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

Cyril Grouin, Nicolas Griffon, Aurélie Névéol

CHU Hôpitaux de Rouen

CiSMeF Catalogue et Index des Sites Médicaux de langue Française

cnrs dépasser les frontières

Limsi

LIMICS

Inserm

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI UPR 3251)

Utilisation secondaire des données cliniques

- Consentement éclairé des patients
  - Difficile à obtenir: contact avec le patient, décès
  - Non requis en cas d'anonymisation des données

- ***Désidentification***
  - Masquer/altérer les informations identifiantes
  - Objectif: protéger l'identité du patient

  → rendre impossible la ré-identification

Principe de la désidentification “hiding in plain sight”

[Sweeney, 1996]

- Repérer les informations identifiantes
  - Noms, prénoms, adresses, etc.
- Remplacer par des substituts plausibles
- Hypothèse: les « oublis » passent inaperçus

Impact de la désidentification

- Analyses de textes désidentifiés
  - étiquetage morpho-syntaxique, extraction d'information, entités nommées et concepts **[Deléger et al. 2013, Meystre et al. 2014b]**
- Repérage d'éléments identifiants
  - Augmentation du rappel effectif des outils de désidentification automatique **[Carrell et al., 2013]**
- Réidentificaion des patients
  - Échec pour les médecins traitants dans les trois mois **[Meystre et al., 2014a]**

Objectif:
évaluer les risques de ré-identification

- Contexte: partage de corpus désidentifié
  - Dans le cadre d'un challenge comme DEFT
  - Participants: chercheurs en TAL, médecins

- Hypothèse de désidentification automatique
  - Un seul ou plusieurs documents par patient?
  - Méthode de ré-introduction des substituts géographiques

Préparation du corpus

- Désidentification automatique avec MEDINA
  - Repérage des données identifiantes: modèle CRF entrainé sur 100 documents et évalué à .88 de F-mesure **[Grouin et Névéol 2014]**
  - Remplacement à base de règles

- Présence de données résiduelles

→ Est-il possible de les repérer?

→ Si oui, cela permet-il de ré-identifier les patients?

Préparation du corpus

- 12 types d'information à désidentifier (cf. HIPAA)

- prénom
- nom
- initiales
- adresse postale
- ville
- code postal
- téléphone
- télécopie
- e-mail
- date
- Identifiant (numéro de sécurité sociale, numéro de série)
- nom d'hôpital

- 3 cibles:
- patients, parents des patients, professionnels de sante

- 3 types de documents les plus fréquents en corpus:
- compte-rendu hospitalier, compte-rendu d'acte, correspondance

Critères d’inclusion

- Documents jugés « difficiles » pour l’outil
  - Nom ou prénom:
    - noms composés
  - Information de contact:
    - déclencheurs « domicilié », « personne de contact »
  - Dates: repérage des dates institutionnelles (loi, etc.)
    - Doivent rester inchangées pour préserver la confidentialité

Hypothèses de travail

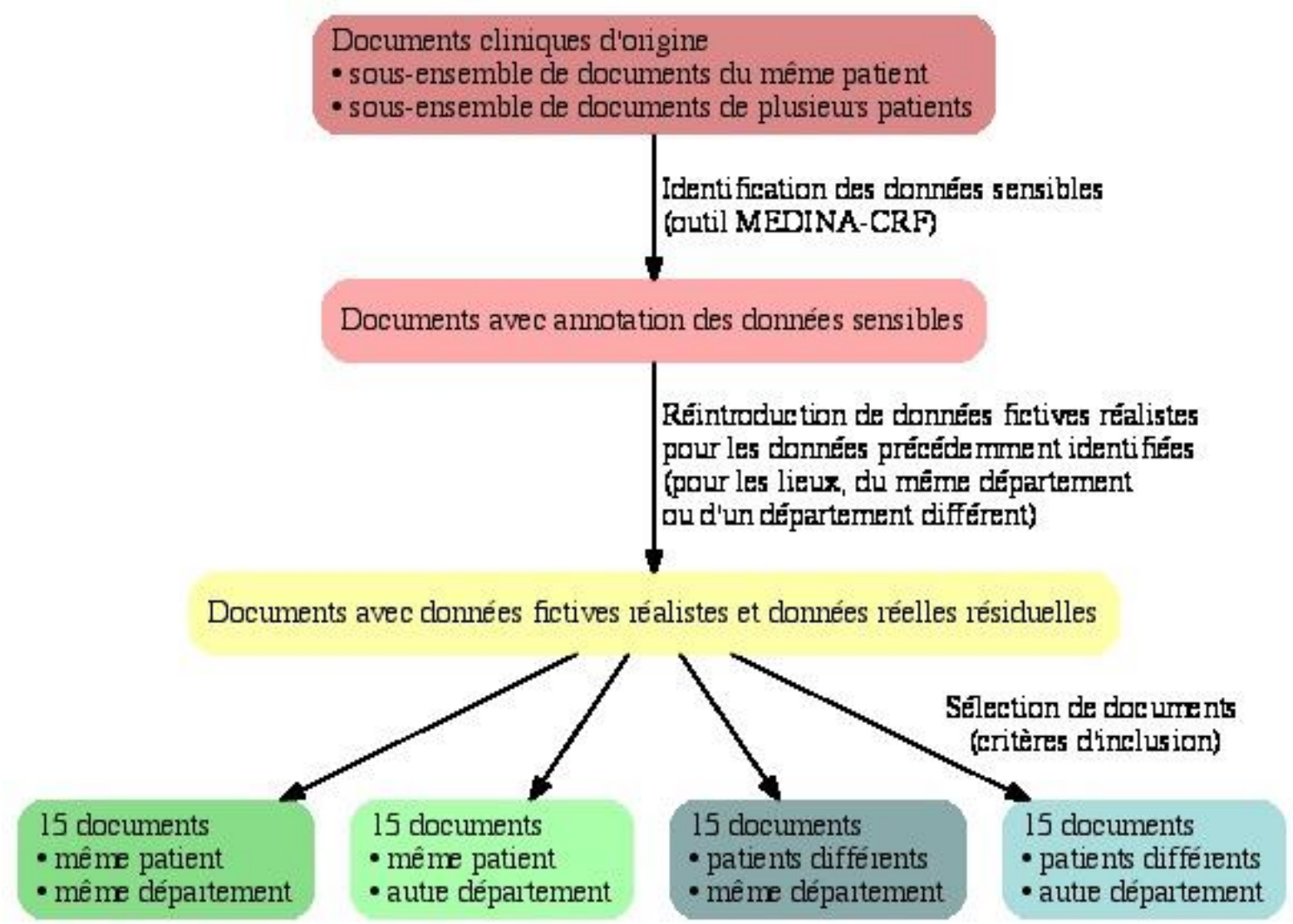
- Un ou plusieurs documents par patient?
  - hypothèse #1 : risque de réidentification plus élevé pour un corpus de documents relatifs à un seul patient

  → possibilité de croiser des informations entre documents

- Origine géographique des documents?
  - hypothèse #2 : identification des données réelles résiduelles plus complexe si données fictives issues du même département

  → origine géographique connue d'emblée

Préparation du corpus



Exemple de document

Original

Martine Dupont, née le 05/08/1928
Mariée, 4 enfants (3 à Nice et 1 en Corse)
Profession : sans profession
...
Personne de confiance : époux
Tel : 06 19 46 13 89
…
Pathologie pancréatique en 1993
...
Dr. Daniel Lucas, Médecin attaché.

Exemple de document

Original, marquage référence

Martine Dupont, née le 05/08/1928
Mariée, 4 enfants (3 à Nice et 1 en Corse)
Profession : sans profession
...
Personne de confiance : époux
Tel : 06 19 46 13 89
…
Pathologie pancréatique en 1993
...
Dr. Daniel Lucas, Médecin attaché.

Exemple de document

Original, marquage référence

Martine Dupont, née le 05/08/1928
Mariée, 4 enfants (3 à Nice et 1 en Corse)
Profession : sans profession
...
Personne de confiance : époux
Tel : 06 19 46 13 89
…
Pathologie pancréatique en 1993
...
Dr. Daniel Lucas, Médecin attaché.

→ Évaluation de l'outil d'annotation automatique

Exemple de document

Original, marquage référence

Martine Dupont, née le 05/08/1928
Mariée, 4 enfants (3 à Nice et 1 en Corse)
Profession : sans profession
...
Personne de confiance : époux
Tel : 06 19 46 13 89
…
Pathologie pancréatique en 1993
...
Dr. Daniel Lucas, Médecin attaché.

→ Évaluation de l'outil d'annotation automatique

Texte désidentifié

Monique Durand, née le 04/07/1927
Mariée, 4 enfants (3 à Nice et 1 en Corse)
Profession : sans profession
...
Personne de confiance : époux
Tel : 06 25 17 46 59
…
Pathologie pancréatique en 1992
...
Dr. Gregory House, Médecin attaché.

Exemple de document

Original, marquage référence

Martine Dupont, née le 05/08/1928
Mariée, 4 enfants (3 à Nice et 1 en Corse)
Profession : sans profession
...
Personne de confiance : époux
Tel : 06 19 46 13 89
…
Pathologie pancréatique en 1993
...
Dr. Daniel Lucas, Médecin attaché.

→ Évaluation de l'outil d'annotation automatique

Texte désidentifié, marquage référence

Monique Durand, née le 04/07/1927
Mariée, 4 enfants (3 à Nice et 1 en Corse)
Profession : sans profession
...
Personne de confiance : époux
Tel : 06 25 17 46 59
…
Pathologie pancréatique en 1992
...
Dr. Gregory House, Médecin attaché.

Exemple de document

Original, marquage référence

Martine Dupont, née le 05/08/1928
Mariée, 4 enfants (3 à Nice et 1 en Corse)
Profession : sans profession
...
Personne de confiance : époux
Tel : 06 19 46 13 89
…
Pathologie pancréatique en 1993
...
Dr. Daniel Lucas, Médecin attaché.

→ Évaluation de l'outil d'annotation automatique

Texte désidentifié, marquage référence

Monique Durand, née le 04/07/1927
Mariée, 4 enfants (3 à Nice et 1 en Corse)
Profession : sans profession
...
Personne de confiance : époux
Tel : 06 25 17 46 59
…
Pathologie pancréatique en 1992
...
Dr. Gregory House, Médecin attaché.

→ Évaluation du repérage des informations identifiantes résiduelles

Protocole expérimental

- 3 catégories d’expérimentateurs
  - 1 médecin de l'hôpital fournisseur du corpus
  - 2 chercheurs ayant conçu l'outil de désidentification
  - 3 autres chercheurs en TAL, sans connaissance particulière du corpus ou de l'outil
- Consignes
  - Annoter les données identifiantes résiduelles
  - Essayer de ré-identifier les patients
- Questions sur les stratégies mises en oeuvre

Distribution des données identifiantes dans le corpus

Catégorie	Total	Réelles
Noms	541	18 (3,3%)
Prénoms	487	17 (3,5%)
Initiales	39	35 (89,7%)
Identifiants	20	16 (80,0%)
Hôpitaux	166	24 (14,5%)
Adresses	60	21 (51,7%)
Codes postaux	67	12 (17,9%)
Villes	153	39 (25,5%)
Dates	233	17 (7,3%)
E-mails	42	0
Télephones	282	0
TOTAL	**2090**	**199 (9,5%)**

Performances du système de désidentification

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
Télephones	0,99	1,00	0,99
Total	**0,96**	**0,90**	**0,93**

→ un peu mieux qu'attendu (comparable)

Reconnaissance des données résiduelles

Corpus	Dev 1 N – P – R - F	Médecin N – P – R - F	Dev 2 N – P – R - F
1 – MP,MD	34 .71 .33 .45	13 .62 .11 .19	285 .16 .64 .26
2 – MP, DD	35 .57 .54 .56	11 .64 .18 .29	59 .19 .30 .23
3 – PD, MD	31 .61 .40 .49	19 .47 .19 .27	28 .71 .43 .53
4 – PD, DD	42 .67 .50 .57	25 .76 .34 .47	41 .51 .38 .43
total	**135 .62 .41 .50**	**66 .61 .20 .30**	**408 .23 .46 .30**

Corpus	Chercheur 1 N – P – R - F	Chercheur 2 N – P – R - F	Chercheur 3 N – P – R - F
1 – MP,MD	30 .47 .19 .27	0 .00 .00 .00	26 .00 .00 .00
2 – MP, DD	8 .50 .11 .18	66 .02 .03 .02	24 .00 .00 .00
3 – PD, MD	6 .33 .04 .08	0 .00 .00 .00	26 .00 .00 .00
4 – PD, DD	15 .80 .21 .34	43 .07 .05 .06	10 .00 .00 .00
total	**56 .54 .15 .23**	**109 .04 .02 .03**	**88 .00 .00 .00**

Accords inter-annotateur (F-mesure)

	Dev 1	Médecin	Dev 2	Ch 1	Ch 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

→ Accords très faibles, même pour les annotateurs avec des connaissances fines des données

Tentatives de ré-identification

- Utilisation d'annuaires inversés (dev 1, ch 1)
  - → identification de l'hôpital d'origine
  - → pas de ré-identification de patient

- Utilisation du système hospitalier (médecin) et de compétences en codage CIM10
  - → ré-identification des deux patients à multiples documents > 30 minutes chacun

Configuration de la désidentification

- Un ou plusieurs documents par patient?
  - résultats peu probants sur la reconnaissances des données résiduelles
  - facilite la ré-identification par le médecin

- Configuration géographique

  → la ré-introduction de données du même département permet une plus faible reconnaissance des données résiduelles

Limites

- Corpus de petite taille
  - taille comparable à Meystre et al. (N=85)
  - sous corpus de 15 documents: résultats indicatifs

  → temps d'annotation raisonnable (2h/annotateur)

- Expérimentateurs
  - pas d'expérimentateurs à « haut-risque » tel que le patient lui-même ou ses proches

Conclusion

- La réidentification n’est pas impossible

- … mais demande:
  - du temps
  - un accès privilégié au système hospitalier
  - des connaissances médicales

Merci!

neveol@limsi.fr

Annotateurs:
Annick Choisier
Kevin Cohen
François Morlane-Hondère

Financement:

CABeRneT ANR-13-JS02-0009-01