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).

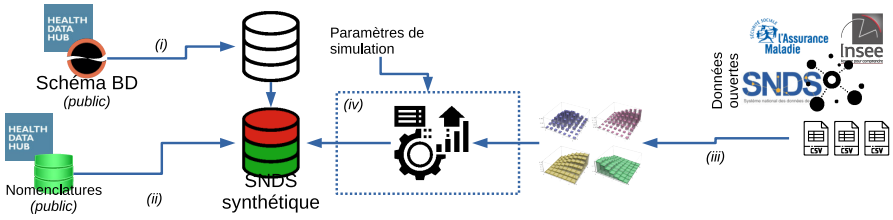

FIG. 1 – *Illustration du processus de génération d'un jeu de données synthétiques.*