Offre de stage de fin d'étude (Bac+5), mention Informatique/Bio-statistique

> LIEU DU STAGE

INSERM (Institut National de la Santé et de la Recherche Médicale)

CépiDc (Centre d'épidémiologie sur les causes médicales de décès)

80, rue du Général Leclerc, 94270 Le Kremlin-Bicêtre

> INTITULE DU STAGE

Anonymisation de base de données médicales et administratives :

> DOMAINE(S) COUVERT(S) PAR LE STAGE

Statistique, Épidémiologie, Mathématiques (optimisation convexe et recherche opérationnelle), Informatique,

Contexte

Les données sur les causes médicales de décès sont considérées comme les données de santé publiques de référence au niveau national et international. Utilisée par la recherche ainsi que par les multiplies acteurs de la santé publique, l'accès à ces données est un enjeu primordial. Le CépiDc est l'unité de service Inserm chargée de produire annuellement de la production de la Base des Causes Médicales de Décès (BCMD), de sa diffusion et du support technico-scientifique relatif à son exploitation. Pour accéder à ses données il existe deux cas de figures : Soit on dispose d'une autorisation CNIL, et dans ce cas l'accès aux données peut se faire au sein d'une infrastructure sécurisée, sinon il est toujours possible d'avoir un accès si les données fournies sont anonymes. Ce dernier point est controversé, en effet pendant longtemps la seule procédure d'anonymisation a consisté à fournir des données agrégées en faisant attention à respecter des critères (principalement trois : la k-anonymité, la l-diversité et l'intégrité des données). Cependant ces mesures ne garantissent pas en théorie la protection de la confidentialité des données tout en dégradant, parfois sévèrement, l'utilité statistique de ces dernières. Au-delà des causes médicales de décès, cette problématique couvre l'ensemble des données dites confidentielles ou sensibles dans le contexte d'un accès ouvert aux données promulgué par le volet open data la Loi pour une république numérique de 2016.

Plus récemment, la formalisation mathématique de ce problème a permis de proposer une solution permettant de fournir des données bruitées au niveau individuel garantissant ainsi la confidentialité différentielle tout en altérant au minimum le contenu. Dans ce cadre, le CépiDc souhaite mettre à disposition un outil open data pour avoir accès à ces données non réidentifiantes des personnes physiques.

Objectifs

Dans un premier temps, le stagiaire aura pour objectifs de recenser l'ensemble des méthodes d'anonymisation adaptées aux données du CépiDc. L'idée est de tester ces méthodes et de les évaluer afin d'en retenir une (ou plusieurs) adaptées à notre activité (le stagiaire devra produire des statistiques de mortalité). Enfin le stagiaire pourra participer à la conception du système qui sera mis à la disposition des utilisateurs de ces données.

Méthodes

Le stagiaire devra faire une synthèse des méthodes d'anonymisation des bases de données respectant la confidentialité différentielle lorsque ces dernières contiennent des données catégorielles. Ces méthodes seront ensuite évaluées sur la base des causes médicale de décès, et comparées selon différents critères à définir en fonction de l'utilité (distribution de la mortalité par causes de décès, distribution spatiale des décès, limiter les biais et la perte de puissance statistiques sur les taux standardisés, ou plus généralement les régressions statistiques). Ces critères seront basés sur une analyse de la dégradation de ces utilités par cause de décès et par unité spatiale sur différentes échelles (France entière, régions, département, etc.).

En fonction du type de sortie (base agrégées ou contenant des microdonnées), l'anonymité effectivement obtenue sera évaluée, d'abord en fonction des critères classiques d'anonymité et de diversité, puis par des méthodes d'inférence (en utilisant des méthodes probabilistes d'appariement dans le cas des microdonnées, et en mesurant la perte de précisions des attaques par inférence pour les bases agrégées).

Enfin, si les résultats sont positifs, le stagiaire devra concevoir (en utilisant R) une solution qui permettra l'extraction de bases de données anonymisées.

Résultats attendus :

- Revue bibliographique sur les méthodes d'anonymisation.
- Choix des fonctions d'utilité adaptées et études comparatives.
- Implémentation d'un algorithme d'appariement
- Implémentation d'un outil de requête de données anonymisées.

Le cas échéant, degré prévisible de confidentialité du rapport de stage

☐ **extrême** ☑ **moyen** ☐ **faible**

Connaissances et aptitudes recherchées chez le stagiaire :

Connaissances des outils suivants :

- *Biostatistique et principes de l'apprentissage statistique.*

Aptitudes :

- *Logiciels : R, Python*
- *Aisance en programmation*
- *Manipulation de bases de données éventuellement volumineuses*
- *Anglais lu et écrit courant*

> ENVIRONNEMENT DE LA MISSION

Intitulé, activité, compétences statistiques de l'unité d'accueil et du maître de stage :

INSERM-CépiDc (Centre d'épidémiologie sur les causes médicales de décès). Les missions du CépiDc sont :

- la production de la base nationale des causes médicales de décès,
- la diffusion de cette base pour des objectifs de recherche et de santé publique,
- la production d'analyses statistiques et de recherche sur cette base de données.

Cette dernière mission a donné lieu à l'application des méthodologies statistiques adaptées pour de nombreuses publications dans des revues scientifiques internationales.

Karim Bounebache, le responsable de ce stage, est statisticien-expert en apprentissage statistique et docteur en mathématiques appliquées. Au sein du CépiDc, la mission du stagiaire sera effectuée en collaboration avec l'équipe en charge des recherches et développements qui comprend deux ingénieurs experts en statistique et épidémiologie ainsi qu'un doctorant spécialisé en apprentissage machine et calcul haute performance.

Ressources mises à la disposition du stagiaire :

Le stagiaire disposera d'un bureau, d'un ordinateur puissant, et du logiciel R et Python. Il aura accès à un serveur de calcul haute performance.

La gratification du stage est d'environ 500€ / mois

Durée du stage : 6 mois

> PERSONNE(S) A CONTACTER

M BOUNEBACHE Saïd-Karim,

INSERM-CépiDc

said.bounebache@inserm.fr

M GHOSN Walid,

INSERM-CépiDc

walid.ghosn @inserm.fr