Université Claude Bernard Lyon 1

DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **23 septembre 2025**

Nom de famille et prénom de l'auteur. e : **Monsieur Oleksandr SOROCHYNSKYI**

Titre de la thèse **:** Contribution à l'analyse des risques de mortalité et de morbidité en assurance à l'aide de données de santé

Résumé

La disponibilité croissante de données médicales offre des opportunités pour l'exploitation secondaire de ces dernières, notamment pour la construction d'indicateurs de santé, sans nécessiter la mise en place coûteuse de nouveaux dispositifs de collecte. Ces indicateurs jouent un rôle crucial dans trois grands enjeux des systèmes de santé contemporains : éclairer les décisions des politiques publiques, améliorer l'efficience du système de soins, et cibler les actions de prévention. Cette thèse s'inscrit dans ce cadre général en explorant différentes stratégies d'exploitation secondaire des données de santé, en mettant l'accent sur les cas où cette exploitation soulève des défis méthodologiques importants. Les contributions proposées relèvent principalement de la méthodologie statistique appliquée, adaptée à des contextes variés, et ont pour but de démontrer le potentiel de telles données dans la production d'indicateurs utiles et fiables. Les travaux sont structurés autour de trois contextes d'application distincts. Le Chapitre 1 propose une méthodologie de calcul de l'espérance de vie en bonne santé à partir des données hospitalières issues du PMSI. En appliquant un modèle de Cox à la durée de vie sans survenue de pathologies majeures, il devient possible d'analyser un indicateur de type disease-free life expectancy, intégrant à la fois la morbidité et la mortalité. L'analyse confirme le rôle central de certains facteurs comportementaux, tout en permettant une quantification fine des disparités entre sexes et territoires. Ce travail illustre ainsi le potentiel des données hospitalières pour produire des indicateurs robustes, exploitables par les décideurs publics, et complémentaires aux indicateurs classiques fondés sur les enquêtes. Le Chapitre 2 s'intéresse aux âges biologiques comme indicateurs individuels de santé, conçus pour refléter l'état physiologique d'un individu mieux que son âge chronologique. À partir des données NHANES, plusieurs méthodes de calcul d'âges biologiques sont comparées selon leurs associations avec la mortalité, la morbidité et divers facteurs de risque. Une procédure d'imputation multiple permet de dépasser les limites du plan d'enquête. La comparaison est menée dans une perspective de prévention, avec l'objectif de construire un indicateur prédictif et compréhensible. L'analyse révèle qu'en l'absence de recalibrage, les âges biologiques ne permettent pas d'améliorer la prédiction du nombre de décès à partir d'une table de mortalité. L'intégration de la loi de mortalité dès la construction de l'indicateur permet de rendre ces âges biologiques opérationnels dans un cadre assurantiel, ouvrant la voie à des stratégies de prévention personnalisées. Le Chapitre 3 prolonge la réflexion dans un contexte assurantiel, en étudiant l'évaluation de la santé à travers les postes de préjudice extraits de rapports d'expertise médicale, à l'aide de techniques de traitement automatique du langage. Après

une phase d'extraction des montants de préjudice, la tâche est formulée comme un problème de régression supervisée. Plusieurs modèles sont testés, allant de modèles linéaires aux réseaux de neurones basés sur CamemBERT. Les résultats montrent des performances modérées (R^2 entre 20% et 40%). Bien que cette tentative ne débouche pas encore sur un modèle opérationnel, elle permet d'explorer les approches possibles et d'identifier les conditions nécessaires à une poursuite réaliste de cette tâche : diversification des sources textuelles, disponibilité d'une référence humaine, et capacité des modèles à se généraliser à des documents produits plus tôt dans le processus d'indemnisation. Cette thèse illustre ainsi la diversité des approches possibles pour l'exploitation secondaire des données de santé. Chaque contexte impose des contraintes méthodologiques spécifiques, mais tous montrent le potentiel de ces données pour renforcer les capacités de pilotage, de prévention et l'efficience des systèmes de santé.

Mots-clés : Données de santé,Indicateurs de santé,Espérance de vie en bonne santé,Âge biologique,Traitement automatique du langage,Assurance santé,