Gestion des données manquantes et des dérives de données pour la réutilisation de données de santé de vie réelle : détection, caractérisation et mise en qualité

Il existe un intérêt croissant pour la réutilisation des données en clinique, notamment au travers de l'exploitation des données des entrepôts, en raison de leur caractère exhaustif, multimodal et de leur profondeur temporelle [Meystre, 2017]. Par son fort partenariat avec le Centre des Données Cliniques, responsable de l'entrepôt de données du CHU de Rennes (CDC, CHU Rennes), le Laboratoire du Traitement du Signal et de l'Image (LTSI, INSERM U1099) contribue aux développements et à l'implémentation de méthodes de structuration, d'exploration et d'exploitation de ce type de données (i.e., dossiers médicaux électroniques, biologie). En particulier, l'équipe DOMASIA est spécialisée dans les systèmes d'information de santé apprenants (SIAS) et ses travaux se déclinent en trois axes (https://ltsi.univ-rennes.fr/domasia) :

1. « Patient to Data » : Interopérabilité et intégration des données médicales de santé dans une perspective de réutilisation secondaire
2. « Data to Knowledge » : Extraction de connaissances à partir des données massives en santé et modèles prédictifs
3. « Knowledge to Patient » : Implémentation, évaluation en vie réelle et mesure d'impact

Dans ce cadre, la gestion de la qualité des données est un élément majeur de confiance qui intervient au sein de chaque axe : lors de la collecte, de l'analyse, de l'entraînement de modèles prédictifs ou lors du déploiement de ces méthodes en situation de soin.

Dans cette thèse, nous nous concentrerons sur deux facteurs courants qui affectent la qualité des jeux de données dans le contexte de la réutilisation des données massives : les données manquantes et les dérives temporelles. Ces facteurs ont un impact significatif sur la fiabilité, la validité et la représentativité des données utilisées pour des études secondaires [Weiskopf, 2013]. Ces deux facteurs sont inhérents à la réutilisation des données massives car, étant collectées à partir du soin clinique, celle-ci reflètent la réalité clinique. En effet, certaines informations, nécessaires pour une étude secondaire, peuvent ne pas être remontées pour l'ensemble d'une population en raison de la nature individuelle de la prise en charge. De la même manière, les pratiques cliniques évoluant (e.g., changement de systèmes de mesures, mise à jour de logiciels, nouvelles directives cliniques), les données recueillies subissent des changements au cours du temps. La mise en place de méthodologies robustes pour gérer ce type de données est un élément clé pour assurer à la fois la sécurité et la validité des outils d'aide à la décision (e.g., modèles prédictifs, analyses populationnelles).

L'objectif de cette thèse est de proposer une méthodologie afin de prendre en compte ces problèmes de qualité en les détectant, les caractérisant et les corrigeant. Elle sera construite en 4 étapes :

1. État de l'art sur les techniques de détection (e.g., dénombrement, tests statistiques, algorithme de machine learning), de caractérisation (e.g., production d'indicateurs, analyses statistiques, visualisation des données) et de gestion (e.g., imputations et transformation classiques, imputation et transformation basées sur les connaissances) [Heymans, 2022, Žliobaitė 2016].

2. Implémentation au sein d'une librairie et adaptation des mécanismes de détection et de gestion rencontrés dans la littérature pour les cas rencontrés dans l'équipe.

3. Intégration de ces mécanismes au sein de chaînes de traitement de données complètes (incluant les modèles prédictifs) pour assurer la robustesse des méthodes (conditions réelles, attaque par empoisonnement de données).

4. Application et évaluation des méthodes sur plusieurs cas d'usage du laboratoire (e.g., correction données biologiques, prédiction du cancer de la vessie à partir de données clinico-biologiques)

Au cours de cette thèse, les travaux seront valorisés par des articles de journaux et/ou de conférences internationale et un dépôt logiciel.

Lors de ce travail, il sera important de prendre en compte les aspects suivants :

Gestion des biais : Identifier et atténuer les biais potentiels présents dans les données de santé (ex : biais de sélection, biais d'information).

Sécurité et confidentialité : Les travaux étant basés sur des données de santé de vie réelles, il sera nécessaire de respecter les mesures strictes pour garantir la confidentialité et la sécurité des données des patients conformément aux réglementations en vigueur.

Collaboration : Impliquer des experts cliniques pour assurer la pertinence clinique des résultats et faciliter leur adoption dans la pratique.

Analyse comparative : Comparer les performances des différentes techniques de gestion des données manquantes et des dérives temporelles pour justifier les choix méthodologiques.

Cette thèse est financée dans le cadre du projet PEPR TraceIA, ayant pour ambition de renforcer la sécurité des SIAS. La remontée automatisée des indicateurs de qualité produits permettra à la fois de garantir la provenance des données [Stoldt, 2021] et de détecter les attaques par empoisonnement de données [Joe, 2022].

Profil recherché :

- Diplômé.e d'un Master 2 ou Ingénieur
- Maîtrise du langage Python ou R
- Bon niveau en mathématiques et statistiques
- Bonnes connaissances en Machine Learning
- Bon niveau en anglais

Contacts : Pr. Marc CUGGIA marc.cuggia@univ-rennes.fr, Dr. Sandie CABON sandie.cabon@univ-rennes.fr, and Dr. Morgane Pierre-Jean morgane.pierre-jean@univ-rennes.fr

Date de démarrage : 1er septembre 2024

Handling missing data and data drift for real-life clinical data reuse: detection, characterization and quality control

There is a growing interest in the reuse of clinical data, particularly through the exploitation of data warehouses, due to their comprehensive, multimodal, and temporal depth [Meystre, 2017]. Through its strong partnership with the Centre des Données Cliniques, responsible for the clinical data warehouse of the Rennes University Hospital (CDC, CHU Rennes), le Laboratoire du Traitement du Signal et de l'Image (LTSI, INSERM U1099) contributes to the development and implementation of methods for structuring, exploring, and exploiting this type of data (i.e., electronic medical records, biology). In particular, the DOMASIA team specializes in learning health information systems (LHIS), and its work is divided into three axes (https://ltsi.univ-rennes.fr/domasia):

1. "Patient to Data": Interoperability and integration of medical data for secondary use.
2. "Data to Knowledge": Knowledge extraction from massive health data and predictive models.
3. "Knowledge to Patient": Implementation, real-world evaluation, and impact measurement.

In this context, data quality management is a major trust element that occurs within each axis: during collection, analysis, predictive model training, or deployment of these methods in care settings.

In this thesis, we will focus on two common factors that affect the quality of datasets in the context of massive data reuse: missing data and temporal drifts. These factors have a significant impact on the reliability, validity, and representativeness of data used for secondary studies [Weiskopf, 2013]. These two factors are inherent in the reuse of massive data because, collected from clinical care, they reflect clinical reality. Indeed, some information necessary for a secondary study may not be available for the entire population due to the individual nature of care. Similarly, evolving clinical practices (e.g., changes in measurement systems, software updates, new clinical guidelines) lead to changes in the collected data over time. Robust methodologies for managing this type of data are key to ensuring both the safety and validity of decision support tools (e.g., predictive models, population analyses).

The objective of this thesis is to propose a methodology for managing these quality issues by detecting, characterizing, and correcting them. It will be built in 4 steps:

1. State of the art on detection techniques (e.g., counting, statistical tests, machine learning), characterization (e.g., production of indicators, statistical analyses), and management (e.g., classical imputations and transformations, knowledge-based imputations and transformations) [Heymans, 2022, Žliobaitė 2016].
2. Implementation within a library and adaptation of detection and management mechanisms encountered in the literature for cases encountered in the team.
3. Integration of these mechanisms within complete data processing chains (including predictive models) to ensure the robustness of methods (real conditions, data poisoning attacks).
4. Application and evaluation of methods on several use cases from the laboratory (e.g., correction of biological data, prediction of bladder cancer from clinical-biological data).

During this work, it will be important to consider the following aspects:

Bias Management: Identify and mitigate potential biases present in health data (e.g., selection bias, information bias).

Security and Confidentiality: Since the work is based on real-life health data, it will be necessary to adhere to strict measures to ensure patient data confidentiality and security in accordance with relevant regulations.

Collaboration: Engage clinical experts to ensure the clinical relevance of the results and facilitate their adoption in practice.

Comparative Analysis: Compare the performance of different techniques for handling missing data and temporal drifts to justify methodological choices.

During this thesis, the work will be valorised by journal and/or international conference articles and software deposition. This thesis is funded as part of the PEPR TraceIA project, which aims to strengthen the security of LHIS. The automated retrieval of quality indicators produced will both guarantee the data's provenance [Stoldt, 2021] and detect data poisoning attacks [Joe, 2022].

Profile sought:

- Holder of a Master's degree or Engineering degree
- Proficiency in Python or R programming language
- Strong background in mathematics and statistics
- Good knowledge of Machine Learning
- Proficient in English

Contact: Prof. Marc CUGGIA marc.cuggia@univ-rennes.fr, Dr. Sandie CABON sandie.cabon@univ-rennes.fr, and Dr. Morgane Pierre-Jean morgane.pierre-jean@univ-rennes.fr

Start date: September 1st, 2024

References:

[Meystre, 2017] Meystre, S. M., Lovis, C., Bürkle, T., Tognola, G., Budrionis, A., & Lehmann, C. U. (2017). Clinical data reuse or secondary use: current status and potential future progress. Yearbook of medical informatics, 26(01), 38-52.

[Weiskopf, 2013] Weiskopf, N. G., & Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), 144-151

[Heymans, 2022] Heymans, M. W., & Twisk, J. W. (2022). Handling missing data in clinical research. Journal of clinical epidemiology, 151, 185-188.

[Žliobaitė 2016] Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. Big data analysis: new algorithms for a new society, 91-114.

[Stoldt, 2021] Stoldt, J. P., & Weber, J. H. (2021). Provenance-based trust model for assessing data quality during clinical decision making. In 2021 IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH) (pp. 24-31). IEEE.

[Joe, 2022] Joe, B., Park, Y., Hamm, J., Shin, I., & Lee, J. (2022). Exploiting missing value patterns for a backdoor attack on machine learning models of electronic health records: Development and validation study. JMIR Medical Informatics, 10(8), e38440.