

# 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 à 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 à elle seule plus de 3 000 milliards<sup>1</sup> 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 raffinement 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'événements. 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](http://www.healthcarefinancenews.com/news/change-healthcare-analysis-shows-262-million-medical-claims-initially-denied-meaning-billions)