Anonymisation de documents médicaux en texte libre et en français via réseaux de neurones

PFIA 2023 - Journée Santé & IA

Antoine RICHARD[1], François TALBOT[1], et David GIMBERT[1]

[1]DSN Bron, Hospices Civil de Lyon, 61 Boulevard Pinel, 69672 Bron, France
antoine.richard@chu-lyon.fr

6 Juillet 2023

1 Introduction
- Contexte
- Problématique
- Objectifs

2 Méthodologie
- Préparation des données
- Apprentissage
- Ré-identification

3 Résultats
- Pré-apprentissage
- Apprentissage
- Ré-identification

4 Conclusion

Les Hospices Civils de Lyon (HCL)

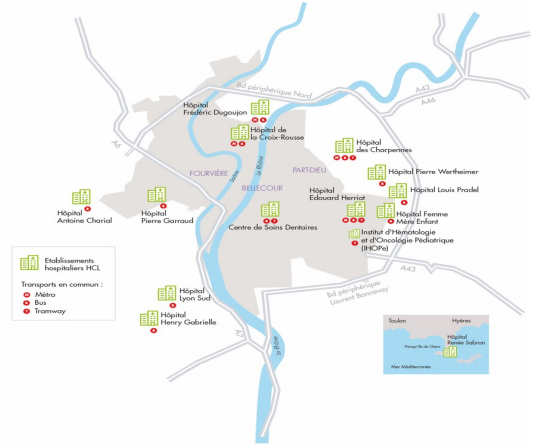


Figure – Établissements affiliés aux HCL

- $>$ 20.000 soignant.e.s
- $>$ 1.000.000 consultations par an
- 4 départements informatiques

Easily® [1]

- Dossier Patient Informatisé développé par les HCL
- Déployé dans plus d'une centaine d'établissements de santé
- Conception par portails Métiers
- Divers modules :
  - Suivi des patients
  - Prescriptions
  - Aide à la décision
  - PMSI
  - Recherche Clinique

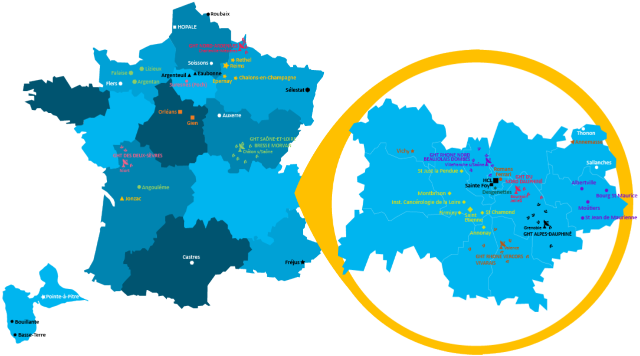


Figure – Hôpitaux utilisant Easily®en France

1. https ://hopsis.org/fr/groupe/carte-des-etablissements/

Entrepôt de Données de Santé (EDS)[2]

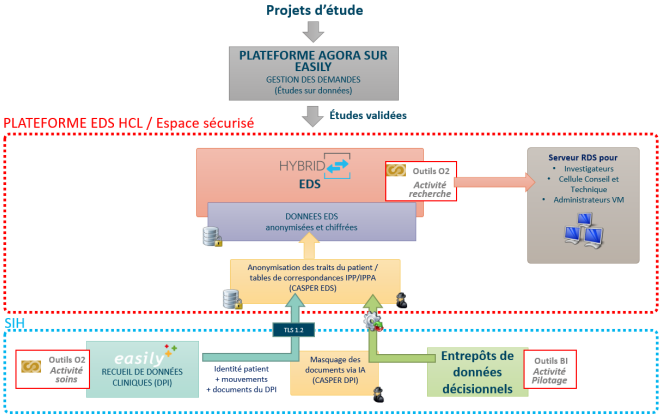


Figure – Interactions entre Easily® et l'EDS des HCL

2. https ://www.cnil.fr/fr/la-cnil-adopte-un-referentiel-sur-les-entrepots-de-donnees-de-sante

Protection des données privées [3]

Risque d'individualisation

Il ne doit pas être possible d'isoler un individu dans un jeu de données

Risque de corrélation

Il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu

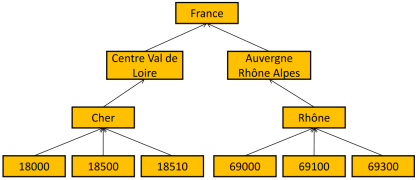
Risque d'inférence

il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu

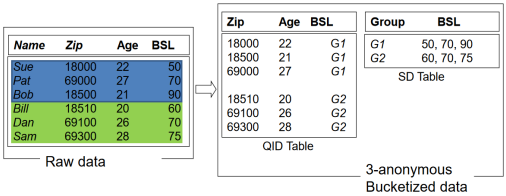
$\Rightarrow$ Anonymisation/Pseudonymisation des données (Olatunji et al. 2022)

3. https ://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

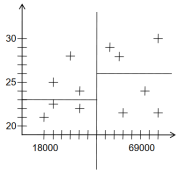
Anonymisation/Pseudonymisation de bases de données [4]



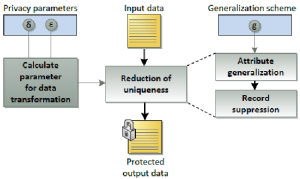
Generalization Algorithm (Sweeney 2002)



Bucketization (Xiao et Tao 2008)



Mondrian Algorithm (LeFevre, DeWitt et Ramakrishnan 2006)



ARX : SafePub Algorithm (Bild, Kuhn et Prasser 2018)

4. https ://arx.deidentifier.org/

Anonymisation/Pseudonymisation de textes libres

GROUPEMENT HOSPITALIER CENTRE
Hôpital Edouard Herriot
5, Place d'Arsonval
69437 Lyon cedex 03 - France

N° FINESS : 690781154 | N° FINESS HCL : 690781810

Lyon, le [redacted]

Mme [redacted]

Cher Confrère,

Veuillez trouver ci-dessous le **compte-rendu d'hospitalisation** du [redacted] au [redacted] votre patiente Mme [redacted]

Motif : Poussée d'urticaire sur fond d'urticaire chronique spontanée.

Histoire de la maladie : Antécédents:
Médicaux:
- Lupus systémique avec atteinte cutanéo-articulaire et pleuropéricardique Suivi Dr [redacted]
- Urticaire(VEMS)
- L'Urticaire/Vasculite urticarienne sous Aerius 4j + Singulair, poussées quotidiennes (urticaire et angioedème)
Chirurgicaux:
- appendicectomie en 2003
- réduction mammaire
- Pas d'atopie personnelle ou familiale

Mode de vie : en instance de divorce, vit seule. Pas d'enfants. A repris une formation d'assistante sociale.

Histoire de la maladie:
- date de début : Urticaire avec angiooedeme début 2015. Biopsies cutanées histologie + IFD en Juillet 2015 et 2017 vasculite urticarienne. Caractère migrateur des lésions.
- Traitements déjà réalisés : AERIUS/POLARAMINE/SINGULAIR/BILASTINE
- facteurs aggravants : asthénie, transpiration, tomates, fraises
- facteurs soulagents : Amélioration dans l'eau froide. Efficacité partielle sous antihistaminiques.
- Sous XOLAIR depuis le 1er juin 2018, initialement une injection/mois puis rapidement une injection/deux semaines. Dernière injection le 19/10/18 à 375mg.

Depuis le 26/10 : fièvre a 38°3 et rhinite : a vu un médecin -> cause virale. Frissons ces 3 derniers jours, dolprane ce matin.
- Depuis 3-4 jours poussées d'urticaire + fortes, passe ses nuits dans la salle de bains sous l'eau car prurit++
- Oedème du visage le matin en se réveillant, qui passe en 3H

Score UCT : 1/16
Score UAS7 : score lésionnel : 18/21 score de prurit : 18/21

Examen clinique :

Lésions erythémateuses prurigineuses, sur tout le corps surtout palmaire plantaire et thoracique

Figure – Exemple de document médical pour une patiente factice

Anonymisation/Pseudonymisation de textes libres

GROUPEMENT HOSPITALIER CENTRE
Hôpital Edouard Herriot
5, Place d'Arsonval
69437 Lyon cedex 03 - France

N° FINESS : 690781154 | N° FINESS HCL : 690781810

Service d'ophtalmologie

Chef de service :

Pr. C. BURILLON
Secrétariat : 04 72 11 42 17
Tél : 04 72 11 42 36
E-mail : marie-odile@chu-lyon.fr

Praticiens Hospitaliers :

Dr. J. FLEURY
Secrétariat : 04 72 11 03 92
E-mail : virginie.comte-marie-magdeleine@chu-lyon.fr

Dr. C. ROGER-BONNAS
Secrétariat : 04 72 11 95 13
E-mail : anne.cucci@chu-lyon.fr

Dr. B. CHARLEUX
Secrétariat : 04 72 11 95 13
E-mail : marie-line-claude.blin@chu-lyon.fr

Assistants - Chefs de Clinique :

Dr. C. ROCHEFEAU
Secrétariat : 04 72 11 95 13
E-mail : marie-line-claude.blin@chu-lyon.fr

Dr. R. MOUCHEL
Secrétariat : 04 72 11 95 13
E-mail : anne.cucci@chu-lyon.fr

Dr. A. TAIEB
Secrétariat : 04 72 11 95 13
E-mail : marie-veronique.vinson@chu-lyon.fr

Consultation Orthoptique :
Secrétariat : 04 72 11 95 13
E-mail : ch-orsdesme@chu-lyon.fr

Examens électro-physiologiques :
Secrétariat : 04 72 11 95 13
E-mail : ghislaine.reguillon@chu-lyon.fr

Consultation ophtalmologique :
Secrétariat : 04 72 11 42 54

Consultation d'Urgence 24h/24h :
Tél. : 04 72 11 42 33

Chirurgie Ambulatoire :
Secrétariat : 04 72 11 09 47
Fax : 04 72 11 09 67

Hospices Civils de Lyon
www.chu-lyon.fr

Renseignements ACI
3 825 0 825 69 (0,15 €/mn)

@source HCL-sante
ref : [redacted]

Lyon, le [redacted]

Mme B.[redacted]
[redacted]
[redacted]
[redacted]ques

Cher Confrère,

Veuillez trouver ci-dessous le **compte-rendu d'hospitalisation** du [redacted] au [redacted] votre patiente Mme [redacted] née le [redacted]

Motif : Poussée d'urticaire sur fond d'urticaire chronique spontanée.

Histoire de la maladie : Antécédents:
Médicaux:
- Lupus systémique avec atteinte cutanéo-articulaire et pleuropéricardique. Suivi Dr VILLAR (CHU)
- L'Urticaire/Vasculite urticarienne sous Aerius 4j + Singulair, poussées quotidiennes (urticaire et angioedème)
Chirurgicaux:
- appendicectomie en 2003
- réduction mammaire
- Pas d'atopie personnelle ou familiale

Mode de vie : en instance de divorce, vit seule. Pas d'enfants. A repris une formation d'assistante sociale.

Histoire de la maladie:
- date de début : Urticaire avec angiooedeme début 2015. Biopsies cutanées histologie + IFD en Juillet 2015 et 2017 vasculite urticarienne. Caractère migrateur des lésions.
- Traitements déjà réalisés : AERIUS/POLARAMINE/SINGULAIR/BILASTINE
- facteurs aggravants : asthénie, transpiration, tomates, fraises
- facteurs améliorants : Amélioration dans l'eau froide. Efficacité partielle sous antihistaminiques.
- Sous XOLAIR depuis le 1er juin 2018, initialement une injection/mois puis rapidement une injection/deux semaines. Dernière injection le 19/10/18 à 375mg.

Depuis le 26/10 : fièvre a 38°3 et rhinite : a vu un médecin -> cause virale. Frissons ces 3 derniers jours, dolprane ce matin
- Depuis 3-4 jours poussées d'urticaire + fortes, passe ses nuits dans la salle de bains sous l'eau car prurit++
- Oedème du visage le matin en se réveillant, qui passe en 3H

Score UAS7 : 1/16
Score UAS7 : score lésionnel : 18/21 score de prurit : 18/21

Examen clinique :

Lésions erythémateuses prurigineuses, sur tout le corps surtout palmaire plantaire et thoracique

Notre service participe au projet ZEPRA (Zéro Papier en Rhône-Alpes) : envoi des courriers par voie électronique.
En tant que médecin correspondant, si cela vous intéresse, transmettez vos coordonnées, numéro RPPS et email à hcl.zepra@chu-lyon.fr
Renseignements HCL : 08253 0 825 69 (0,15 €/mn)
www.chu-lyon.fr

Traitement Automatique du Langage :

- En anglais :
  - Chazard et al. 2014 (RegEx + Stopwords)
- En espagnol :
  - Lima et al. 2020 (RegEx + Stopwords)
- En français :
  - Grouin 2013 (RegEx + Statistiques)

Figure – Exemple de document médical pour une patiente factice

Anonymisation/Pseudonymisation de textes libres

GROUPEMENT HOSPITALIER CENTRE
Hôpital Edouard Herriot
5, Place d'Arsonval
69437 Lyon cedex 03 - France

N° FINESS : 690781154 | N° FINESS HCL : 690781810

Service d'ophtalmologie

Chef de service :

Pr. C. BURILLON
Secrétariat : 04 72 11 42 17
Tél. : 04 72 11 42 36
E-mail : marie-odile@chu-lyon.fr

Praticiens Hospitaliers :

Dr. J. FLEURY
Secrétariat : 04 72 11 69 92
E-mail : virginie.comte-marie-magdeleine@chu-lyon.fr

Dr. C. ROGER-BONNAS
Secrétariat : 04 72 11 95 13
E-mail : anne.cucci@chu-lyon.fr

Dr. B. CHARLEUX
Secrétariat : 04 72 11 95 13
E-mail : marie-line-claude.blin@chu-lyon.fr

Assistants - Chefs de Clinique :

Dr C. ROCHEFEAU
Secrétariat : 04 72 11 95 13
E-mail : marie-line-claude.blin@chu-lyon.fr

Dr. R. MOUCHEL
Secrétariat : 04 72 11 95 13
E-mail : anne.cucci@chu-lyon.fr

Dr. A. TAIEB
Secrétariat : 04 72 11 95 13
E-mail : marie-veronique.vinson@chu-lyon.fr

Consultation Orthoptique :
Secrétariat : 04 72 11 95 13
E-mail : ch-orsdesse@chu-lyon.fr

Examens électro-physiologiques :
Secrétariat : 04 72 11 95 13
E-mail : gilbert.regnault@chu-lyon.fr

Consultation ophtalmologique :
Secrétariat : 04 72 11 42 54

Consultation d'Urgence 24h/24h :
Tél. : 04 72 11 42 33

Chirurgie Ambulatoire :
Secrétariat : 04 72 11 89 47
Fax : 04 72 11 89 67

Hospices Civils de Lyon
www.chu-lyon.fr

Renseignements HCL
3 825 0 825 69 (0,15 €/mn)

@source HCL-easily
ppf : [redacted]

Lyon, le [redacted]

Mme B.A [redacted]
[redacted]
[redacted]

[redacted]QUES

Cher Confrère,

Veuillez trouver ci-dessous le **compte-rendu d'hospitalisation** du [redacted] au [redacted] votre patiente Mme [redacted] née le [redacted].

Motif : Poussée d'urticaire sur fond d'urticaire chronique spontanée.

Histoire de la maladie : Antécédents:
Médicaux:
- Lupus systémique avec atteinte cutanéo-articulaire et pleuropéricardique. Suivi Dr VILLAR-HDHJ
- L'Urticaire/Vasculite urticarienne sous Aerius 4j + Singulair, poussées quotidiennes (urticaire et angioedèmes)
Chirurgicaux:
- appendicectomie en 2003
- réduction mammaire
- Pas d'allergie personnelle ou familiale

Mode de vie : en instance de divorce, vit seule. Pas d'enfants. A repris une formation d'assistante sociale.

Histoire de la maladie:
- date de début : Urticaire avec angiooedeme début 2015. Biopsies cutanées histologie + IFD en Juillet 2015 et 2017 vasculite urticarienne. Caractère migrateur des lésions.
- Traitements déjà réalisés : AERIUS/POLARAMINE/SINGULAIR/BILASTINE
- facteurs aggravants : asthénie, transpiration, tomates, fraises
- facteurs favorisants : Amélioration dans l'eau froide. Efficacité partielle sous antihistaminiques.
- Sous XOLAIR depuis le 1er juin 2018, initialement une injection/mois puis rapidement une injection/deux semaines. Dernière injection le 19/10/18 à 375mg.

Depuis le 26/10 : fièvre a 38°3 et rhinite : a vu un médecin -> cause virale. Frissons ces 3 derniers jours, dolprane ce matin.
- Depuis 3-4 jours poussées d'urticaire + fortes, passe ses nuits dans la salle de bains sous l'eau car prurit++
- Oedème du visage le matin en se réveillant, qui passe en 3H
Score UCT : 1/16
Score UAS7 : score lésionnel : 18/21 score de prurit : 18/21

Examen clinique :

Lésions erythémateuses prurigineuses, sur tout le corps surtout palmaire plantaire et thoracique

Notre service participe au projet ZERMA (Zéro Papier en Rhône-Alpes) : envoi des courriers par voie électronique.
En tant que médecin correspondant, si cette voie vous intéresse, transmettez vos coordonnées, numéro RPPS et email à hcl.zerma@chu-lyon.fr
Renseignements HCL : 08253 69 825 (0,15 €/mn)
www.chu-lyon.fr

Traitement Automatique du Langage :

- En anglais :
  - Chazard et al. 2014 (RegEx + Stopwords)
- En espagnol :
  - Lima et al. 2020 (RegEx + Stopwords)
- En français :
  - Grouin 2013 (RegEx + Statistiques)

Machine Learning :

- En anglais :
  - Johnson, Bulgarelli et Pollard 2020 (BERT)
- En japonais :
  - Kajiyama et al. 2020 (Bi-LSTM)
- En français :
  - Bourdois et al. 2021 (FlauBERT)
  - Tannier et al. 2023 (CamemBERT)

Figure – Exemple de document médical pour une patiente factice

Anonymisation/Pseudonymisation de textes libres

GROUPEMENT HOSPITALIER CENTRE
Hôpital Edouard Herriot
5, Place d'Arsonval
69437 Lyon cedex 03 - France

N° FINESS : 690781154 | N° FINESS HCL : 690781103

Service d'ophtalmologie

Chef de service :

Pr. C. BURILLON
Secrétariat : 04 72 11 42 17
Tél : 04 72 11 42 36
E-mail : marie-odile.d@chu-lyon.fr

Praticiens Hospitaliers :

Dr. J. FLEURY
Secrétariat : 04 72 11 69 92
E-mail : virginie.comte-marie-magdeleine@chu-lyon.fr

Dr. C. ROGER-BONNAS
Secrétariat : 04 72 11 95 13
E-mail : anne.cucci@chu-lyon.fr

Dr. B. CHARLEUX
Secrétariat : 04 72 11 95 13
E-mail : marie-line-claude.bin@chu-lyon.fr

Assistants - Chefs de Clinique :

Dr C. ROCHEFEAU
Secrétariat : 04 72 11 95 13
E-mail : marie-line-claude.bin@chu-lyon.fr

Dr. R. MOUCHEL
Secrétariat : 04 72 11 95 13
E-mail : anne.cucci@chu-lyon.fr

Dr. A. TAIEB
Secrétariat : 04 72 11 95 13
E-mail : marie-veronique.vinson@chu-lyon.fr

Consultation Orthoptique :
Secrétariat : 04 72 11 95 13
E-mail : ch-orsdonne@chu-lyon.fr

Examens électro-physiologiques :
Secrétariat : 04 72 11 95 13
E-mail : ghe.eeg.opthalmo@chu-lyon.fr

Consultation ophtalmologique :
Secrétariat : 04 72 11 42 94

Consultation d'Urgence 24h/24h :
Tél. : 04 72 11 42 33

Chirurgie Ambulatoire :
Secrétariat : 04 72 11 89 47
Fax : 04 72 11 89 87

Hospices Civils de Lyon
www.chu-lyon.fr
Renseignements HCL
3 825 0 825 69 (0,15 €/mn)

Notre service participe au projet ZERNA (Zéro Papier en Rhône-Alpes) : envoi des courriers par voie électronique.
En tant que médecin correspondant, si cela vous intéresse, transmettez vos coordonnées, numéro RPPS et email à hcl.zerna@chu-lyon.fr
Renseignements HCL : 08253 69 825 (0,15 €/mn)
www.chu-lyon.fr

@source HCL-easily
ppr [redacted]

Lyon, le [redacted]

Mme B.A. [redacted]
[redacted]
[redacted]

[redacted]QUES

Cher Confrère,

Veuillez trouver ci-dessous le **compte-rendu d'hospitalisation** du [redacted] au [redacted] votre patiente Mme [redacted] née le [redacted].

Motif : Poussée d'urticaire sur fond d'urticaire chronique spontanée.

Histoire de la maladie : Antécédents:
Médicaux:
- Lupus systémique avec atteinte cutanéo-articulaire et pleuropéricardique. Suivi Dr VILLAR-HDHJ
- L'Urticaire/Vasculite urticarienne sous Aeriius 4j + Singulair, poussées quotidiennes (urticaire et angioedèmes)
Chirurgicaux:
- appendicectomie en 2003
- réduction mammaire
- Pas d'allergie personnelle ou familiale

Mode de vie : en instance de divorce, vit seule. Pas d'enfants. A repris une formation d'assistante sociale.

Histoire de la maladie:
- date de début : Urticaire avec angiooedeme début 2015. Biopsies cutanées histologie + IFD en Juillet 2015 et 2017 vasculite urticarienne. Caractère migrateur des lésions.
- Traitements déjà réalisés : AERIUS/POLARAMINE/SINGULAIR/BILASTINE
- facteurs aggravants : asthénie, transpiration, tomates, fraises
- facteurs favorisants : Amélioration dans l'eau froide. Efficacité partielle sous antihistaminiques.
- Sous XOLAIR depuis le 1er juin 2018, initialement une injection/mois puis rapidement une injection/deux semaines. Dernière injection le 19/10/18 à 375mg.

Depuis le 26/10 : fièvre a 38°3 et rhinite : a vu un médecin -> cause virale. Frissons ces 3 derniers jours, dolprane ce matin.
- Depuis 3-4 jours poussées d'urticaire + fortes, passe ses nuits dans la salle de bains sous l'eau car prurit++
- Oedème du visage le matin en se réveillant, qui passe en 3H
Score UAS7 : 1/16
Score UAS7 : score lésionnel : 18/21 score de prurit : 18/21

Examen clinique :

Lésions erythémateuses prurigineuses, sur tout le corps surtout palmaire plantaire et thoracique

Figure – Exemple de document médical pour une patiente factice

Traitement Automatique du Langage :

- En anglais :
  - Chazard et al. 2014 (RegEx + Stopwords)
- En espagnol :
  - Lima et al. 2020 (RegEx + Stopwords)
- En français :
  - Grouin 2013 (RegEx + Statistiques)

Machine Learning :

- En anglais :
  - Johnson, Bulgarelli et Pollard 2020 (BERT)
- En japonais :
  - Kajiyama et al. 2020 (Bi-LSTM)
- En français :
  - Bourdois et al. 2021 (FlauBERT)
  - Tannier et al. 2023 (CamemBERT)

⇒ Risques de perte d'utilité (Obeid et al. 2019)

Détecter les éléments identifiants d'un texte libre

Un problème de Reconnaissance d'Entités Nommées (NER)

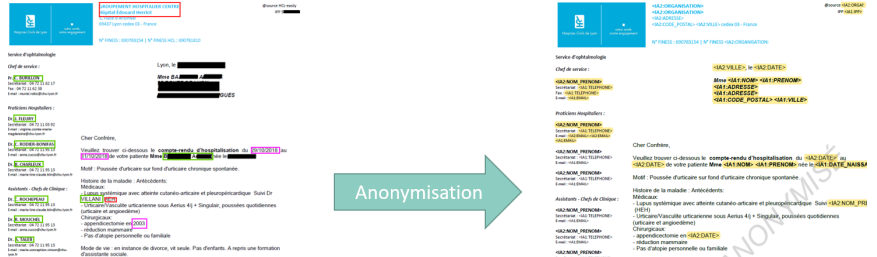


Figure – Exemple d'anonymisation d'un document médical en texte libre

Problématiques traitées

1. La détection des éléments identifiants est-elle faisable grâce à CamemBERT ? (Martin et al. 2020)
2. Le pré-apprentissage du jargon médical améliore-t-il la détection ?
3. Y-a-t-il un risque de ré-identification si le modèle venait à être diffusé ?

1 Introduction
- Contexte
- Problématique
- Objectifs

2 Méthodologie
- Préparation des données
- Apprentissage
- Ré-identification

3 Résultats
- Pré-apprentissage
- Apprentissage
- Ré-identification

4 Conclusion

Données de pré-apprentissage du jargon médical

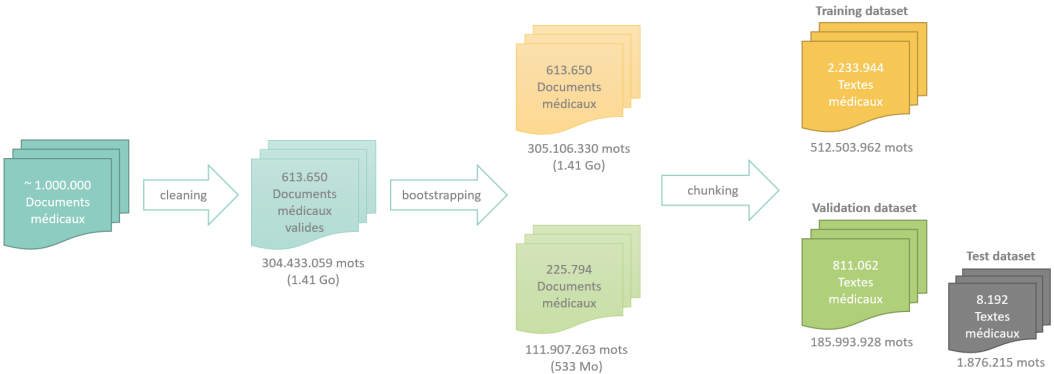


Figure – Processus de préparation des données pour le pré-apprentissage du jargon médical

Données d'apprentissage des éléments identifiants

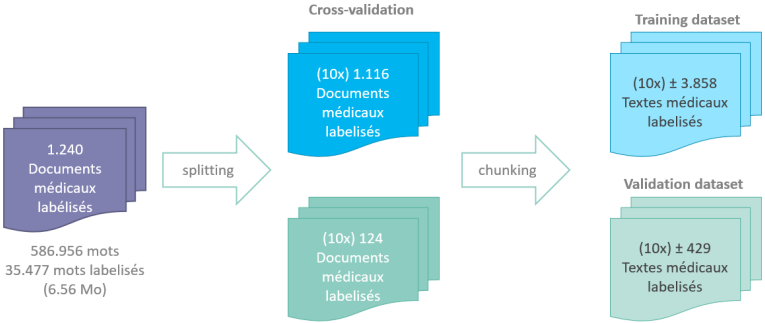


Figure – Processus de préparation des données pour l'apprentissage des éléments identifiants

12 Labels :

- Nom/Prénom (10.475)
- Identifiant Patient Permanent (405)
- Numéro de Dossier (761)
- Téléphone (5.327)
- Email (1.704)
- Date (6.698)
- CodePostal (1.081)
- Ville (1.524)
- Voie (1.033)
- Localité (472)
- Organisation (5.506)
- SiteWeb (491)

Pipeline d'apprentissage

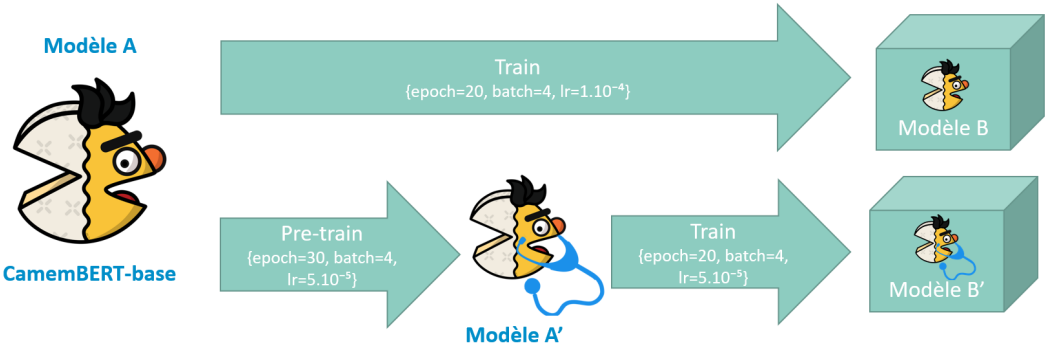


Figure – Processus de pré-apprentissage et d'apprentissage

Évaluation du risque de ré-identification

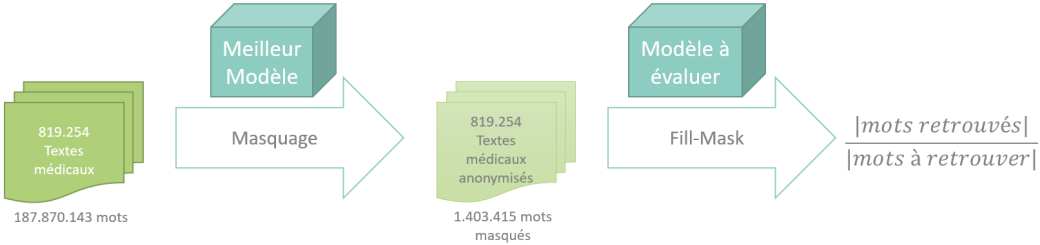


Figure – Processus d'évaluation du risque de ré-identification de documents anonymisés

1 Introduction
- Contexte
- Problématique
- Objectifs

2 Méthodologie
- Préparation des données
- Apprentissage
- Ré-identification

3 Résultats
- Pré-apprentissage
- Apprentissage
- Ré-identification

4 Conclusion

Perplexité avant et après pré-apprentissage

$$Perplexity(W) = 2^{H(W)} \tag{1}$$

$$H(W) = -\sum_{w \in W} p(w) \times log(p(w)) \tag{2}$$

Modèle	Perplexité		
	moyenne	minimum	maximum
Modèle A (CamemBERT)	9.48 ± 0.85	5.34	14.22
Modèle A' (CamemBERT pré-entrainé)	0.66 ± 0.3	0.0001	2.64

Table – Résultats du pré-apprentissage du jargon médical par CamemBERT

Exemples de résultats

Texte	Solution	Modèle entrainé	Propositions (scores)
Contrôle 1 an 1/2 après <mask> radical avec Bricker	cystoprostatectomie	non	rupture (0.29) séparation (0.04)
		oui	prostatectomie (0.49) chirurgie (0.31)
A ce stade, il existe des ondes lentes <mask> dans les deux régions frontales	diphasiques	non	réparties (0.09) principalement (0.04)
		oui	delta (0.34) diffuses (0.09)
Le contrôle de la <mask> huméro-basilique gauche est plutôt bon	fistule	non	ceinture (0.2) zone (0.16)
		oui	fistule (0.6) prothèse (0.13)

Table – Exemples de complétions de textes par CamemBERT avec et sans pré-apprentissage du jargon médical

Détection des éléments identifiants

	Modèle B (sans pré-apprentissage)			Modèle B' (avec pré-apprentissage)		
Label	Précision	Rappel	F1-Score	Précision	Rappel	F1-Score
Nom/Prénom	**0.967 ± 0.009**	**0.965 ± 0.018**	**0.966 ± 0.013**	0.957 ± 0.009	0.957 ± 0.018	0.957 ± 0.013
IPP	**0.964 ± 0.028**	**0.967 ± 0.032**	**0.966 ± 0.027**	0.94 ± 0.068	0.962 ± 0.033	0.95 ± 0.046
NoDossier	**0.845 ± 0.053**	**0.889 ± 0.052**	**0.866 ± 0.046**	0.8 ± 0.045	0.837 ± 0.075	0.817 ± 0.048
Téléphone	**0.983 ± 0.013**	**0.988 ± 0.013**	**0.985 ± 0.012**	0.976 ± 0.021	0.987 ± 0.012	0.981 ± 0.014
EMail	**0.985 ± 0.013**	**0.992 ± 0.009**	**0.988 ± 0.01**	0.96 ± 0.027	0.987 ± 0.008	0.973 ± 0.014
Date	**0.965 ± 0.015**	**0.969 ± 0.016**	**0.967 ± 0.013**	0.953 ± 0.022	0.961 ± 0.026	0.957 ± 0.023
Code Postal	**0.989 ± 0.011**	**0.989 ± 0.013**	**0.989 ± 0.01**	0.98 ± 0.02	0.988 ± 0.011	0.984 ± 0.011
Ville	**0.944 ± 0.025**	**0.943 ± 0.026**	**0.943 ± 0.02**	0.908 ± 0.027	0.929 ± 0.033	0.918 ± 0.021
Voie	**0.936 ± 0.026**	**0.959 ± 0.02**	**0.947 ± 0.02**	0.91 ± 0.039	0.945 ± 0.022	0.927 ± 0.028
Localité	**0.946 ± 0.05**	**0.946 ± 0.037**	**0.945 ± 0.024**	0.93 ± 0.048	0.933 ± 0.035	0.931 ± 0.027
Organisation	**0.806 ± 0.047**	**0.825 ± 0.024**	**0.815 ± 0.035**	0.774 ± 0.054	0.805 ± 0.023	0.789 ± 0.037
Site Web	**0.968 ± 0.041**	**0.98 ± 0.026**	**0.974 ± 0.031**	0.926 ± 0.05	0.975 ± 0.033	0.95 ± 0.035
Moyenne	**0.942 ± 0.057**	**0.951 ± 0.049**	**0.946 ± 0.053**	0.918 ± 0.066	0.939 ± 0.059	0.928 ± 0.062

Table – Résultats de l'évaluation des modèles B et B' pour la détection d'éléments identifiants

Risques de ré-identification

Modèle	Nombre de mots retrouvés
Modèle A (sans pré-apprentissage)	1513 (0.1%)
Modèle A' (avec pré-apprentissage)	115.191 (8%)
Modèle B (sans pré-apprentissage)	**3 (0.0002%)**
Modèle B' (avec pré-apprentissage)	19 (0.001%)

Table – Résultats de l'évaluation du risque de ré-identification des données patient

1 Introduction
- Contexte
- Problématique
- Objectifs

2 Méthodologie
- Préparation des données
- Apprentissage
- Ré-identification

3 Résultats
- Pré-apprentissage
- Apprentissage
- Ré-identification

4 Conclusion

Synthèse et perspectives

Conclusion

1. Détection d'éléments identifiants faisable grâce à CamemBERT

Synthèse et perspectives

Conclusion

1. Détection d'éléments identifiants faisable grâce à CamemBERT
2. Le pré-apprentissage n'a pas amélioré les performances

Synthèse et perspectives

Conclusion

1. Détection d'éléments identifiants faisable grâce à CamemBERT
2. Le pré-apprentissage n'a pas amélioré les performances
3. Le pré-apprentissage comporte des risques de ré-identification

Synthèse et perspectives

Conclusion

1. Détection d'éléments identifiants faisable grâce à CamemBERT
2. Le pré-apprentissage n'a pas amélioré les performances
3. Le pré-apprentissage comporte des risques de ré-identification

$\Rightarrow$ une première étape de travaux à plus long terme

Synthèse et perspectives

Conclusion

1. Détection d'éléments identifiants faisable grâce à CamemBERT
2. Le pré-apprentissage n'a pas amélioré les performances
3. Le pré-apprentissage comporte des risques de ré-identification

$\Rightarrow$ une première étape de travaux à plus long terme

Perspectives

Synthèse et perspectives

Conclusion

1. Détection d'éléments identifiants faisable grâce à CamemBERT
2. Le pré-apprentissage n'a pas amélioré les performances
3. Le pré-apprentissage comporte des risques de ré-identification

$\Rightarrow$ une première étape de travaux à plus long terme

Perspectives

1. Amélioration du modèle
   - Entrainement/Validation sur plus de données
   - Apprentissage avec FlauBERT (Le et al. 2020), DrBERT (Labrak et al. 2023), ou des Large Language Models (Zhao et al. 2023)
   - Alignement avec l'AP-HP (Tannier et al. 2023)

Synthèse et perspectives

Conclusion

1. Détection d'éléments identifiants faisable grâce à CamemBERT
2. Le pré-apprentissage n'a pas amélioré les performances
3. Le pré-apprentissage comporte des risques de ré-identification

$\Rightarrow$ une première étape de travaux à plus long terme

Perspectives

1. Amélioration du modèle
   - Entrainement/Validation sur plus de données
   - Apprentissage avec FlauBERT (Le et al. 2020), DrBERT (Labrak et al. 2023), ou des Large Language Models (Zhao et al. 2023)
   - Alignement avec l'AP-HP (Tannier et al. 2023)
2. Améliorer la validation de nos modèles
   - Attaques sur réseaux de neurones (Berthelier, Boutet et Richard 2023)
   - Évaluer les risques de perte d'utilité

Synthèse et perspectives

Conclusion

1. Détection d'éléments identifiants faisable grâce à CamemBERT
2. Le pré-apprentissage n'a pas amélioré les performances
3. Le pré-apprentissage comporte des risques de ré-identification

$\Rightarrow$ une première étape de travaux à plus long terme

Perspectives

1. Amélioration du modèle
   - Entrainement/Validation sur plus de données
   - Apprentissage avec FlauBERT (Le et al. 2020), DrBERT (Labrak et al. 2023), ou des Large Language Models (Zhao et al. 2023)
   - Alignement avec l'AP-HP (Tannier et al. 2023)
2. Améliorer la validation de nos modèles
   - Attaques sur réseaux de neurones (Berthelier, Boutet et Richard 2023)
   - Évaluer les risques de perte d'utilité
3. Travaux Inter-CHU

Merci pour votre attention =)

5 Comparatif avec les résultats de l'AP-HP (Tannier et al. 2023)

6 Résultats avec DrBERT (Labrak et al. 2023)

Détection d'éléments identifiants

Modèle B			EDS-NLP	
Label	F1-Score		Label	F1-Score
Nom/Prénom	0.966 ± 0.013		First Name	0.986
			Last Name	0.984
IPP	0.966 ± 0.027		Patient ID	0.964
			NSS	0.923
NoDossier	0.866 ± 0.046		Visit ID	0.902
Téléphone	0.985 ± 0.012		Phone	0.964
EMail	0.988 ± 0.01		EMail	0.992
Date	0.967 ± 0.013		Date	0.995
			Birthdate	0.981
Code Postal	0.989 ± 0.01		ZIP	0.996
Ville	0.943 ± 0.02		City	0.985
Voie	0.947 ± 0.02		Address	0.989
Localité	0.945 ± 0.024			
Organisation	0.815 ± 0.035			
Site Web	0.974 ± 0.031			

Table – Résultats de l'évaluation du modèle B et de la partie Machine Learning de EDS-NLP entrainé sur ∼3400 textes

5 Comparatif avec les résultats de l'AP-HP (Tannier et al. 2023)

6 Résultats avec DrBERT (Labrak et al. 2023)

Connaissance du lexique/jargon médical

Texte	Solution	Modèle	Propositions (scores)
Contrôle 1 an 1/2 après <mask> radical avec Bricker	cystoprostatectomie	DrBERT-4GB	prostatectomie (0.88) mastectomie (0.05)
		DrBERT-7GB	résection (0.11) chirurgie (0.08)
A ce stade, il existe des ondes lentes <mask> dans les deux régions frontales	diphasiques	DrBERT-4GB	négatives (0.14) , (0.06)
		DrBERT-7GB	localisées (0.11) fortes (0.07)
Le contrôle de la <mask> huméro-basilique gauche est plutôt bon	fistule	DrBERT-4GB	conduction (0.08) pression (0.06)
		DrBERT-7GB	fonction (0.09) région (0.09)

Table – Exemples de complétions de textes par DrBERT-4GB et DrBERT-7GB

Détection des éléments identifiants

	Modèle *B*	DrBERT-4GB	DrBERT-7GB
Label	F1-Score	F1-Score	F1-Score
Nom/Prénom	**0.966 ± 0.013**	0.947 ± 0.015	0.946 ± 0.018
IPP	**0.966 ± 0.027**	0.948 ± 0.051	0.954 ± 0.05
NoDossier	**0.866 ± 0.046**	0.814 ± 0.056	0.799 ± 0.04
Téléphone	**0.985 ± 0.012**	0.978 ± 0.015	0.98 ± 0.014
EMail	**0.988 ± 0.01**	0.976 ± 0.019	0.976 ± 0.018
Date	**0.967 ± 0.013**	0.957 ± 0.021	0.953 ± 0.024
Code Postal	**0.989 ± 0.01**	0.976 ± 0.018	0.982 ± 0.012
Ville	**0.943 ± 0.02**	0.906 ± 0.026	0.899 ± 0.03
Voie	**0.947 ± 0.02**	0.924 ± 0.028	0.923 ± 0.039
Localité	**0.945 ± 0.024**	0.938 ± 0.026	0.939 ± 0.022
Organisation	**0.815 ± 0.035**	0.794 ± 0.032	0.795 ± 0.029
Site Web	**0.974 ± 0.031**	0.961 ± 0.033	0.966 ± 0.03
Moyenne	**0.946 ± 0.053**	0.927 ± 0.061	0.926 ± 0.064

Table – Résultats de l'évaluation des modèles *B*, DrBERT-4GB et DrBERT-7GB pour la détection d'éléments identifiants

Références I


Berthelier, Gaspard, Antoine Boutet et Antoine Richard (juill. 2023). "Privacy leakages on NLP models and mitigations through a use case on medical data". In : **COMPAS 2023 - Conférence francophone d'informatique en Parallélisme, Architecture et Système**. LISTIC - Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance / USBM - Université Savoie Mont Blanc. Annecy, France. URL : https://inria.hal.science/hal-04138528.


Bild, Raffael, Klaus A Kuhn et Fabian Prasser (2018). "SafePub : A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees.". In : **Proc. Priv. Enhancing Technol.** 2018.1, p. 67-87. DOI : 10.1515/popets-2018-0004.


Bourdois, Loick et al. (avr. 2021). "De-identification of Emergency Medical Records in French : Survey and Comparison of State-of-the-Art Automated Systems". en. In : **The International FLAIRS Conference Proceedings** 34. ISSN : 2334-0762. DOI : 10.32473/flairs.v34i1.128480. URL : https://journals.flvc.org/FLAIRS/article/view/128480 (visité le 02/02/2023).


Chazard, Emmanuel et al. (2014). "Proposal and evaluation of FASDIM, a Fast And Simple De-Identification Method for unstructured free-text clinical records". In : **International journal of medical informatics** 83.4, p. 303-312. DOI : 10.1016/j.ijmedinf.2013.11.005.

Références II


Grouin, Cyril (2013). "Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique. (Clinical Records De-Identification : Performances and Limits of Rule-based and Machine-Learning based Approaches)". Thèse de doct. Pierre et Marie Curie University, Paris, France. URL : https://tel.archives-ouvertes.fr/tel-00848672.


Johnson, Alistair E. W., Lucas Bulgarelli et Tom J. Pollard (2020). "Deidentification of Free-Text Medical Records Using Pre-Trained Bidirectional Transformers". In : **Proceedings of the ACM Conference on Health, Inference, and Learning**. CHIL '20. Toronto, Ontario, Canada : Association for Computing Machinery, p. 214–221. ISBN : 9781450370462. DOI : 10.1145/3368555.3384455.


Kajiyama, Kohei et al. (2020). "De-identifying free text of Japanese electronic health records". In : **Journal of Biomedical Semantics** 11.1, p. 1-12. DOI : 10.1186/s13326-020-00227-9.


Labrak, Yanis et al. (2023). "DrBERT : A Robust Pre-trained Model in French for Biomedical and Clinical domains". In : **Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL'23), Long Paper**. Toronto, Canada : Association for Computational Linguistics. DOI : 10.48550/arXiv.2304.00958.

Références III

Le, Hang et al. (2020). "FlauBERT : Unsupervised Language Model Pre-training for French". In : **Proceedings of The 12th Language Resources and Evaluation Conference**. Marseille, France : European Language Resources Association, p. 2479-2490. URL : https://www.aclweb.org/anthology/2020.lrec-1.302.

LeFevre, Kristen, David J DeWitt et Raghu Ramakrishnan (2006). "Mondrian multidimensional k-anonymity". In : **22nd International conference on data engineering (ICDE'06)**. IEEE, p. 25-25. DOI : 10.1109/ICDE.2006.101.

Lima, Salvador et al. (2020). "HitzalMed : Anonymisation of Clinical Text in Spanish". In : **Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020**. Sous la dir. de Nicoletta Calzolari et al. European Language Resources Association, p. 7038-7043. URL : https://aclanthology.org/2020.lrec-1.870/.

Martin, Louis et al. (juill. 2020). "CamemBERT : a Tasty French Language Model". In : **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online : Association for Computational Linguistics, p. 7203-7219. DOI : 10.18653/v1/2020.acl-main.645. URL : https://www.aclweb.org/anthology/2020.acl-main.645.

Références IV

Obeid, Jihad S et al. (2019). "Impact of de-identification on clinical text classification using traditional and deep learning classifiers". In : **Studies in health technology and informatics** 264, p. 283. DOI : 10.3233/SHTI190228.

Olatunji, Iyiola E et al. (2022). "A review of anonymization for healthcare data". In : **Big data**. DOI : 10.1089/big.2021.0169.

Sweeney, Latanya (2002). "k-anonymity : A model for protecting privacy". In : **International journal of uncertainty, fuzziness and knowledge-based systems** 10.05, p. 557-570. DOI : 10.1142/S0218488502001648.

Tannier, Xavier et al. (2023). "Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse". In : **arXiv preprint arXiv :2303.13451**. DOI : 10.48550/arXiv.2303.13451.

Xiao, Xiaokui et Yufei Tao (2008). "Dynamic anonymization : Accurate statistical analysis with privacy preservation". In : **Proceedings of the 2008 ACM SIGMOD international conference on Management of data**, p. 107-120. DOI : 10.1145/1376616.1376630.

Zhao, Wayne Xin et al. (2023). **A Survey of Large Language Models**. arXiv : 2303.18223 [cs.CL].