Retex d'une thèse sur les données de santé hospitalières

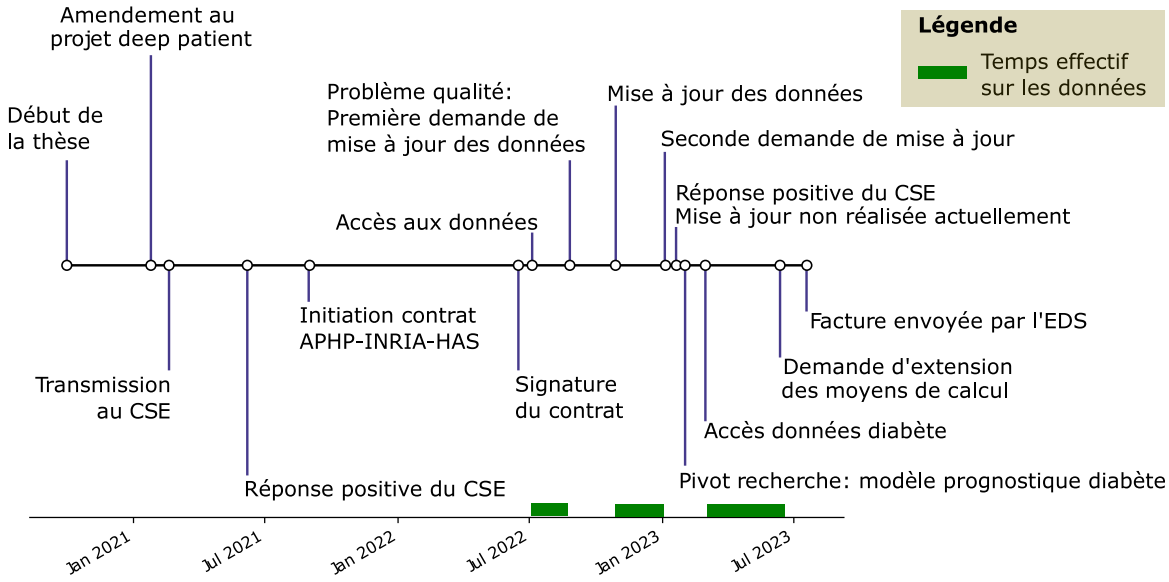
March 22, 2023

Matthieu Doutreligne
m.doutreligne@has-sante.fr
HAS, Inria

Contexte de la thèse

- Doctorant : Matthieu Doutreligne, corps de l'INSEE venant de la DREES, détaché à la HAS pour la thèse
- Encadrant Machine Learning : Gaël Varoquaux, directeur de recherche Inria, équipe Social Data, co-fondateur de la librairie open-source de machine learning scikit-learn
- Encadrant médical : Claire Morgand, Responsable du département données et études en santé ARS IDF, ancienne directrice adjointe CépiDc

Principales étapes



Calendrier détaillé

- 1er octobre 2021 : Début de la thèse.
- Octobre - décembre : Axe de recherche identifié sur la pertinence des dossiers médicaux électroniques pour la mesure d'effet causal de traitement. Possibilité de collaboration avec le professeur Antoine Neuraz sur le projet pré-existant deep patient sur l'Entrepôt de Données de Santé (EDS) de l'AP-HP.
- 25 janvier 2021 : Amendement au projet deep patient écrit et transmis à l'investigateur principal Antoine Neuraz afin d'ajouter l'axe de recherche causal et assurer un accès aux données pour le doctorant.
- 19 février 2021 : Transmission pour passage en Comité Scientifique et Ethique (CSE) par Antoine Neuraz.
- 7 Juin 2021 : Passage effectif de l'amendement en CSE.

- 30 juin 2021 : Nouvelles questions du CSE sur le projet.
- septembre 2021 : Début de la mise en oeuvre du contrat APHP-HAS-Inria.
- 16 juin 2022 : Signature du contrat de collaboration APHP-HAS-Inria.
- 30 juin 2022 : Création du compte EDS et accès délivrés.
- 5 juillet 2022 : Problèmes de connexion résolus, accès effectif aux données.
- 26 août 2022 : Première demande de mise à jour des données par Matthieu Doutreligne, suite à des incohérences constatées sur l'historique et des données manifestement manquantes en très grand nombre.
- 16 septembre 2022 : Demande de mise à jour des données par Antoine Neuraz.
- 28 octobre 2022 : Mise à jour effective des données.
- 8 novembre 2022 : Constat et documentation de la persistence d'incohérences temporelles et des données manquantes. Voir les détails sur cet url.
- Décembre - janvier : Investigation de la source d'incohérence avec la cellule qualité de l'EDS.
- 4 janvier 2023 : Confirmation par la cellule qualité d'un mélange aléatoire des dates dans l'extraction.
- 5 janvier 2023 : Second amendement pour mettre à jour les données avec des dates non mélangées transmis au CSE.
- 20 janvier 2023 : Réponse positive du CSE. Cette mise à jour n'a pourtant pas été effectuée actuellement: Pas de mail reçu de la DSI et pas de base retrouvée sur l'espace projet.
- Février 2023 : Changement de question de recherche. Projet de modèle prognostique de risque non causal sur les données de l'APHP. Utilisation d'une cohorte de diabétologie (projet codia) et d'un échantillon aléatoire de 200,000 patients.
- 1er mars 2023 : Demande d'ajout de Matthieu Doutreligne à l'espace projet codia, ajout pris en compte immédiatemment.
- 31 mai 2023 : Présentation de l'équipe data de l'EDS sur l'appariement au statut vital INSEE. Manque manifeste de certains statuts vitaux dans les données. Pas d'estimation disponible à ce jour.
- 12 juin 2023 : Demande d'extension de ressources de calcul au delà d'un espace de travail classique (16GB RAM = ordinateur de bureau) afin d'évaluer des modèles de deep learning et d'évaluer plus rigoureusement les modèles de machine learning pour le risque cardio-vasculaire. Demande d'un GPU, 80GB de mémoire RAM, 40 CPUs.
- 19 juillet 2023 : Facture envoyée par la DSI.
- 4 août 2023 : Demande de contact pour effectuer le financement nécessaire de la part de l'investigateur principal du projet diabète (Louis Potier).