Étude des risques de réidentification des patients à partir d'un corpus désidentifié de comptes-rendus cliniques en français

Cyril Grouin[1] Nicolas Griffon[2, 3] Aurélie Névéol[1]
(1) CNRS, LIMSI, UPR 3251, Campus universitaire d'Orsay, rue John von Neumann, 91405 Orsay
(2) INSERM, LIMICS, UMR_S 1142, 15 rue de l'École de Médecine, 75006 Paris
(3) CISMeF-TIBS-LITIS EA 4108, CHU de Rouen 76031 Rouen
prenom.nom@limsi.fr, prenom.nom@chu-rouen.fr

Résumé. La désidentification permet de préserver le secret médical lors de l'utilisation de documents cliniques pour faire avancer la recherche médicale. Cet article présente une évaluation des risques de réidentification des patients sur un corpus désidentifié de comptes-rendus cliniques en français. Les informations identifiantes sont marquées automatiquement dans le corpus, puis remplacées par des substituts plausibles. Les documents ainsi désidentifiés sont présentés à six évaluateurs avec une connaissance variable des documents et de la méthode de désidentification employée, afin qu'ils réidentifient les patients. La quantité d'informations identifiantes retrouvées semble liée à la familiarité des évaluateurs avec les documents et la méthode de désidentification. L'introduction de substituts géographiques de la même provenance que les documents originaux semble mieux préserver la confidentialité. Les informations retrouvées par les évaluateurs ne permettent pas de réidentifier les patients, sauf en cas d'accès privilégié au système d'information hospitalier de l'établissement d'origine des documents.

Abstract.

Chance of reidentification of patients from a de-identified corpus of clinical records in French
De-identification aims at preserving patient confidentiality while enabling the use of clinical documents for furthering medical research. Herein, we evaluate patient re-identification risks on a corpus of clinical documents in French. Personal Health Identifiers are automatically marked by a de-identification system applied to the corpus, followed by reintroduction of plausible surrogates. The resulting documents are shown to individuals with varying knowledge of the documents and de-identification method. The individuals are asked to re-identify the patients. The amount of information recovered increases with familiarity with the documents and/or de-identification method. Surrogate re-introduction with localization from the same (vs. different) geographical area as the original documents is more effective. The amount of information recovered was not sufficient to re-identify any of the patients, except when privileged access to the hospital health information system and several documents about the same patient were available.

Mots-clés : Désidentification, réidentification, dossiers médicaux éléctroniques, vie privée.

Keywords: De-identification, Re-identification, Electronic Health Records, Privacy.

1 Introduction

Les recherches fondées sur des données cliniques supposent l'obtention du consentement des patients concernés. En France, les règles et recommandations en matière de respect de la vie privée impliquent que, dans les cas où il est impossible d'obtenir le consentement des patients (patient décédé, difficulté d'identifier les ayant-droits, etc.), les comptes-rendus cliniques doivent être anonymisés pour pouvoir être utilisés à des fins de recherche, en dehors du parcours de soin classique. La désidentification consiste à masquer les informations identifiantes relatives au patient, de telle sorte qu'il n'est plus possible de retrouver l'identité du patient (réidentification) sur la base des informations qui auront été laissées en clair. Les méthodes de désidentification automatique sont souvent évaluées sur leur capacité à identifier des éléments relevant de catégories prédéfinies depuis des comptes-rendus cliniques (Meystre *et al.*, 2010). Aux États-Unis, dix-huit catégories ont été définies dans le cadre de la loi HIPAA [1] de 1996 (US Department of Health Human Services, 1996). En

1. HIPAA : Health Insurance Portability and Accountability Act

l'absence d'une loi Européenne équivalente, nous utilisons ce cadre juridique pour désidentifier les documents cliniques français.

Évaluer le risque de réidentification des patients à partir de documents désidentifiés est une tâche complexe, dans la mesure où la combinaison d'éléments d'information en apparence inoffensifs peut néanmoins remettre en cause le respect de la vie privée du patient (Benitez & Malin, 2010; Barbaro & Zeller Jr, 2006; Grouin, 2013). La création de corpus de comptes-rendus cliniques réalistes a été réalisée avec succès (Neamatullah *et al.*, 2008) en enchaînant deux systèmes, un premier système d'identification des éléments à désidentifier dans les comptes-rendus cliniques suivi d'un deuxième système de remplacement des informations précédemment identifiées par des éléments fictifs plausibles. Le résultat produit un corpus valable sur les plans cliniques et linguistiques.

Alors que l'impact de la désidentification sur des traitements ultérieurs (étiquetage en parties du discours, extraction d'information, repérage d'entités nommées, etc.) a été étudié sur un corpus de comptes-rendus cliniques (Deléger *et al.*, 2013; Meystre *et al.*, 2014b), il n'existe que peu d'études sur l'impact réel concernant la vie privée des patients. Il a récemment été démontré que les médecins ne sont plus capables de ré-identifier les patients qu'ils ont récemment soignés au-delà d'un délai de trois mois, lorsqu'ils se fondent sur les comptes-rendus cliniques (Meystre *et al.*, 2014a). Il est cependant nécessaire d'aller au-delà de ces premières expériences, pour évaluer la nature et la possibilité des risques de réidentification des patients.

Dans cet article, nous présentons les expériences que nous avons menées en matière d'évaluation des risques de réidentification des patients par des humains, à partir de comptes-rendus cliniques rédigés en français et désidentifiés automatiquement. Les personnes impliquées dans cette étude (chercheurs en informatiques et médecin) présentent différents niveaux de connaissances, tant du point de vue des comptes-rendus cliniques étudiés que de celui des méthodes utilisées pour désidentifier automatiquement les comptes-rendus. Nous avons réalisé nos expériences sur différents types de données (documents relatifs au même patient ou à des patients différents) en faisant également varier les méthodes de réintroduction d'informations fictives (en utilisant soit des informations géographiques similaires à celles d'origine, soit des informations différentes).

2 État de l'art

Toute mise à disposition de données contenant des informations personnelles implique de respecter la vie privée des personnes mentionnées dans les données. Lorsque des documents désidentifiés sont créés à partir de documents réels pour être mis à disposition à des fins de recherche, il est nécessaire d'évaluer les risques de non respect de la vie privée au regard des bénéfices attendus par les résultats de la recherche utilisant ces données.

Lors des premières mises à disposition de données aux États-Unis, une évaluation inadéquate des risques de non respect de la vie privée a conduit à des situations délicates qui ont impliqué des actions en justice (Barbaro & Zeller Jr, 2006). À la lumière de cette expérience, des précautions extrêmes sont désormais requises avant toute mise à disposition de données dites « sensibles ». Le cas des données médicales, notamment celles contenues dans les comptes-rendus hospitaliers, nécessite une attention particulière. Fournir des données médicales qui ne respecteraient pas la vie privée du patient constituerait une violation du secret médical garanti dans le serment d'Hippocrate, le code de déontologie des médecins et le code pénal.

La base de données MIMIC II [2] (Saeed *et al.*, 2002, 2011; Lee *et al.*, 2011) constitue un exemple de réussite en matière de partage de données cliniques à grande échelle [3] respectant la vie privée des patients. En plus d'appliquer une méthode de désidentification performante, les concepteurs de la base de données ont mis en place un accord d'utilisation des données qui impose aux futurs utilisateurs d'être informés de la nature sensible de ces données et de veiller au respect de la protection de la vie privée au cas où ils identifieraient, dans les données fournies, des éléments nominatifs. À notre connaissance, il s'agit de la seule base de données de cette taille disponible pour la recherche clinique ou le traitement automatique des langues. De plus petits jeux de données cliniques ont également été distribués dans des conditions similaires à celles de la base MIMIC lors des campagnes d'évaluations, telle que les campagnes i2b2 dont la première édition en 2006 a porté sur la problématique de la désidentification des comptes-rendus cliniques (Uzuner *et al.*, 2007).

Nous estimons que l'étude des risques de réidentification des patients à partir de données désidentifiées et porteuses d'informations réalistes permet de mieux comprendre l'équilibre bénéfice/risque préalablement à la mise à disposition de

2. MIMIC : Multiparameter Intelligent Monitoring in Intensive Care
3. La version 3 de MIMIC comporte 23 180 dossiers.

données « sensibles ». D'autre part, nous pensons que ce type d'étude peut également contribuer à améliorer la conception et l'évaluation des systèmes de désidentification automatique, jusqu'à présent uniquement évalués en termes quantitatifs.

3 Matériel et méthodes

3.1 Corpus

Nous précisons que le corpus utilisé dans cette étude a reçu une autorisation de la CNIL [4], pour réaliser des recherches en matière de recherche d'information depuis un volume important de comptes-rendus électroniques patients. Dans notre étude, nous avons ciblé douze types d'informations à désidentifier, relatifs aux patients, aux parents des patients, et aux professionnels de santé : *prénoms, noms, initiales, adresses postales, villes, codes postaux, numéros de téléphone et de télécopie, adresses électroniques, noms d'hôpitaux, identifiants* (numéros de sécurité sociale ou numéros de série d'appareillage médical), et *dates* (y compris les dates de naissance) [5]. En matière de documents, nous avons sélectionné les trois types de documents les plus fréquents dans le corpus global d'où proviennent les données : compte-rendu hospitalier, compte-rendus d'acte, correspondance.

La désidentification a été réalisée au moyen d'approches à base d'apprentissage statistique, en appliquant l'outil MEDINA (conçu par l'utilisateur dénommé « Dev 2 » dans la suite de nos expériences). Nous avons construit un modèle statistique CRF (champs aléatoires conditionnels (Lafferty *et al.*, 2001))) adapté aux données à désidentifier sur la base d'un corpus de 100 documents annotés et vérifiés par des humains. L'approche utilisée, réalisée par les utilisateurs « Dev 1 » et « Dev 2 », a été décrite dans (Grouin & Névéol, 2014).

Dans la suite de cet article, nous appelons « données identifiantes réelles » l'ensemble des données présentes dans les documents d'origine qui correspondent aux douze types d'informations précédemment listés. Les « données réelles résiduelles » constituent un sous-ensemble. Elles correspondent aux données d'origine qui n'ont pas été identifiées automatiquement par notre système de désidentification MEDINA et qui n'ont donc pas fait l'objet d'un remplacement par des données fictives réalistes.

3.1.1 Critères d'inclusion

Grâce à l'outil MEDINA, nous avons volontairement constitué un corpus de travail comportant des documents pour lesquels il est fortement probable que des informations personnelles n'aient pas été identifiées, et restent visibles dans les documents transformés. Ainsi, notre étude permet d'évaluer les risques de ré-identification dans un contexte où ce risque peut être considéré comme élevé. Nous avons extrait du corpus désidentifié 60 documents pour lesquels nous savons que des informations ont échappé à l'outil de désidentification. Cette extraction repose sur différents critères jugés difficiles à appréhender pour un outil de désidentification automatique, critères que nous avons établis suite à une analyse des erreurs du système de désidentification :

- **noms, prénoms :** l'outil échoue à identifier des noms complexes, ou des portions de noms complexes, qui intègrent des traits-d'union ou des espaces (*Dorothy Jane, Watterman-Smith*) ;
- **informations de contact :** l'outil échoue également à identifier les informations de contact qui apparaissent dans le contenu même du document (c.-à-d., en dehors des entêtes et pieds-de-page), quand bien même ces informations sont introduites par des déclencheurs (*domicilié, personne de confiance*) ;
- **dates :** l'outil ne permet pas de faire la différence entre les dates liées au patient et les dates plus générales mentionnées dans le document (dates de procédure légale ou de changement de numérotation téléphonique). Appliquer le processus d'antédatation sur ces dates générales (remplacement des dates par des dates situées dans le passé en conservant le même écart pour toutes les dates d'un dossier) peut compromettre le processus global de désidentification puisque l'écart appliqué sur l'ensemble des dates du document permet alors de retrouver les dates d'origine.

4. CNIL : Commission nationale de l'informatique et des libertés `http:www.cnil.fr`

5. Ces douze catégories correspondent majoritairement aux catégories d'origine du HIPAA (US Department of Health Human Services, 1996). Certaines catégories définies par le HIPAA ne se retrouvent pas dans les documents français *(numéros de permis de conduire/carte d'identité, identifiants des véhicules, adresses Internet et adresses IP, identifiants biométriques, photographie)*. D'autre part, nous avons pris en compte les catégories *initiales* (des médecins) et *noms d'hôpitaux*, hors HIPAA, car nous considérons que ces informations, si elles ne permettent pas une réidentification directe, permettent néanmoins de réduire la population aux seuls patients traités par un ensemble donné d'hôpitaux.

3.1.2 Hypothèses d'évaluation du risque de réidentification

Les informations personnelles identifiées automatiquement ont été remplacées par des données fictives réalistes. Parce que l'outil de repérage automatique des informations à désidentifier aura manqué certaines de ces informations, en l'absence de vérification humaine, il reste donc dans le corpus désidentifié des données identifiantes provenant des documents d'origine. L'idée sous-jacente dans la réintroduction de données fictives réalistes repose sur le principe « *caché au vu de tous* ». Nous émettons l'hypothèse que les données identifiantes d'origines seront moins visibles si elles figurent au milieu d'autres données fictives réalistes. Le module de remplacement des données identifiantes par des données fictives a été mis au point par l'un des auteurs (« Dev 2 ») et étendu et adapté au corpus par un autre auteur (« Dev 1 »).

Nous évaluons le risque de réidentification en tenant compte de deux situations différentes, toutes deux pertinentes pour les besoins de la recherche médicale réalisée à partir de comptes-rendus désidentifiés.

Restriction à un ou plusieurs patients Selon l'objectif médical poursuivi dans une étude, il peut être nécessaire d'utiliser un corpus de documents relatifs au même patient (pour étudier la chronologie d'un patient ou l'apparition et l'évolution d'une maladie), par opposition à des documents sélectionnés aléatoirement et appartenants à différents patients. Nous émettons l'hypothèse que le risque de réidentification est plus élevé sur un corpus de documents relatifs à un même patient, dans la mesure où le corpus fournit plus d'informations sur le patient, et qu'il offre également la possibilité de croiser des informations entre documents.

Origine géographique des documents L'outil de réintroduction de données fictives repose sur des listes préétablies pour chaque catégorie d'information à désidentifier. Il est possible de configurer l'outil pour restreindre la zone géographique à un département lors du remplacement des codes postaux, villes et hôpitaux. Nous envisageons deux expériences, l'une réintroduit des données du même département que dans les données d'origine, l'autre réintroduit des données géographiques d'un autre département. Nous émettons l'hypothèse qu'il est plus complexe d'identifier des données réelles identifiantes au milieu de données fictives issues du même département.

3.1.3 Organisation du corpus

En raison de ces différentes configurations, nous avons divisé le corpus en quatre parties :

1. Quinze documents relatifs au même patient, avec réintroduction de données géographiques du même département ;
2. Quinze documents relatifs au même patient, avec réintroduction de données géographiques d'un autre département que celui d'origine ;
3. Quinze documents relatifs à différents patients, avec réintroduction de données géographiques du même département ;
4. Quinze documents relatifs à différents patients, avec réintroduction de données géographiques d'un autre département que celui d'origine.

Nous avons collecté les documents relatifs au même patient de la manière suivante :
– parmi tous les documents d'un patient, nous sélectionnons aléatoirement les documents correspondant à au moins l'un des trois critères de désidentification jugés difficiles (voir section 3.1.1) ;
– puis nous sélectionnons aléatoirement différents documents parmi les précédents documents jusqu'à atteindre le nombre souhaité (15 documents).
Pour le sous-corpus de documents correspondant à plusieurs patients, nous avons appliqué la sélection suivante :
– sélection de trois documents pour chacun des trois critères difficiles ;
– sélection aléatoire des autres documents, qu'ils contiennent ou non des critères difficiles à traiter.
La sub-division entre département identique et département différent de celui d'origine se fait après avoir constitué ces sous-corpus de documents, en substituant, ou non, le numéro de département présent dans les codes postaux du document.

Nous avons réalisé cette sélection aléatoire sans intervention humaine, de telle sorte que les auteurs de l'étude qui ont participé à la phase d'annotation (« Dev 1 » et « Dev 2 ») connaissaient les critères d'inclusion, mais ne savaient quelles informations dans les documents extraits relevaient de cette sélection. Le schéma 1 résume le processus suivi pour constituer les quatre sous-ensembles de corpus utilisés dans le cadre de cette étude.

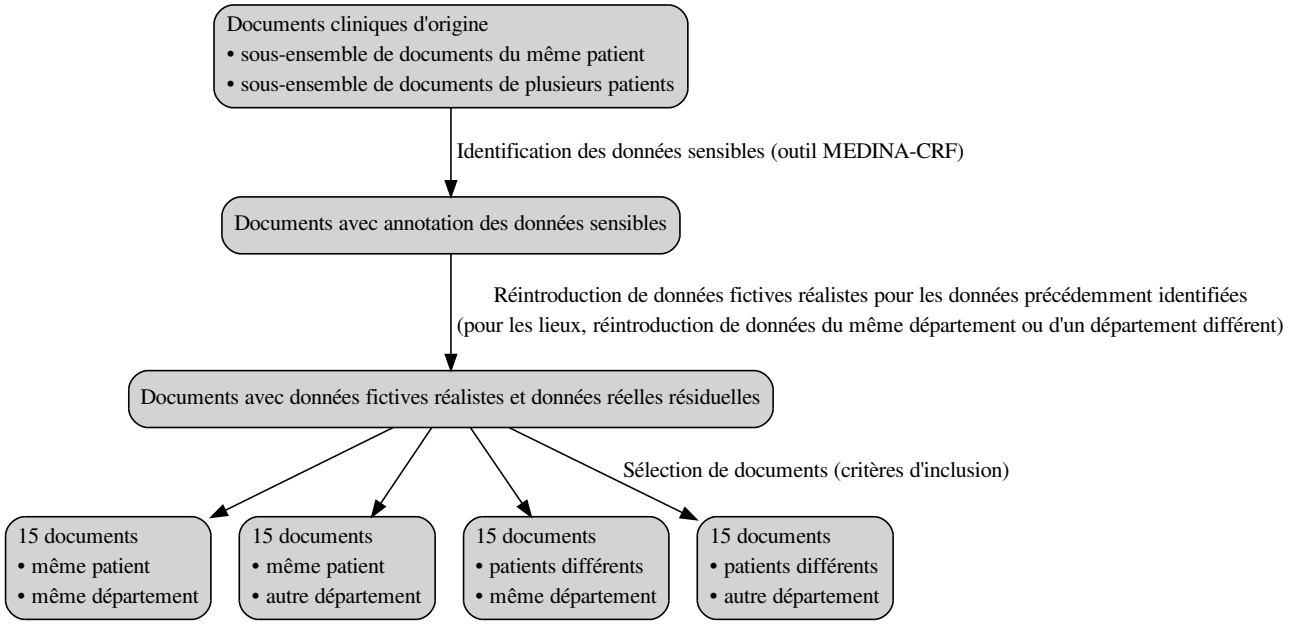

FIGURE 1 – Processus suivi pour constituer les corpus utilisés dans cette étude

3.1.4 Constitution des corpus de références

Les 60 documents du corpus correspondent tous au résultat de l'identification des informations identifiantes par MEDINA avec substitution de ces informations par des informations fictives réalistes non-identifiantes. C'est sur les documents issus de ce traitement que les évaluateurs ont effectué leur travail de detection des données identifiantes originales (non substituées).

Après l'évaluation, deux des auteurs ont examiné l'ensemble des documents du corpus, dans leur version originale et leur version substituée, afin d'identifier toutes les données identifiantes. Deux versions de référence du corpus ont été créées, de manière à évaluer : (i) les performances du système de désidentification automatique (évaluation qui ne peut se faire qu'à partir de la version d'origine du corpus [6]), et (ii) les performances des annotateurs humains sur leur capacité à détecter les données identifiantes réelles (évaluation réalisée sur la version du corpus après réintroduction de données fictives). Si les deux versions du corpus de référence se rapportent aux 60 mêmes documents, la principale différence repose sur le fait que des données fictives ont été introduites en remplacement des données identifiantes réelles pour permettre la diffusion du corpus auprès des annotateurs humains. Parce que les entités ne sont plus les mêmes (*Pierre Fontaine* devient *Paul Martin*) d'une part, et que les coordonnées de début et de fin des portions à traiter changent [7] d'autre part, il n'est plus possible d'utiliser le même corpus de référence pour réaliser les deux évaluations. Nous donnons dans le tableau 1 un exemple de différences entre les deux versions de référence du corpus.

	Texte d'origine (données identifiantes réelles)	Texte avec substitutions (données fictives)
Texte	*Pierre Fontaine né le 03/05/1979*	*Paul Martin né le 01/09/1977*
Annotations de référence (offsets de début et de fin)	Prénom : Pierre (0-6) Nom : Fontaine (7-15) Date : 03/05/1979 (22-32)	Prénom : Paul (0-4) Nom : Martin (5-10) Date : 01/09/1977 (17-27)
Utilisation	Evaluation de l'outil de désidentification automatique MEDINA	Evaluation des annotations humaines

TABLE 1 – Exemples d'annotations de référence sur les deux versions du corpus

6. La version d'origine du corpus contient des données identifiantes réelles.

7. En raison du remplacement des données identifiantes réelles du corpus d'origine par des données fictives, les coordonnées de début et de fin de chaque portion contenant des données à traiter ne sont plus les mêmes. En effet, les offsets de caractères (depuis le début du fichier) diffèrent entre les deux version en raison des différences de taille (nombre de caractères) entre données identifiantes d'origine (*Pierre*, 6 caractères) et données fictives réintroduites (*Paul*, 4 caractères). L'outil utilisé pour l'évaluation (BRATeval) étant sensible aux formes de surface et aux différences d'offsets de caractères, il est donc nécessaire de disposer de deux versions de référence des corpus traités.

Données identifiantes réelles vs. données non-identifiantes fictives Une première version identifie les informations identifiantes réelles que l'outil de désidentification a échoué à identifier, et qui n'ont donc pas été remplacées par des données fictives réalistes (non-identifiantes). Cette version a été obtenue par comparaison des documents avant et après réintroduction des données fictives. Cette version comporte la liste des données identifiantes non retrouvées par MEDINA, ainsi que quelques cas de données identifiantes retrouvées par MEDINA mais remplacées à l'identique. Elle est utilisée pour évaluer la capacité des annotateurs humains à identifier les données identifiantes réelles dans le corpus désidentifié.

Données à désidentifier La deuxième version de référence identifie toutes les données à désidentifier dans le corpus d'origine. Cette version a été obtenue par comparaison du corpus avant et après identification par MEDINA des données à désidentifier. Cette version comporte la liste exhaustive des données identifiantes contenues dans le corpus, et sert à évaluer les performances du système de désidentification MEDINA sur ce corpus.

3.2 Expériences de réidentification

Protocole expérimental Nous avons soumis le corpus à trois catégories d'expérimentateurs humains, ayant différents niveaux de connaissance des documents (donc des informations d'origine susceptibles d'être présentes) et du fonctionnement de l'outil de désidentification (et des limites du systèmes sur certaines catégories) : (i) un médecin de l'hôpital ayant fourni les données d'origine (nommé « Médecin » dans les expériences), (ii) deux chercheurs en informatique ayant conçu et adapté l'outil de désidentification (nommés « Dev 1 » et « Dev 2 »), et (iii) trois autres chercheurs ayant des compétences en informatique ou en linguistique, sans connaissance particulière du corpus ou de l'outil de désidentification (nommés « Chercheur 1 » à « Chercheur 3 »).

Consignes Il a été demandé à chaque expérimentateur d'annoter, dans les documents désidentifiés avec données fictives et présence résiduelle de données identifiantes réelles, toutes les données identifiantes qu'il estimait réelles, donc susceptibles de réidentifier pleinement ou partiellement le patient. Les annotations ont été réalisées au moyen de l'interface d'annotation BRAT (Stenetorp *et al.*, 2012). Pour les annotateurs qui n'étaient familiers, ni avec les données d'origine, ni avec l'outil de désidentification (« Chercheur » 1 à 3), nous leur avons présenté le processus global de constitution du corpus (sections 3.1 et 3.1.4). Contrairement aux autres annotateurs, ces annotateurs n'ont pas eu connaissance du département géographique d'où proviennent les données.

Une fois ce travail d'annotation terminé, nous avons interrogé chaque expérimentateur sur les indices qu'il a estimé utiles pour identifier les données identifiantes réelles. Ces indices sont variables selon les annotateurs et concernent n'importe quel élément ou combinaison d'éléments du texte susceptible d'être réel et permettant une réidentification potentielle. Nous renseignons en section 5.2 de ces tentatives de réidentification et des indices utilisés par les personnes interrogées.

4 Résultats

En moyenne, les expérimentateurs ont mis 2 heures pour traiter l'ensemble du corpus. Le travail d'annotation a ensuite été évalué par rapport aux annotations de référence, mais également en termes d'accord inter-annotateurs (F-mesure), tant sur le corpus global que sur chacun des quatre sous-corpus.

4.1 Performances initiales du système de désidentification

Nous présentons dans le tableau 2 la distribution des données identifiantes (nombre total et nombre de données identifiantes réelles) pour chaque catégorie dans le corpus final de 60 documents. Environ 9,5 % des données sont des données identifiantes réelles, alors que 90,5 % constituent des données fictives qui ont été réintroduites. On observe que la présence résiduelle d'informations identifiantes réelles n'est pas la même selon les catégories d'information, et qu'elle concerne majoritairement les *initiales* (89,7 %) et les *identifiants* (80,0 %), deux catégories qui se rapportent aux médecins mentionnés dans les comptes-rendus. Il en est de même pour les *noms* (3,3 %) et *prénoms* (3,5 %) dont aucun élément ne concerne un patient. Les informations résiduelles des catégories *adresses* (51,7 %), *codes postaux* (17,9 %) et *villes* (25,5 %) renvoient majoritairement aux coordonnées d'hôpitaux, en lien avec les informations résiduelles réelles de la catégorie *hôpitaux* (14,5 %). La catégorie *dates* est particulière, car nous recensons comme "données identifiantes réelles"

les dates qui permettent une identification des dates originales. Il s'agit donc à la fois des dates médicales non substituées (la date d'un examen effectué sur un patient) et des dates administratives de notoriété publique substituées (la date de changement de numérotation téléphonique).

	NOM	**PRE**	**INI**	**ADR**	**VIL**	**CP**	**TEL**	**MAIL**	**ID**	**DATE**	**HOP**	**Total**
Total	541	487	39	60	153	67	282	42	20	233	166	2090
Réelles	18	17	35	21	39	12	0	0	16	17	24	199
	(3,3%)	(3,5%)	(89,7%)	(51,7%)	(25,5%)	(17,9%)	(0,0%)	(0,0%)	(80,0%)	(7,3%)	(14,5%)	(9,5%)

TABLE 2 – Distribution des données identifiantes totales et réelles par catégorie (NOM=noms, PRE=prénoms, INI=initiales, ADR=adresses, VIL=villes, CP=codes postaux, TEL=téléphones, MAIL=e-mails, ID=identifiants, DATE=dates, HOP=hôpitaux) dans le corpus final

Nous rapportons dans le tableau 3 les performances détaillées de l'outil de désidentification MEDINA en termes d'appariements à l'identique avec les données de référence [8]. La performance globale est de 0,93 de F-mesure, pour une précision de 0,96 et un rappel de 0,90, ce qui classe l'outil parmi les systèmes état de l'art du domaine.

Catégorie	**Précision**	**Rappel**	**F-mesure**
Noms	0,97	0,95	0,96
Prénoms	0,98	0,96	0,97
Initiales	0,67	0,05	0,09
Identifiants	1,00	0,25	0,40
Hôpitaux	0,74	0,53	0,62
Adresses	0,98	0,82	0,89
Codes postaux	1,00	0,79	0,88
Villes	0,99	0,95	0,97
Dates	0,94	0,97	0,96
E-mails	1,00	1,00	1,00
Téléphones	0,99	1,00	0,99
Total (micro-moyenne)	0,96	0,90	0,93

TABLE 3 – Performance de l'outil MEDINA, utilisant un modèle CRF appris sur un corpus de 100 documents

4.2 Exemple de document annoté

Nous présentons un extrait document (figure 2) qui a été jugé difficile à désidentifier automatiquement pour le critère "informations de contact" (voir section 3.1.1), en raison de la présence du déclencheur *personne de confiance*. Alors que le numéro de téléphone du mari de la patiente a correctement été détecté par MEDINA, les autres informations relatives à la famille de la patiente n'ont pas été détectées. Sur cet exemple, nous soulignons en vert les données fictives réintroduites, et encadrons en violet les données identifiantes réelles (qui n'ont donc pas été identifiées ni substituées). Nous précisons que les documents présentés aux expérimentateurs n'étaient porteurs d'aucune indication.

Dans cet exemple, les données identifiantes réelles concernent le lieu de résidence des enfants de la patiente : « *2 à Marseille et 1 en Corse* ». Deux annotateurs (« Dev 1 » et « Dev 2 ») ont correctement identifié ces deux informations comme étant réelles, et un annotateur (« Chercheur 1 ») n'a identifié que l'information *Corse*. Sur le base de ses connaissances, l'expérimentateur « Dev 1 » a cru identifier le numéro de téléphone comme étant réel, alors que dans le cas présent il s'agissait bien d'une donnée fictive réaliste.

8. Nous précisons que des données identifiantes d'origine ont pu être correctement identifiées par l'outil de désidentification automatique MEDINA, mais qu'elles auront été substituées par elles-mêmes sous l'effet du tirage au hasard (par exemple, lorsque la substitution substitue un nom de ville « *MARSEILLE* » par lui-même avec une différence de casse typographique « *Marseille* »). Dans ce cas, nous considérons qu'il s'agit toujours d'une donnée identifiante réelle malgré la substitution. En conséquence, il peut exister un écart entre les chiffres présentés dans les tableaux 2 et 3.

nom Huet prénom Arnaud
Née le : date 05/08/1928
Mode de vie :
- Situation familiale : Mariée. 3 Enfants (2 à ville Marseille et 1 en pays Corse).
- Lieu de vie : appartement au 5ème étage avec ascenseur.
- Ancienne profession : secrétaire de direction.

Réseau de soutien :
- Nom personne de confiance : époux N° Tél. : téléphone 06 19 46 13 89
- Aides à domicile : Aide ménagère 3 heures par semaine en chèque emploi service pour le gros ménage.
- APA : non.
ANTECEDENTS :
- Prothèse de hanche.
- Pathologie pancréatique en date 1993.
CONDUITE PROPOSEE :
- Consultation de suivi dans 1 mois pour évaluer la tolérance au traitement (date 26 novembre 2006).

Dr prénom Daniel nom Lucas
Médecin attaché.

FIGURE 2 – Extrait du corpus étudié. Les données fictives réintroduites sont encadrées en vert, les données identifiantes réelles résiduelles sont encadrées en violet (dans cet exemple, elles ont été remplacées par des données fictives)

4.3 Performances individuelles

Le tableau 4 présente les résultats détaillés par catégorie, au niveau global (ligne 6) et par sous-corpus (lignes 2 à 5), pour chaque expérimentateur, classé en fonction des connaissances de chacun sur les données et la méthode de désidentification. Trois groupes peuvent être distingués, séparés par des doubles barres : (i) connaissances avancées à la fois sur les données et sur l'outil de désidentification, (ii) connaissances avancées, soit sur les données, soit sur l'outil de désidentification, et (iii) peu de connaissances sur les données et l'outil.

Corpus	**Dev 1** N - P - R - F	**Médecin** N - P - R - F	**Dev 2** N - P - R - F	**Chercheur 1** N - P - R - F	**Chercheur 2** N - P - R - F	**Chercheur 3** N - P - R - F
1	34 .71 .33 .45	13 .62 .11 .19	285 .16 .64 .26	30 .47 .19 .27	0 .00 .00 .00	26 .00 .00 .00
2	35 .57 .54 .56	11 .64 .18 .29	59 .19 .30 .23	8 .50 .11 .18	66 .02 .03 .02	24 .00 .00 .00
3	31 .61 .40 .49	19 .47 .19 .27	28 .71 .43 .53	6 .33 .04 .08	0 .00 .00 .00	26 .00 .00 .00
4	42 .67 .50 .57	25 .76 .34 .47	41 .51 .38 .43	15 .80 .21 .34	43 .07 .05 .06	10 .00 .00 .00
total	135 .62 .41 .50	66 .61 .20 .30	408 .23 .46 .30	56 .54 .15 .23	109 .04 .02 .03	84 .00 .00 .00

TABLE 4 – Performances sur l'identification des données identifiantes réelles en termes de nombre de données annotées (N), précision (P), rappel (R) et F-mesure (F). Les sous-corpus sont décrits en section 3.1.3 : 1 = même patient, même département, 2 = même patient, autre département, 3 = patients différents, même département, 4 = patients différents, autre département

Le tableau 5 présente les accords inter-annotateurs sur l'identification des données identifiantes réelles, calculés en termes de F-mesure.

	Dev 1	Médecin	Dev 2	Chercheur 1	Chercheur 2
Médecin	0,32	-			
Dev 2	0,21	0,10	-		
Chercheur 1	0,21	0,11	0,18	-	
Chercheur 2	0,00	0,00	0,00	0,01	-
Chercheur 3	0,01	0,01	0,03	0,01	0,00

TABLE 5 – Accords inter-annotateurs (F-mesure)

5 Discussion

5.1 Performances individuelles d'identification des données identifiantes réelles

Le tableau 4 montre que globalement, l'identification de données identifiantes réelles est faible, et que la capacité de l'expérimentateur à détecter ces données est fortement lié aux connaissances antérieures qu'il a des documents et/ou de l'outil de désidentification. La F-mesure la plus élevée est de 0,50 (expérimentateur cumulant connaissances fines sur le corpus et sur l'outil de désidentification), ce qui ne constitue pour autant pas une bonne performance. Alors que les chercheurs 1 à 3 n'avaient aucune connaissance des données et de l'outil, le chercheur 1 a néanmoins deviné l'origine géographique des documents une fois l'expérience terminée. Cependant, personne n'a été en mesure d'identifier plus d'informations spécifiques aux patients.

Le tableau 5 montre que l'accord inter-annotateur le plus élevé est observé entre l'expérimentateur ayant obtenu les meilleurs résultats (« Dev 1 ») et le médecin, autrement dit, les deux expérimentateurs ayant des connaissances fines des données d'origine. L'accord est cependant faible à 0,33 de F-mesure, ce qui ne permet pas de conclure à un accord entre annotateurs (Artstein & Poesio, 2008). Ainsi, même en ayant des connaissances précises des données et/ou de l'outil de désidentification, deux humains ne parviennent pas à se mettre d'accord sur ce qui constitue une donnée identifiante réelle ou non. D'après ces résultats, la stratégie « *caché au vu de tous* » semble correctement fonctionner, et les données identifiantes réelles ne sont pas évidentes pour les différents expérimentateurs impliqués.

5.2 Tentatives de réidentification

Les outils disponibles pour tenter de réidentifier les patients se composent essentiellement des informations disponibles sur internet. L'un des expérimentateur (« Chercheur 1 ») a systématiquement vérifié les noms d'hôpitaux, de personnes et les adresses, en utilisant un moteur de recherche classique, ce qui lui a permis d'identifier l'hôpital d'où proviennent les données. Deux expérimentateurs (« Dev 1 » et « Médecin ») ont utilisé un annuaire inversé pour vérifier tous les numéros de téléphone et adresses qu'ils estimaient être réels. Les requêtes n'ont cependant renvoyé aucun résultat. Les autres annotateurs n'ont pas indiqué avoir utilisé d'autres sources d'informations que celles présentes dans les documents traités.

Un expérimentateur (« Médecin ») a eu accès au système d'information patient (SIP) de l'hôpital d'où sont extraites les données. Sur le sous-corpus mélangeant les documents relatifs à plusieurs patients, il ne lui a pas été possible de réidentifier le moindre patient, faute de pouvoir établir une requête valable dans le système [9] ; des essais de ré-identifications soutenus ont été effectués sur trois documents, et ont été abandonnés au bout de 30 minutes de recherches infructueuses. En revanche, pour les sous-corpus proposant plusieurs documents sur le même patient, sur la base d'un recoupement d'informations entre la date de séjour approximative et les codes d'actes médicaux trouvés dans les documents ou inferrés à partir de connaissances médicales (ces codes cliniques ne constituent pas des informations identifiantes), il a été possible au médecin de constituer une requête valide et de réidentifier correctement le patient. Pour les deux patients du corpus (sous-corpus 1 et 3), les expériences de réidentification dans le SIP ont cependant demandé 20 minutes pour un patient et 30 minutes pour le deuxième avant de pouvoir effectivement retrouver les identités d'origine.

L'outil le plus efficace pour réidentifier un patient se révèle donc être le système d'information de l'hôpital, à condition de disposer des droits d'accès à cet outil et de savoir l'utiliser. Le SIP est conçu de telle sorte que l'utilisateur doit fournir

9. Le système d'information patient est conçu de telle sorte qu'il n'est possible de l'interroger qu'au moyen d'une requête combinant suffisamment d'éléments pour correspondre à un patient enregistré dans la base. Une recherche en plein texte – comme dans un moteur de recherche classique – n'existe pas, car elle ne correspond pas à un besoin des médecins. Aucun outil déployé dans l'hôpital ne permet donc de faire une telle interrogation.

suffisamment d'informations en entrée sur un patient pour pouvoir accéder aux documents de ce patient. Dans notre étude, un seul document désidentifié avec données fictives ne permet pas de retrouver le patient d'origine. À l'inverse, lorsque plusieurs documents sont disponibles pour un même patient, le patient concerné peut être retrouvé. Mais dans ce cas, la réidentification du patient suppose : (i) un accès au SIP, (ii) une connaissance de la manière dont les documents sont codés et stockés dans le SIP, et (iii) des connaissances médicales pour inférer le code des diagnostics à partir des documents, de manière à disposer d'informations complémentaires pour enrichir la requête dans le SIP.

5.3 Performances selon la configuration des expériences

Les résultats présentés dans la tableau 4 montrent que les performances globales sont généralement meilleures lorsque la réintroduction de données fictives se fait sur un département différent de celui des données d'origine (corpus 2 et 4). Cette observation révèle que le principe « *caché au vu de tous* » est conforté puisque les données identifiantes réelles résiduelles auront été « noyées » parmi les données fictives réintroduites sur le même département.

Un argument en défaveur de l'utilisation du même département que celui des données d'origine lors de la réintroduction de données fictives portait sur le fait que, potentiellement, il existe un risque non nul de tirer aléatoirement la même information que celle que l'on cherche à masquer *(un nom d'hôpital dans une liste d'hôpitaux, une ville parmi toutes les villes du département, etc.)*, ce qui réduit à néant les efforts de désidentification. Nous avons constaté ce phénomène dans le cas où l'information à réintroduire est dans une forme typographique différente de celle d'origine (par exemple, « *Beau-Mont* » vs. « *BEAUMONT* »).

Au regard des résultats présentés dans le tableau 4, il n'est pas évident de déterminer que l'identification est rendue plus simple lorsque le travail porte sur le même patient (corpus 1 et 2) que sur des patients différents (corpus 3 et 4). Cependant, la disponibilité de plusieurs documents pour un même patient facilite la réidentification du patient dans le SIP de l'hôpital.

Pour ce qui concerne les dates, l'objectif de la désidentification et de la pseudonymisation consiste à préserver la vie privée du patient tout en maintenant un caractère réaliste aux données contenues dans les documents. Dans le domaine médical, les préconisations en matière d'anonymisation précisent qu'« *un dispositif d'anonymisation doit permettre de suivre le dossier d'une même personne, non identifiable, dans la durée* » (Belleil, 2008).

5.4 Limites

L'une des limites de cette étude concerne la taille du corpus. Nous avons restreint la taille du corpus à 60 documents, de manière à proposer une durée de réalisation de l'expérience acceptable pour les différents expérimentateurs. Ce corpus se révèle comparable en termes de taille par rapport à celui utilisé par (Meystre *et al.*, 2014a), constitué de 85 documents. De fait, les configurations que nous souhaitions étudier nous ont conduit à diviser le corpus en sous-ensembles de 15 documents, ce qui permet uniquement de dégager des résultats indicatifs. Nous estimons que cette étude mériterait d'être poursuivie sur une plus grande échelle.

D'autre part, une autre catégorie importante de personnes capables de réidentifier les patients à partir du contenu des documents désidentifiés sont les patients eux-mêmes, ou les parents et proches des patients. Par exemple, une personne qui connait personnellement le patient présenté dans notre exemple (Figure 2) pourrait lire ce document et réaliser que les informations disponibles *(mère de 4 enfants, sans profession, avec une prothèse de hanche et un problème pancréatique dans le passé)* correspondent à l'une de ses connaissances. Nous n'avons cependant pas été en mesure de mettre en place un protocole expérimental adéquat pour mesurer ce risque, la démarche étant longue et coûteuse (recherche des patients, obtention de leur accord pour participer, explication des objectifs et de la procédure suivie, éventuels défraiements, etc.). Nous estimons que la possibilité de réidentifier un tel patient doit être similaire à celle qu'un médecin réidentifie un patient qu'il a soigné dans les trois derniers mois. Il a ainsi été démontré que les médecins ne sont pas capables de réidentifier leurs propres patients à partir de comptes-rendus cliniques désidentifiés au-delà de trois mois (Meystre *et al.*, 2014a).

6 Conclusion

Dans cet article, nous avons présenté les expériences que nous avons menées en matière d'identification, par des expérimentateurs humains, de données identifiantes réelles résiduelles dans un corpus de comptes-rendus cliniques désidentifiés

et porteur d'informations fictives réalistes. Nous avons également étendu ces expériences à des tentatives de réidentification des patients, compte-tenu des informations disponibles dans les documents ou des informations qu'il est possible d'inférer au moyen de connaissances médicales (ici, le codage des actes médicaux).

En dépit de l'absence de désidentification de certains éléments par notre outil de désidentification d'une part, et de la connaissance des faiblesses de l'outil par les développeurs d'autre part, jamais la protection de la vie privée des patients n'a été remise en cause sans disposer d'un accès privilégié au système d'information patient (SIP) de l'hôpital d'où sont issues les données, accès strictement réservé au personnel médical de l'établissement. Lorsqu'un accès au SIP est possible, les patients peuvent être réidentifiés par le biais d'un recoupement d'informations trouvées dans plusieurs documents et par la mobilisation de connaissances médicales sur le codage des actes médicaux. Le respect de la vie privée des patients semble néanmoins respecté lorsque n'est fourni qu'un seul document par patient, y compris pour le personnel médical disposant des accès au SIP.

Enfin, il est moins évident de retrouver des informations lorsque la réintroduction de données géographiques se fait à partir de données issues du même département que celui des données d'origine.

Remerciements

Ce travail a reçu le soutien de l'Agence Nationale pour la Recherche (ANR) dans le cadre du projet CABeRneT (Compréhension Automatique de Textes Biomédicaux pour la Recherche Translationnelle) ANR-13-JS02-0009-01. Les auteurs remercient le service d'informatique medicale du CHU de Rouen pour l'accès au corpus LERUDI, ainsi que les annotateurs qui on participé à cette étude (Annick Choisier, François Morlane-Hondère, Kevin B. Cohen).

Références

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4), 555–96.

BARBARO M. & ZELLER JR T. (2006). A face is exposed for aol searcher no. 4417749. *The New York Times*.

BELLEIL A. (2008). *Référentiel AFCDP des dispositifs d'anonymisation*. Rapport interne, AFCDP.

BENITEZ K. & MALIN B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc*, **17**(2), 169–77.

DELÉGER L., MOLNAR K., SAVOVA G., XIA F., LINGREN T., LI Q., MARSOLO K., JEGGA A., KAISER M., STOUTENBOROUGH L. & SOLTI I. (2013). Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *J Am Med Inform Assoc*, **20**(1), 84–94.

GROUIN C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. Thèse de doctorat, Université Pierre et Marie Curie, Paris, France.

GROUIN C. & NÉVÉOL A. (2014). De-identification of clinical notes in french : towards a protocol for reference corpus developpement. *J Biomed Inform*, **46**(3), 506–515.

LAFFERTY J. D., MCCALLUM A. & PEREIRA F. C. N. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, p. 282–289, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LEE J., SCOTT D. J., VILLARROEL M., CLIFFORD G. D., SAEED M. & MARK R. G. (2011). Open-access MIMIC-II database for intensive care research. In *Proc IEEE Eng Med Biol Soc*, p. 8315–8.

MEYSTRE S., SHEN S., HOFMANN D. & GUNDLAPALLI A. (2014a). Can physicians recognize their own patients in de-identified notes ? In *Stud Health Technol Inform*, volume 205, p. 778–82.

MEYSTRE S. M., FERRÁNDEZ O., FRIEDLIN F. J., SOUTH B. R., SHEN S. & SAMORE M. H. (2014b). Text de-identification for privacy protection : a study of its impact on clinical text information content. *J Biomed Inform*, **50**, 142–50.

MEYSTRE S. M., FRIEDLIN F. J., SOUTH B. R., SHEN S. & SAMORE M. H. (2010). Automatic de-identification of textual documents in the electronic health record : a review of recent research. *BMC Med Res Methodol*, **10**(70).

NEAMATULLAH I., DOUGLASS M. M., LEHMAN L.-W. H., REISNER A., VILLAROEL M., LONG W. J., SZOLOVITS P., MOODY G. B., MARK R. G. & CLIFFORD G. D. (2008). Automated de-identification of free-text medical records. *BMC Med Inform Decis Mak*, **8**(32).

SAEED M., LIEU C., RABER G. & MARK R. G. (2002). MIMIC II : a massive temporal ICU patient database to support research in intelligent patient monitoring. *Comput Cardiol*, **29**, 641–4.

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.

STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). BRAT : a web-based tool for NLP-assisted text annotation. In *Proc of EACL Demonstrations*, p. 102–7, Avignon, France : ACL.

US DEPARTMENT OF HEALTH HUMAN SERVICES S. . (1996). Health Insurance Portability and Accountability Act. http ://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf. §164.514.

UZUNER O., LUO Y. & SZOLOVITS P. (2007). Evaluating the state-of-the-art in automatic de-identification. *J Am Med Inform Assoc*, **14**(5), 550–63.