Génération d'un SNDS synthétique à partir de données ouvertes

Thomas Guyet*

*Institut Agro / IRISA UMR 6974
thomas.guyet@irisa.fr

1 Introduction

Le SNDS (Système National des Données de Santé) désigne ici la base de données contenant les informations de remboursement de soins par l'assurance maladie également connue sous le nom SNIIRAM (Bezin et al., 2017). Cette base de données contient des informations riches permettant de répondre à de nombreuses questions épidémiologiques et médico-économiques. De part son contenu médical sensible et personnel, leur usage est restreint, ce qui limite les possibilités d'expérimentation de nouveaux algorithmes sur ces données.

L'approche proposée dans cet article vise à générer des données synthétiques pour alimenter une base de données, d'une part, respectant la structure originale du SNDS et, d'autre part, reproduisant des statistiques connues sur les agrégats de variables épidémiologiques en s'appuyant pour cela sur les données ouvertes (*open data*). Les mesures de protection statistique mises en œuvre sur les données ouvertes librement accessibles assurent ainsi leur réutilisabilité dans le respect de la vie privée.

2 Génération d'un SNDS synthétique

Le processus de génération en quatre phases principales est illustré dans la Figure 1 : (i) création de la structure générale de la base de données à partir du schéma de la base de données, (ii) chargement des nomenclatures qui alimentent 416 tables, (iii) reconstruction de distributions des variables à partir de données ouvertes, (iv) simulation de nouvelles bases aléatoires (12 tables).

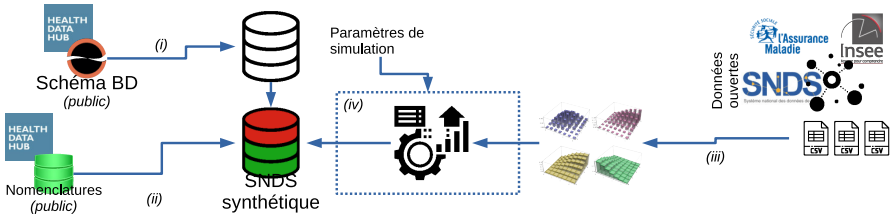

FIG. 1 – *Illustration du processus de génération d'un jeu de données synthétiques.*

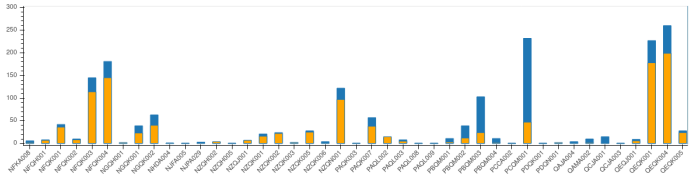

FIG. 2 – *Répartition des actes CCAM de NFKA008 à QEQK005 en population synthétique (en Orange) pour les département 22 et 35 et en population réelle nationale (en bleu).*

Simulation Le génération de données synthétiques est configurée en précisant la liste des départements à considérer ainsi que l'année à simuler. Ces paramètres servent à générer une population synthétique dont on connaît, à une date donnée la structure par sa localisation (code commune), son sexe et son âge quinquennal (par classe d'âge de 5 ans). Pour tous les bénéficiaires de cette population, on génère ensuite leurs éventuelles affections de longue durée (ALD) de manière aléatoire. Finalement, un médecin traitant est attribué à chaque bénéficiaire parmi les médecins de sa commune de résidence (sinon de son département).

Les étapes suivantes génèrent l'historique médical de chaque patient : soin à l'hôpital et « en ville ». Dans les deux cas, le principe de la génération synthétique d'une collection de soins comprend trois dimensions : (i) déterminer si et combien de prestations d'une nature donnée (hospitalisation, visite, acte, etc.) un bénéficiaire a eu dans l'année, (ii) déterminer le code de chacune de ces prestations (code CIM, CCAM, CIP ou NABM) en utilisant des probabilités conditionnelles d'observer ces codes, (iii) lier cette prestation aux informations complémentaires (*p. ex.* spécialité de l'exécutant). Les quantités et les probabilités sont conditionnées aux caractéristiques du patient : son âge, son sexe ou sa localisation de résidence. La granularité de la localisation est par défaut celle du département.

Estimation des distributions Les distributions statistiques et des estimations de quantités de prestations sont estimées à partir des données ouvertes. Le défi est lié à l'incomplétude des informations disponibles. Pour le résoudre, différentes techniques Bayésiennes ont été utilisées pour estimer des distributions jointes à partir de distributions marginales.

3 Résultats

Les outils sont mis à disposition en ligne [1] et permettent à chacun de générer un SNDS synthétique au format SQLite. La Figure 2 illustre les distributions réelles et synthétiques des actes techniques médicaux (CCAM). On constate que les proportions des actes sont plutôt fidèles à la réalité, mais les quantités sont globalement sous-estimées par rapport à la réalité.

Références

Bezin, J., M. Duong, R. Lassalle, C. Droz, A. Pariente, P. Blin, et N. Moore (2017). The national healthcare system claims databases in France, SNIIRAM and EGB : Powerful tools for pharmacoepidemiology. *Pharmacoepidemiology and drug safety 26*(8), 954–962.

1. `https://gitlab.inria.fr/tguyet/medtrajectory_datagen`