THESE DE DOCTORAT DE

NANTES UNIVERSITE

ECOLE DOCTORALE N° 605
Biologie-Santé
Spécialité : Epidémiologie

Par **Matilde KARAKACHOFF**

Des bases de données aux entrepôts hospitaliers : Évolution, intégration et valorisation de données massives pour la recherche en santé

Une thèse sur l'évolution des bases de données en santé et leur mobilisation

Thèse présentée et soutenue à Nantes, le 1 décembre 2026
Unité de recherche : Inserm UMR 1087 / CNRS UMR 6291, Unité de recherche de l'institut du thorax

Rapporteurs avant soutenance :

Pr Leslie Guillon-Grammatico	PU-PH, Université de Tours
Pr Anne-Sophie Jannot	PU-PH, Université Paris Cité

Composition du Jury :

Président :	--	-- *(à préciser après la soutenance)*
Examinateurs :	Pr Corinne Miral	Professeur Nantes Université
	Dr Antoine Neurz	CDO, Stane Group
Dir. de thèse :	M Jean-Jacques Schott	Directeur de recherche, Institut du Thorax, Inserm

Invité(s)

Pr Pierre-Antoine Gourraud	PU-PH, Nantes Université
M Adrien Bazoge	IR, CHU de Nantes

TITRE : Des bases de données aux entrepôts hospitaliers : Évolution, intégration et valorisation de données massives pour la recherche en santé

Mots clés : entrepôt de données de santé, données massives, données de vie réelle, gouvernance, exploitation, valorisation

Résumé : Cette thèse retrace l'évolution de la production et de l'exploitation des données en santé, depuis la construction de bases dédiées jusqu'au développement des entrepôts de données de santé (EDS). Les travaux sur la surveillance des addictions, la génomique et les anévrismes intracrâniens montrent comment l'accroissement du volume, de l'hétérogénéité et de la dimensionnalité des données nécessite d'adapter leur structuration, les infrastructures et les méthodes d'analyse. Le développement de l'Entrepôt de Données de Santé du CHU de Nantes ouvre de nouvelles possibilités de réutilisation des données de soin pour la recherche, notamment pour construire des populations d'étude et exploiter les données non structurées par le traitement automatique du langage et les modèles de langage. Ces travaux soulignent que la production de connaissances robustes repose sur la qualité et la contextualisation des données, la maîtrise des biais, la validation des méthodes et une gouvernance adaptée. Les perspectives concernent l'interopérabilité et la fédération des EDS, l'articulation des données de soin et de recherche, ainsi que l'intégration de données cliniques, génomiques, environnementales, sociales et territoriales pour une recherche reproductible et utile en santé publique.

Title : From databases to hospital data warehouses : Evolution, Integration and use of big data in health research

Keywords : health data warehouse, big data, real-world data, governance, data use, data valorisation

Abstract : This thesis traces the evolution of health data production and use, from the development of dedicated databases to the implementation of health data warehouses (HDWs). Studies on addiction surveillance, genomics, and intracranial aneurysms illustrate how the increasing volume, heterogeneity, and dimensionality of data require adaptations in data structuring, infrastructures, and analytical methods. The development of the Nantes University Hospital Health Data Warehouse opens new opportunities for reusing healthcare data for research, particularly to construct study populations and leverage unstructured data through natural language processing and language models. These studies highlight that producing robust knowledge relies on data quality and contextualization, control of bias, methodological validation, and appropriate data governance. Future perspectives include improving interoperability and the federation of HDWs, bridging healthcare and research data, and integrating clinical, genomic, environmental, social, and geographical data. Together, these developments aim to support reproducible research and the generation of robust and meaningful knowledge for clinical research and public health.