En revanche, elle est inscrite dans le temps par le biais du *statut* du tabagisme parmi cinq valeurs *(en cours, passé, toujours, jamais, inconnu)*.

Événement	Attribut : *Valeurs possibles*	Positionnement temporel
CAD	indicator :	
	mention	before, during, after
	event, symptom, test	before
Diabetes	indicator :	
	mention	before, during, after
	A1C, glucose	before
Hyperlipidemia	indicator :	
	high chol., high LDL	before
	mention	before, during, after
Hypertension	indicator :	
	high bp	during
	mention	before, during, after
Medication	section_type :	
	Allergies, Discharge, HPI, Medications, Medications_On_Admission, Past_Medical_History, Plan, Problems, Subjective	before, during, after
	Assessment, Brief_Hospital_Course, Chief_Complaint, Conclusions, Consultations, Diagnosis, Family_History, Follow_Up, General, Hospital_Course, Impression_and_Plan, Interpretation, Major_Surgical_Or_Invasive_Procedure, Microbiology, Objective, Patient_Test_Information, Pertinent_Results, Physical_Examination, _Prologue_, Reason_For_This_Examination, Review_of_Systems, Social_History, Underlying_Medical_Condition, Vital_Signs	after
Obese	relative_position :	
	0, 1, 2	before, during, after
	3, 4	during

TABLE 3 – Règles apprises par le classifieur OneRule pour le calcul du positionnement temporel

5.4 Configurations des expériences

Nous avons défini trois configurations expérimentales distinctes de manière à tester différentes hypothèses de travail. Nous résumons ces configurations dans le tableau 4.

Config	Identification des maladies, facteurs de risque et traitements	Calcul de la temporalité	Hypothèse testée
S-1R	Modèle CRF simplifié et règles de post-traitement	OneRule	Approche statistique de base
C-1R	Modèle CRF complet et règles de post-traitement	OneRule	Les informations temporelles apportent une information utile
C-E	Modèle CRF complet et règles de post-traitement	Règles empiriques	Le calcul de la temporalité par des règles améliore la précision

TABLE 4 – Résumé des trois configurations expérimentales et hypothèse testée

La figure 4 représente les configurations testées en mettant en évidence les approches utilisées pour identifier les maladies, facteurs de risque et traitements (boîtes sur fond rouge) et pour calculer la temporalité (boîtes sur fond bleu). Sur ce schéma, la boîte de l'outil HeidelTime permet de rappeler en quoi diffère la construction des deux modèles CRF.

6 Résultats

Nous indiquons dans le tableau 5 les résultats que nous avons obtenus, sur les sous-corpus de développement et de test interne (modèles CRF construits sur le corpus d'apprentissage de 390 documents) et sur le corpus de test officiel (modèles

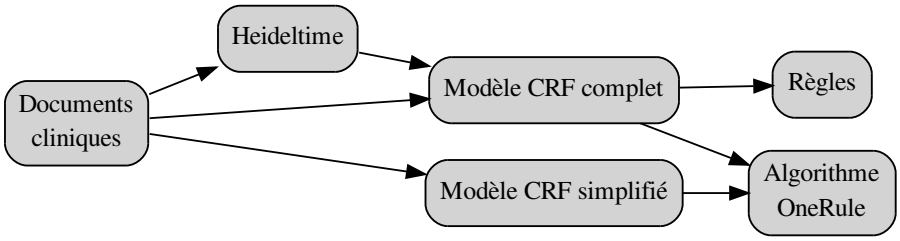

FIGURE 4 – Représentation graphique des trois configurations expérimentales

CRF construits sur l'ensemble du corpus d'entraînement, soit 790 documents). Les résultats sont donnés en terme de micro F-mesure, calculés par le script d'évaluation fourni par les organisateurs de la campagne d'évaluation.

Corpus	Test interne (269 docs)				Test officiel (514 docs)						
Modèle CRF	Corpus d'apprentissage (390 docs)				Corpus d'entraînement (790 docs)						
Expérience	Baseline	S-1R	C-1R	C-E	Baseline	S-1R				C-1R	C-E
Précision	0,6206	0,7609	0,7609	0,8037	0,6503	0,9057				0,9069	0,8753
Rappel	0,8849	0,9445	0,9445	0,9484	0,9005	0,7922		DCT		0,7621	0,7689
F-mesure	0,7295	**0,8428**	**0,8428**	0,8700	0,7552	**0,8451**	before	during	after	0,8282	0,8187
CAD	0,4742	0,5558	0,5558	0,6015	0,4500	**0,7387**	0,6625	0,7637	0,8607	0,6971	0,6387
Diabetes	0,7580	0,9158	0,9158	0,9257	0,7679	**0,8996**	0,8826	0,8921	0,9281	0,8689	0,8528
Family_Hist	0,0131	1,000	1,000	1,000	0,9630	0,9630	—	—	—	0,9630	0,9630
Hyperlipidemia	0,7749	0,8386	0,8386	0,8639	0,7955	**0,8315**	0,8043	0,8416	0,8517	0,8199	0,8167
Hypertension	0,8668	0,8399	0,8399	0,9212	0,8806	0,9172	0,9139	0,9012	0,9444	**0,9190**	0,8753
Medication	0,8128	0,8927	0,8927	0,9050	**0,8401**	0,8389	0,8435	0,8353	0,8378	0,8208	0,8208
Obese	**0,8013**	0,5536	0,5536	0,7440	**0,8259**	0,6991	0,6331	0,8046	0,6331	0,6800	0,7817
Smoker	0,5838	0,7514	0,7514	0,7454	0,5857	**0,7237**	—	—	—	0,7083	0,7096

TABLE 5 – Résultats globaux et détaillés (micro mesures) sur le sous-corpus de test interne et sur le corpus de test officiel, pour chacune des trois configurations expérimentales envisagées (colonnes nommées S-1R, C-1R, C-E, voir tableau 4). L'évaluation du positionnement temporel est fournie pour l'expérience produisant les meilleurs résultats. Les meilleurs résultats sont représentés en gras

7 Discussion

Nous obtenons nos meilleurs résultats au moyen de la première configuration, fondée sur la combinaison du CRF avec les règles de post-traitements pour le calcul du positionnement temporel. Alors que nous avons obtenu de meilleurs résultats sur le sous-corpus de développement au moyen de la deuxième configuration (c.-à-d. en prenant les normalisations d'expressions temporelles fournies par HeidelTime comme caractéristiques pour construire notre modèle CRF), cette configuration s'est révélée la moins efficace sur le corpus de test officiel. Cette différence peut s'expliquer : (i) par le fait que le modèle CRF utilisé pour identifier les facteurs de risque n'a pas été construit sur le même ensemble de fichiers (390 documents pour le test interne vs. 790 documents pour le test officiel), et/ou (ii) parce que les propriétés présentes dans le sous-corpus de test interne et le corpus de test officiel ne se retrouvent pas selon les mêmes distributions, malgré la répartition aléatoire des documents lors de la constitution de nos sous-corpus de travail (voir section 4).

Contrairement aux résultats obtenus sur le sous-corpus de développement, l'utilisation des normalisations fournies par l'outil HeidelTime n'est pas pertinente pour le corpus de test officiel ($F_{C\text{-}1R} = 0,8282 < F_{S\text{-}1R} = 0,8451$). De manière similaire aux résultats obtenus sur le sous-corpus de développement, le calcul des relations temporelles permet d'améliorer les résultats ($F_{C\text{-}E} = 0,8187 < F_{C\text{-}1R} = 0,8282$).

Dans le détail, notre approche permet de traiter efficacement les facteurs de risque relatifs à deux catégories : *hypertension* (F=0,9190) et *diabète* (F=0,8996). Ces résultats s'expliquent par le nombre élevé de mentions pour ces deux catégories

dans le corpus d'entraînement, ce qui permet de produire des modèles CRF, pour la prédiction des facteurs de risque, et OneRule, pour la prédiction du positionnement temporel, particulièrement robustes. Le résultat élevé (F=0,9630) pour le facteur *family history of CAD* n'est pas pertinent dans la mesure où nous n'avons pas traité ce facteur de risque et avons simplement utilisé la valeur par défaut "not present" sur chacun des documents cliniques. Nous obtenons nos moins bons résultats sur les catégories *obese* (F=0,6991, jusqu'à 0,7817 sous la troisième configuration) et *smoker* (F=0,7237).

En ce qui concerne le calcul de la valeur du positionnement temporel associé à chaque facteur de risque, nous observons que les résultats par valeurs temporelles ("before", "during", "after") divergent selon les facteurs de risque : (i) pour les médicaments, les résultats sont homogènes entre les trois valeurs temporelles ; (ii) pour l'obésité, nous obtenons de meilleurs résultats avec la relation "during", et (iii) pour les autres facteurs de risque, nous réalisons de meilleures performances sur la relation "after". Une deuxième observation concernant le facteur d'obésité est que nous obtenons une F-mesure plus élevée avec la configuration C-E (F=0,7817), c'est-à-dire en calculant la temporalité au moyen de règles. Cela signifie que le modèle OneRule n'est pas efficace sur ce facteur puisqu'il conduit à dégrader les résultats de 10 points de F-mesure : l'information de position relative dans le texte est moins prédictive que les règles définies empiriquement. La comparaison des configurations C-1R et C-E pour les autres facteurs montre cependant que les modèles OneRule (C-1R) sont plus efficaces pour les catégories *CAD* (+5.8pt), *Hypertension* (+4.4pt), et *Diabetes* (+1.6pt).

Sur l'identification des facteurs de risque, nous obtenons des résultats différents en fonction de la manière dont sont représentés ces facteurs dans les documents. Nous avons ainsi mieux identifié les mentions que les informations devant être inférées de valeurs numériques supérieures à des seuils prédéterminés. Pour le facteur *hyperlipidemia*, nous obtenons par exemple des F-mesures de 0,8517 sur les "mention" (711 entités dans le corpus de test), de 0,4444 sur les valeurs "high cholesterol" (un taux de cholestérol total supérieur à 240 mg/dL, soit seulement 11 entités), et de 0,2941 pour les valeurs "high LDL" (un taux de cholestérol LDL supérieur à 100 mg/dL, soit 29 entités).

8 Conclusion

Dans cet article, nous avons présenté les expériences que nous avons menées pour répondre à la double problématique de (i) l'identification des facteurs de risque de développement de maladies cardio-vasculaires pour des patients diabétiques et (ii) de positionnement de ces facteurs de risque par rapport à la date de consultation, depuis les informations exprimées dans des documents cliniques. Sur l'identification des facteurs de risque, notre approche repose sur un système par apprentissage statistique (CRF) complété par des règles de post-traitements. Nous avons déterminé le positionnement temporel des éléments précédemment identifiés vis à vis de la date de création du document (DCT) au moyen d'un ensemble de six règles. L'enchaînement de ces deux étapes nous permet d'obtenir une micro F-mesure globale de 0,8451 sur le corpus de test. Il est difficile de dire à quelle distance ce score est de l'optimum atteignable par rapport au corpus d'entraînement dans la mesure où les meilleurs systèmes (F=0,9277) ont été entraînés par des équipes qui ont d'abord complété les annotations de ce corpus. Un travail de réannotation de corpus constitue une étape utile au vu de ces résultats.

Le calcul de la temporalité par une méthode d'apprentissage, même simple comme OneRule, conduit à de meilleurs résultats que les cinq règles définies empiriquement. Nous comptons tester une autre approche du problème de classification multi-labels en entraînant séparément un classifieur pour chaque positionnement temporel. Nous envisageons également d'adopter des méthodes différentes pour le traitement de certains facteurs comme le statut de tabagisme, en considérant qu'il s'agit également d'une tâche de classification que la présence d'indices dans les documents pourrait aider à résoudre.

Remerciements

Ce travail a été financé par l'Agence Nationale de Sécurité du Médicament (ANSM) dans le cadre du projet Vigi4MED (Vigilance dans les forums sur les Médicaments) ANSM-2013-S-060 et par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet Accordys (Agrégation de Contenus et de COnnaissances pour Raisonner à partir de cas dans la DYSmorphologie fœtale) ANR-12-CORD-0007-03.

Références

CHANG A. X. & MANNING C. D. (2012). SUTime : a library for recognizing and normalizing time expressions. *Language Resources and Evaluation*.

CHAPMAN W. W., BRIDEWELL W., HANBURY P., COOPER G. F. & BUCHANAN B. G. (2001). A simple algorithm for identifying negated findings and siseases in discharge summaries. *J Biomed Inform*, **34**(5), 301–10.

CHERRY C., ZHU X., MARTIN J. & DE BRUIJN B. (2013). A la recherche du temps perdu : extractiong temporal relations from medical text in the 2012 i2b2 NLP challenge. *J Am Med Inform Assoc*, **20**(5), 843–48.

CHILDS L. C., TAYLOR R. J., SIMONSEN L., HEINTZELMAN N. H., KOWALSKI K. M. & TAYLOR R. J. (2008). Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. *J Am Med Inform Assoc*, **16**(4), 571–5.

CLARK C., GOOD K., JEZIERNY L., MACPHERSON M., WILSON B. & CHAJEWSKA U. (2008). Identifying smokers with a medical extraction system. *J Am Med Inform Assoc*, **15**(1), 36–9.

CORMACK J., NATH C., MILWARD D., RAJA K. & JONNALAGADDA S. (2014). Agile text mining for the i2b2 2014 cardiac risk factors challenge. In *i2b2 Work Proc*, Washington, DC.

DOAN S., BASTARACHE L., KLIMKOWSKI S., DENNY J. C. & XU H. (2010). Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inform Assoc*, **17**(5), 528–31.

HALL M. A., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P. & WITTEN I. H. (2009). The WEKA data mining software : An update. *SIGKDD Explor Newsl*, **11**(1).

HAMON T., GROUIN C. & ZWEIGENBAUM P. (2014). Disease and disorder template filling using rule-based and statistical approaches. In *Working notes of the ShARe/CLEF eHealth Evaluation Lab*.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proc of ACL*, p. 504–13, Uppsala, Sweden.

MORICEAU V. & TANNIER X. (2014). French resources for extraction and normalization of temporal expressions with HeidelTime. In *Proc of LREC*, p. 3239–43, Reykjavik, Iceland.

ROBERTS K., SHOOSHAN S. E., RODRIGUEZ L., ABHYANKAR S., KILICOGLU H. & DEMNER-FUSHMAN D. (2014). NLM : Machine learning methods for detecting risk factors for heart disease in EHRs. In *i2b2 Work Proc*, Washington, DC.

SAEED M., VILLAROEL M., REISNER A. T., CLIFFORD G., LEHMAN L.-W., MOODY G., HELDT T., KYAW T. H., MOODY B. & MARK R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) : A public-access intensive care unit database. *Crit Care Med*, **39**(5), 952–60.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language*.

STRÖTGEN J. & GERTZ M. (2010). Heideltime : high quality rule-based extraction and normalization of temporal expressions. In *Proc of SemEval*.

STRÖTGEN J. & GERTZ M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, **47**(2), 269–298.

STUBBS A., KOTFILA C., XU H. & UZUNER O. (2014a). Practical applications for NLP in clinical research : the 2014 i2b2/UTHealth shared tasks. In *Proc of i2b2/UTHealth NLP Challenge*.

STUBBS A., UZUNER O., KUMAR V. & SHAW S. (2014b). *Annotation guidelines : risk factors for heart disease in diabetic patients*. i2b2/UTHealth NLP Challenge.

SUN W., RUMSHISKY A. & UZUNER O. (2013). Temporal reasoning over clinical text : the state of the art. *J Am Med Inform Assoc*, **20**(5), 814–9.

SUTTON C. & MCCALLUM A. (2006). An introduction to conditional random fields for relational learning. In L. GETOOR & B. TASKAR, Eds., *Introduction to Statistical Relational Learning*. MIT Press.

TORII M., WEI FAN J., LI YANG W., LEE T., WILEY M. T., ZISOOK D. & HUANG Y. (2014). De-identification and risk factor detection in medical records. In *i2b2 Work Proc*, Washington, DC.

UZUNER O. (2009). Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc*, **16**(4), 561–70.

UZUNER O., GOLDSTEIN I., LUO Y. & KOHANE I. (2008). Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc*, **15**(1), 14–24.

UZUNER O., SOLTI I. & CADAG E. (2010). Extracting medication information from clinical text. *J Am Med Inform Assoc*, **17**(5), 514–518.

VERHAGEN M., MANI I., SAURI R., KNIPPEN R., JANG S. B., LITTMAN J., RUMSHISKY A., PHILLIPS J. & PUSTEJOVSKY J. (2005). Automating temporal annotation with TARSQI. In *Proc of ACL, Interactive Poster and Demonstration Sessions*, Stroudsburgh, PA.