Résolution d'entités pour améliorer la qualité des données transactionnelles dans un système de santé

Tarek Benkhelif*, Wissam Siblini*

*komodo health france 115 Rue de l'Abbé Groult 75015 Paris
prénom.nom@komodohealth.com

Résumé. Les données de santé impliquent un réseau complexe d'entités telles que les patients, les prestataires de soin et les payeurs. Suivre chaque entité du système avec un haut degré de confiance est l'un des principaux défis en matière de qualité de données dans le domaine de la santé. Souvent désigné par "résolution d'entités", l'association précise des épisodes de soins de chaque patient est essentielle pour récupérer des historiques complets. Dans cet article applicatif sur les données transactionnelles du système de santé, nous dressons d'abord un inventaire des problèmes liés a la désambiguïsation des patients comme les dissociations d'identifiants et les collisions. Ensuite, sur un jeu de données réel enregistrant plus de 150 milliards d'interactions patient-professionnel de santé, nous proposons une approche pour reconnaître les identifiants de patients issus d'erreurs ou de dédoublements. Leur filtrage nous permet d'observer une réduction de 93% de l'écart entre le nombre de patients dans nos données et le nombre de patients attendus d'après le recensement Census.

1 Introduction

Chaque jour, un volume considérable de données de santé est généré aux États-Unis. Par exemple, l'année 2016 représente a elle seule plus de 3 000 milliards [1] de dollars de frais médicaux soumis par les hôpitaux. Les données se présentent sous de nombreuses formes : essais cliniques, publications scientifiques, dossier de santé électronique, facturation, demandes de remboursement, etc. Certaines de ces données sont accessibles au public, tandis que la plupart restent restreintes, isolées, par exemple, dans les systèmes hospitaliers ou d'assureurs. Avec pour mission de réduire le fardeau global de la maladie, notre initiative s'inscrit dans l'acquisition, le traitement, le raffinage et l'analyse d'un volume important de données transactionnelles de santé. L'objectif est d'avoir un processus de collecte donnant l'aperçu le plus complet et fiable des patients du système et de leur historique d'évènements. La capacité à relier ces entités à travers de multiples sources est alors une pierre angulaire de notre activité.

Dans cet article nous commençons par présenter le cadre général du système de santé en insistant sur les éléments qui font entrave à l'agrégation des sources de données comme

1. www.healthcarefinancenews.com/news/change-healthcare-analysis-shows-262-million-medical-claims-initially-denied-meaning-billions

la forte volumétrie, la pseudonymisation et les erreurs lors de saisies manuelles. Nous définissons ensuite la problématique, visant à détecter et à résoudre les faux identifiants. Nous dressons ensuite un panorama des approches de la littérature connexes à notre objectif. Enfin, nous décrivons en détail les modèles et heuristiques proposés pour la détection qui permettent d'améliorer significativement la qualité des les données.

2 Définition du problème

Un système de santé est constitué au minimum de l'interaction entre deux entités : un patient nécessitant des soins et un praticien de santé les fournissant (Wikipedia, 2022). Cette interaction est conclue par une rétribution financière. Dans une logique de gestion du risque pour les individus, la santé implique généralement des institutions intermédiaires (e.g. assurance maladie, mutuelles) qui contribuent au financement. Enfin, de nombreuses entités structurantes (hôpitaux, pharmacies, organisations gouvernementales, producteurs de médicaments/dispositifs, etc...) sont également acteurs dans ce système (Lameire et al., 1999).

L'accomplissement de soins génère de nombreuses données médicales (imageries, rapports, bilans biologiques) et un nombre important d'applications de l'apprentissage automatique (Miotto et al., 2018). En parallèle, lorsque le système de santé est mature et structuré, l'administratif constitue une source de données tout aussi exploitable (Milea, 2010; Sikka et al., 2005). Par exemple dans le système américain, chaque interaction donne lieu à une facture normalisée (*administrative claim*) identifiant le patient, le professionnel, et encodant les procédures effectuées, les médicaments prescrits, et les diagnostics selon des classes normalisées (ICD, NDC, PCS, etc... - Hirsch et al. (2016)). Ces données, combinées avec la littérature et l'open data, facilitent des applications autour de la prise en charge des patients, des maladies, et permettent d'améliorer le système de soin dans sa globalité. Par exemple certains s'intéressent à détecter des maladies rares, étudient l'utilisation de traitements, mesurent des statistiques, font des projections, etc. (Sikka et al., 2005) Cependant, le succès de ces différentes applications est conditionné par la qualité, l'exhaustivité et le bon traitement des données récoltées au travers de différentes sources (Milea, 2010).

2.1 Source de données et anonymisation

Tout acteur qui a un rôle dans le paiement des soins observe des données administratives : professionnel de santé, hôpital (*Provider*), mutuelle (*Payer*) et patient (Sanglier, 2011). Dans une situation idéale les données seraient homogènes et regroupées mais, en pratique, chaque acteur a une vue partielle et dans un format personnalisé. Des groupements de *Payers*/*Providers* ou intermédiaires supplémentaires effectuent déjà une première agrégation de vues. Dans ce papier, on se place dans un cadre où l'on a accès aux données de plusieurs de ces aggregateurs, que l'on appellera *sources*, et l'objectif est de les rassembler en résolvant les entités impliquées.

Au delà des hétérogénéités sur le format et la disponibilité des données (Morris et al., 2014) : (1) Les sources ont des intersections non nulles de patients et d'évènements (car, par exemple, un patient peut avoir plusieurs mutuelles, ou une visite peut être vue à la fois du côté *Provider* et *Payer*). (2) Des données sont masquées. Pour chaque patient, un identifiant primaire est généré par le hachage d'informations personnelles (e.g. date de naissance, nom, initiales). Ces dernières (i) ne sont parfois pas exclusives, ainsi un même identifiant peut être

attribué à plusieurs patients (*collisions*), ou (ii) sont parfois amenées à changer (e.g. mariage), ainsi un patient peut avoir plusieurs identifiants (*dissociations*). (3) Parfois des données temporaires (e.g. devis) générées entraînent l'apparition de nouveaux identifiants (*faux identifiants*) dans les données de certaines sources. Il est important de pouvoir résoudre les patients du système pour éviter d'avoir des vues partielles et ne pas limiter la performance des algorithmes d'apprentissage. Le problème de désambiguïsation de l'identité apparaît alors comme un des plus grands défis en matière de qualité des données de santé (Sukumar et al., 2015).

2.2 Problème : résolutions d'entités

On considère qu'on dispose, comme point de départ, de la concaténation brute $\mathcal{D}$ des données fournies par un ensemble de sources $\mathcal{S} = \{S_1, S_2, ..., S_s\}$. Chaque évènement $e \in \mathcal{D}$ (interaction patient - professionnel de santé) est associé à un identifiant patient $i = \text{id}(e)$.

On note $\mathcal{I}(\mathcal{D}) = \{i_1, i_2, ..., i_{n_I}\}$ l'ensemble des n_I identifiants patients présents dans $\mathcal{D}$ et $\mathcal{P}(\mathcal{D}) = \{p_1, p_2, ..., p_{n_P}\}$ l'ensemble des n_P véritables patients représentés par $\mathcal{D}$. On note $f : p \mapsto \mathcal{I}_p$ la fonction de correspondance entre un véritable patient et un sous ensemble d'identifiants de $\mathcal{I}(\mathcal{D})$, et $\overline{f} : i \mapsto \mathcal{P}_i$ la fonction de correspondance entre un identifiant et un sous ensemble de patients de $\mathcal{P}(\mathcal{D})$. On définit également les fonctions de cardinal associées $g : p \mapsto |f(p)|$ et $\overline{g} : i \mapsto |\overline{f}(i)|$. Dans ce cadre, on a alors :

- $\mathcal{F}(\mathcal{I}) = \{i \mid i \in \mathcal{I}(\mathcal{D}), \overline{g}(i) = 0\}$ l'ensemble des *faux identifiants*, i.e. n'ayant pas de correspondance avec de véritables patients.
- $\mathcal{C}(\mathcal{I}) = \{i \mid i \in \mathcal{I}(\mathcal{D}), \overline{g}(i) > 1\}$ l'ensemble des *collisions*.
- $\mathcal{H}(\mathcal{P}) = \{p \mid p \in \mathcal{P}(\mathcal{D}), g(p) > 1\}$ l'ensemble des *dissociations*. On notera $\mathcal{H}(\mathcal{I})$ l'union des identifiants associés aux patients de $\mathcal{H}(\mathcal{P})$.

Le véritable ensemble de patients $\mathcal{P}(\mathcal{D})$ ainsi que les fonctions de correspondance sont des inconnus. Cependant, pour certaines sous-parties de la population $\mathcal{I}_1 \subset \mathcal{I}(\mathcal{D})$, on dispose d'un registre de vérification d'existence (*ID verifier*) que l'on notera h_1. Pour chaque identifiant i de $\mathcal{I}_1$, $h_1(i)$ nous indique si i est l'identifiant d'un véritable patient ($h_1(i) = 1$ pour $i \in \mathcal{I}_1 \setminus (\mathcal{F}(\mathcal{I}) \cup \mathcal{H}(\mathcal{I}))$) ou non ($h_1(i) = 0$ pour $i \in \mathcal{I}_1 \cap \mathcal{F}(\mathcal{I})$). Pour les patients dissociés de $\mathcal{H}(\mathcal{P})$, l'*ID verifier* ne reconnaît qu'un seul identifiant ($h_1(i) = 1$) et pas les autres ($h_1(i) = 0$). Dans ce papier, on notre objectif sera de prédire automatiquement le label h_1 à partir des données associées à l'identifiant. Plus précisément, chaque identifiant i de $\mathcal{I}(\mathcal{D})$ a un ensemble de variables descriptives a_i (e.g. âge, genre) et un sous ensemble d'évènements $\mathcal{E}_i = \{e \mid e \in \mathcal{D}, \text{id}(e) = i\}$ de $\mathcal{D}$. Chaque évènement est lui-même décrit par plusieurs variables (e.g. date, zone géographique, code de diagnostic, montant payé, etc.). En appliquant des fonctions d'agrégation sur les éléments de $\mathcal{E}_i$, on peut construire un second ensemble de variables b_i pour l'identifiant i et obtenir un ensemble global x_i par concaténation avec a_i. Formellement, on souhaite construire un modèle l tel que, pour tout i dans $\mathcal{I}_1$, $l(x_i) = 1 - h_1(i)$ i.e. qui détecte les identifiants non connus par l'*ID verifier*.

3 État de l'art

La résolution d'entités vise à rassembler toutes les informations pertinentes sur une personne, une entreprise ou une entité et est souvent constituée de quatre étapes (Binette et Steorts, 2020). Dans la première, "l'alignement des attributs et des schémas", les enregistrements sont

analysés pour identifier les ensembles d'attributs communs entre les différentes sources. Dans la deuxième, "le blocage", les enregistrements similaires sont regroupés en blocs. Les éléments apparaissant dans le même bloc seront comparés ; les autres sont automatiquement considérés comme des non-concordances. À l'étape suivante, "la résolution des entités", les enregistrements co-référents sont identifiés. Enfin, à la quatrième étape, de "fusion, ou canonisation", les entités jugées correspondantes à l'étape 3 sont fusionnées pour produire un seul enregistrement représentatif. Plus spécifique à notre problème, la résolution d'entités pseudonymisées, ou dé-anonymisation, consiste, elle, à réconcilier des enregistrements avec la particularité que les attributs identifiants (e.g. le nom) et quasi-identifiants (e.g. le genre) sont obfusqués. Wang et al. (2018) classent les approches en trois catégories, en fonction des données utilisateur exploitées : (i) contenu (e.g. activités de l'utilisateur), (ii) profil (e.g. nom, genre et 'âge), et (ii) réseau (relations entre les utilisateurs). Nous pouvons difficilement appliquer : (1) Les approches de blocage naïves, car nous n'avons pas à disposition les clés de blocage généralement utilisées (date de naissance, adresse, etc.). (2) Les approches de dé-anonymisation qui sont principalement orientées vers des données de mobilité. (3) Les approches pour les données de santé car elles visent surtout à quantifier le problème et préconiser des améliorations à l'étape de génération des identifiants (inaccessible à notre niveau). Dans la suite, nous proposons donc une approche spécifique à notre application.

4 Travaux réalisés

4.1 Description et préparation des données

Medical et pharmacy claims L'ensemble de données $\mathcal{D}$ est composé principalement de deux types d'évènements (ou *claims*) : *medical* et *pharmacy*. Les *medical claims* décrivent le passage d'un patient auprès d'un médecin et contiennent des informations sur le "patient" (identifiant, informations démographiques, mutuelle, etc.), le professionnel de santé (numéro d'identification national, organisation parente, zone géographique, etc.) et leur interaction (date, procédures réalisées, diagnostics, facturation, etc.). Les *pharmacy claims* décrivent le passage d'un patient auprès d'un fournisseur d'équipement médical et de médicaments (e.g. une pharmacie) et contiennent des informations sur le patient, sur le fournisseur et sur leur interaction (date, numéro NDC de médicament/équipement, facturation, etc.). La liste des variables contenues dans une *claim* est publiquement disponible. On s'appuiera sur ces variables pour construire les vecteurs x_i descriptifs des identifiants. Notons que chaque *claim* est associée à une source $S \in \mathcal{S}$.

Construction du jeu de données On s'attend à ce qu'un faux identifiant soit peu représenté dans les données. Ainsi, trois variables semblent intéressantes à calculer : le nombre de pharmacy claims, de medical claims et de sources. On peut aller plus loin et même fournir au modèle la liste des sources de l'identifiant (sous forme d'un encodage disjonctif), car on remarque sur la Figure 5 que la proportion de mauvais identifiants varie d'une source à l'autre. Cela est dû au fait qu'elles ont elles-mêmes accès à une vue plus ou moins fine des données, et utilisent des mécanismes internes de nettoyage. Les trois premières variables seront utilisées dans la suite pour définir des heuristiques simples servant de baselines. On les complétera, pour notre modèle, de variables basées sur des informations temporelles (date du

premier et dernier évènement, entendue temporelle), des informations médicales (nombre et entropies des codes diagnostics, procédures et médicaments), des informations sur les professionnels de santé (localisations, spécialités), sur le coût financier (statut des transactions, dépenses moyennes/totales) et sur le patient (code zip tronqué, genre). Au total, on crée un premier ensemble de $\dim(x_i) \approx 10^3$ variables. Concernant la variable à prédire, on dispose de la vérité terrain h_1 et on notera la proportion de mauvais identifiants $\delta = \frac{\sum_{i \in \mathcal{I}_1}(1-h_1(i))}{|\mathcal{I}_1|}$. Pour des raisons de confidentialité, les données et leurs dimensions précises ne peuvent être partagées. Cependant, le tableau 1 fournit des ordres de grandeur, qui permettent notamment d'expliquer de rendre compte de la magnitude du problème à résoudre ($\delta = 0.535$).

Sélection de variables Les données ont une forte volumétrie ($\sim$ une centaine de milliards d'évènements) représentant des dizaines de téraoctets de mémoire. Construire les x_i avec des agrégations coûteuses et des jointures entre bases de données de millions/milliards d'enregistrements, représente un défi technique. Nous souhaitons donc réduire l'ensemble de variables. Plus précisément, sur un échantillon réduit $\mathcal{I}'_1 \subset \mathcal{I}_1$ du jeu de données, nous entraînons des modèles à base d'arbres de décision (Random Forest, XGBoost) et utilisons l'importance des variables (réduction moyenne de l'impureté des noeuds où elles sont impliquées) pour effectuer une sélection. Les variables importantes sont par exemple liées à l'étendue temporelle où l'identifiant est observé, à son nombre de sources, au nombre distinct de praticiens ou d'hôpitaux dans ses claims. Ces comptages d'informations brutes sont souvent rapides à calculer. Ainsi, en sélectionnant itérativement les variables les plus importantes, nous parvenous sur $\mathcal{I}'_1$ à réduire le temps de calcul de 96% tout en conservant 98% de la performance prédictive en validation croisée. L'ensemble x_i obtenu finalement compte seulement ≈ 20 variables.

	Notation	Ordre de grandeur
Nombre d'évènements	$\lvert\mathcal{D}\rvert$	10^{11}
Plus petite source	$\min_{S \in \mathcal{S}} \lvert\mathcal{D}_S\rvert$	10^8
Plus grande source	$\max_{S \in \mathcal{S}} \lvert\mathcal{D}_S\rvert$	10^{10}
Nombre d'identifiants	$n_I = \lvert\mathcal{I}(\mathcal{D})\rvert$	10^8
Nombre d'identifiants dans $\mathcal{I}_1$	$\lvert\mathcal{I}_1\rvert$	10^7
Proportion de mauvais identifiants ou dissocations	δ	0.535
Nombre de variables descriptives d'un évènement	$\dim(e)$	30

TAB. 1 – Dimensions principales du jeu de données utilisé.

4.2 Identification des faux identifiants et des patients dissociés

Baseline 1 En analysant les données, on note un nombre important d'identifiants $i \in \mathcal{I}(\mathcal{D})$ tels que $|\mathcal{E}_i| = 1$, i.e. n'apparaissant qu'une fois. Ce comportement est très rare pour un véritable patient, mais représente plutôt la signature d'un identifiant généré par une transaction temporaire ou une erreur ponctuelle. En effet, la Figure 1 montre la proportion d'identifiants de $\{i \mid i \in \mathcal{I}_1, |\mathcal{E}_i| = 1\}$, appelés singletons, pour lesquels $h_1(i) = 1$ et $h_1(i) = 0$. Au total, 94.6% (précision) des singletons sont des faux identifiants ou des dissociations. Par ailleurs, ces identifiants représentent une proportion $R = \frac{\sum_{i \in \mathcal{I}_1, |\mathcal{E}_i|=1}(1-h_1(i))}{\sum_{i \in \mathcal{I}_1}(1-h_1(i))} = 0.327$ (rappel) de la

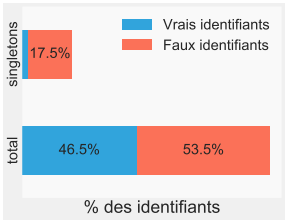

FIG. 1 – Taux de faux identifiants parmi les singletons.

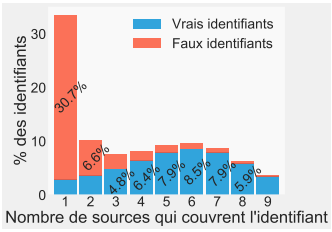

FIG. 2 – Taux de faux identifiants en fonction de la couverture (nombre de sources).

cible du problème.

Baseline 2 En complément nous proposons une seconde heuristique. On observe que les identifiants i cibles, ayant un ou plusieurs évènements, n'apparaissent très souvent que dans une seule source (Figure 2). Ainsi, une baseline qui considère un tel identifiant comme cible, permet d'atteindre un niveau de rappel de 57.4% avec une précision de 85.3%.

Modèle proposé Le jeu de données est découpé en deux parties apprentissage (80%) et test (20%). On considère uniquement les modèles Random Forest (RF) et XGBoost, qui se sont montrés plus performants que des alternatives (régression logistique, SVM, NN) dans des essais préliminaires. On réalise une sélection d'hyperparamètres avec validation croisée (5-fold) sur la partie apprentissage. Le modèle XGBoost est finalement retenu et atteint sur le jeu de test, le meilleur rappel de 87% (pour une précision de 95%), le meilleur temps d'apprentissage (5m contre 2h pour RF), la meilleur utilisation mémoire (environ 1Mb seulement contre 20Gb pour RF), et se compare favorablement aux baselines (Tableau 2). La Figure 3 détaille les taux de vrai et faux négatifs/positifs. On peut enrichir ces résultats en les analysant par source de données (Figure 5). On note d'abord que, bien que le taux global d'identifiants cibles δ soit de 53.5%, le taux pour chaque source est moindre (de quelques pourcents à environ 34% pour la moins 'propre'). Cela est dû au fait que, comme mentionné précédemment, les mauvais identifiants tendent à être présents que dans une seule source. Ainsi, lors de l'union de sources, le nombre de mauvais identifiants augmente plus vite que le nombre d'identifiants légitimes. La précision étant dépendante du taux d'exemples positifs Siblini et al. (2020), la performance du modèle est la meilleure pour la source la moins propre mais reste en dessous de la performance observée globalement.

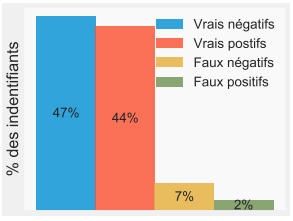

FIG. 3 – Performance de XGBoost

	Precision	Rappel
Baseline 1	0.946	0.327
Baseline 2	0.853	0.574
Modèle	**0.950**	**0.870**

TAB. 2 – Comparaison entre le modèle proposé et les deux baselines.

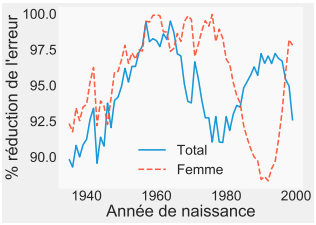

FIG. 4 – Réduction du taux d'erreur entre le compte d'identifiants et Census grâce au modèle.

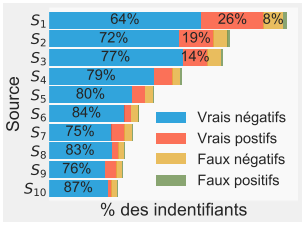

FIG. 5 – Performance du modèle pour les 10 plus grandes sources. La largeur de la barre représente la taille de la source.

Outre la performance mesurable sur le sous ensemble $\mathcal{I}_1$, on peut analyser les prédictions du modèle sur la population globale. On ne pourra pas vérifier leur pertinence vis-à-vis d'une vérité terrain exacte mais constater si le compte d'identifiants prédits comme non-cible se rapproche de la population attendue pour les États-Unis (en utilisant les données du recensement *Census*). Après mesure de l'erreur initiale entre le nombre d'identifiants et la population attendue, on montre, pour chaque année de naissance, la réduction relative de cette erreur grâce à la suppression des identifiants prédits comme cible (Figure 4). Cette réduction est drastique (de presque 100% parfois) et permet d'atteindre un taux d'erreur inférieur à δ rendant les données exploitables. De façon intéressante, on observe que l'erreur reste plus importante pour les identifiants de sexe féminin nés entre 1975 et 2000 [2]. De plus, en inspectant le profil des vrais positifs et des faux négatifs, on remarque que les premiers ont souvent peu d'évènements et ressemblent plus à des erreurs ponctuelles ou des faux identifiants alors que les seconds ont beaucoup d'évènements décrits par plusieurs sources et ressemblent à de véritables patients. On pourra alors (1) considérer en première approximation que les profils identifiés par le modèle peuvent être "supprimés" car ils ajoutent peu de valeur et sont plus probablement des faux identifiants, et (2) s'intéresser aux faux négatifs comme source de vérité pour résoudre le problème plus "difficile" des dissociations.

5 Conclusion

Dans cet article, nous avons identifié et formalisé les problèmes de qualité de données qui surviennent dans l'agrégation de données médicales transactionnelles. Nous avons traité la tâche primordiale de détection de mauvais identifiants en proposant un processus de construction de variables et un modèle d'extreme gradient boosting, économes en ressources. Nous nous intéresserons dans de prochains travaux à la résolution des collisions et des dissociations.

Références

Binette, O. et R. C. Steorts (2020). (almost) all of entity resolution. *arXiv e-prints*, arXiv–2008.

2. Cette sous population est plus susceptible d'avoir changé de nom (e.g. mariage/divorce) que la reste de la population sur la fenêtre temporelle dans laquelle nos données ont été générées (dernière décennie).

Hirsch, J., G. Nicola, G. McGinty, R. Liu, R. Barr, M. Chittle, et L. Manchikanti (2016). Icd-10 : history and context. *American Journal of Neuroradiology 37*(4), 596–599.

Lameire, N., P. Joffe, et M. Wiedemann (1999). Healthcare systems—an international review : an overview. *Nephrology Dialysis Transplantation 14*(suppl_6), 3–9.

Milea, D. D. (2010). *Usage et mésusage dans la prescription des antidépresseurs : l'apport des bases de données*. Ph. D. thesis, Université Claude Bernard-Lyon I.

Miotto, R., F. Wang, S. Wang, X. Jiang, et J. T. Dudley (2018). Deep learning for healthcare : review, opportunities and challenges. *Briefings in bioinformatics 19*(6), 1236–1246.

Morris, G., G. Farnum, S. Afzal, C. Robinson, J. Greene, et C. Coughlin (2014). Patient identification and matching final report. *Office of the National Coordinator for Health Information Technology*, 1–93.

Sanglier, T. (2011). *Comparaison de la prise en charge de la dépression chez le sujet âgé et l'adulte non âgé par l'utilisation de systèmes administratifs automatisés.* Ph. D. thesis, Université Claude Bernard-Lyon I.

Siblini, W., J. Fréry, L. He-Guelton, F. Oblé, et Y.-Q. Wang (2020). Master your metrics with calibration. In *International Symposium on Intelligent Data Analysis*, pp. 457–469. Springer.

Sikka, R., F. Xia, et R. E. Aubert (2005). Estimating medication persistency using administrative claims data. *Am J Manag Care 11*(7), 449–457.

Sukumar, S. R., R. Natarajan, et R. K. Ferrell (2015). Quality of big data in health care. *International journal of health care quality assurance*.

Wang, H., C. Gao, Y. Li, G. Wang, D. Jin, et J. Sun (2018). De-anonymization of mobility trajectories : Dissecting the gaps between theory and practice. In *The 25th Annual Network & Distributed System Security Symposium (NDSS'18)*.

Wikipedia (2022). Health system — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=Health\%20system&oldid=1112316532`. [Online ; accessed 03-October-2022].

Summary

Healthcare data involves a complex network of entities such as patients, providers and payers. Tracking every entity in the system with a high degree of confidence is one of the biggest data quality challenges in healthcare. Often referred to as "entity resolution", the precise association of each patient's care episodes is essential to retrieving complete histories. In this applicative paper on transactional data of the healthcare system, we first draw up an inventory of problems related to patient disambiguation, such as identifier dissociations and collisions. Then, on a real dataset with more than 150 billion patient-healthcare professional interactions, we propose approaches to correctly re-associate the interactions to a unique patient identifier. The results obtained show a reduction of 93% in the gap between the number of patients observed and the number of patients expected according to Census.