Table des matières

[Introduction 3](#_Toc154043983)

[Premiere partie 3](#_Toc154043984)

[Compréhension besoin client 3](#_Toc154043985)

[Etat de l’art 3](#_Toc154043986)

[Elements de conception technique 3](#_Toc154043987)

[Choix techniques liés au projet 3](#_Toc154043988)

[Réponse finale apporté (ce qui a été réalisé) 3](#_Toc154043989)

[Seconde partie (mise en œuvre) 3](#_Toc154043990)

[L’organisation technique et l’environnement de développement tout au long de la production 3](#_Toc154043991)

[La gestion de projet 3](#_Toc154043992)

[Retours d’expérience sur les outils, techniques et compétences à l'oeuvre tout au long du projet 3](#_Toc154043993)

[Troisième partie bilan de projet et améliorations envisageables 3](#_Toc154043994)

[Conclusion 3](#_Toc154043995)

## Introduction

Mon alternance s’est déroulée au CHU de Caen, en médecine nucléaire. La médecine nucléaire permet d’interpréter des images obtenues à partir d’injection de traceurs radioactifs dans le corps.

Le but de ce stage était d’utiliser l’IA pour structurer des comptes rendus médicaux en jeux de données analysables. Ces comptes rendus proviennent d’examens de scintigraphies myocardiques de perfusion (fonctionnement artères du cœur), et de coronarographie (anatomie artères du cœur). L’analyse des comptes rendus permettra de distinguer l’importance de différentes caractéristiques pour la détermination d’une maladie des artères coronaires.

Pour les coronarographies, les données d’entraînement n’ont pas été générées à la main mais avec un programme, cela soulève plusieurs questions quant à l’utilisation de l’IA.

Des modèles d’IA « Scikit-learn » ont été utilisés, en combinaison avec des expressions régulières pour extraire les différentes caractéristiques des scintigraphies. Ces dernières ont permies d’entraîner un modèle de prédiction du risque de maladie.

# Partie 1

## Compréhension besoin client

Notre but est de développer un programme d’IA performant, scalable sur plusieurs années, d’extraction de caractéristiques sur des comptes rendus de scintigraphies. Ceci ayant pour but de contribuer à un projet de recherche de « comparaison des performances diagnostiques de la scintigraphie avec la coronarographie ».

Un jeu d’entraînement fait avec un programme d’expressions régulières est réalisé pour 2019 sur les scintigraphies, et pour les scintigraphies le jeu est réalisé à la main sur cette même année.

### Enjeux réglementaires

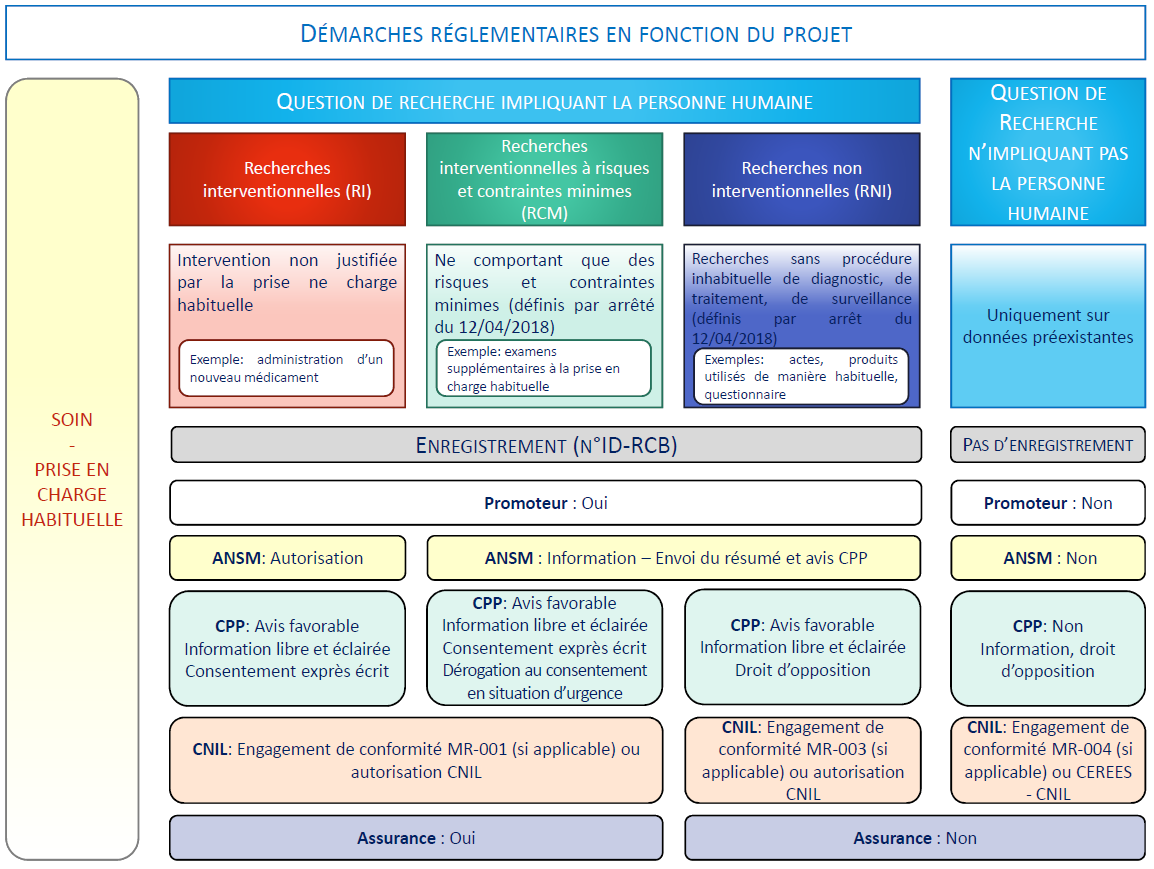
Pour réaliser un travail de recherche en santé, il faut obtenir une autorisation d’un comité d’éthique : le CLERS (Comité Local d’Ethique pour la Recherche en Santé).

Figure : Schéma des réglementation pour les travaux de recherche impliquant les données des patients

Comme on le voit sur la figure 1, dans le cas de recherches portant uniquement sur des données préexistantes, pas besoin de demander une autorisation au CPP (Comité de Protection des Personnes). Il faut être conforme à la MR-004, qui est une liste de règles établies par la CNIL (Commission nationale de l'informatique et des libertés).

Obligations MR-004 :

* Données collectées strictement nécessaires aux objectifs de la recherche.
* Transfert hors UE doit-être strictement nécessaire. Le cas échéant, les données doivent êtres pseudonymisées au minimum.
* Le projet doit-être enregistré sur le « Health Data Hub » (plateforme d’intérêt public de données de santé).
* Données des patients pouvant être conservées jusqu’à 2 ans après la dernière publication des résultats.
* Seuls les professionnels et leurs collaborateurs intervenant dans la recherche, dans un lieu de recherche, peuvent conserver le lien entre l’identité codée des patients et leurs noms et prénoms.
* Une information générale et individuelle concernant la réutilisation des données doit être effectuée vis-à-vis des patients (possibilité de demander une dérogation en cas d’effort disproportionné).
* Anonymiser au maximum les données lors de la publication.

### Utilisateurs projet

L’objectif côté utilisateurs, serait de développer une application qui permette aux médecins chercheurs de structurer les données en sortie de la base de données.

Il est aussi envisagé de développer un modèle de prédiction de risque de maladie en fonction du compte rendu, pour assister les médecins dans leurs diagnostics.

### Schéma fonctionnel utilisé pour le projet

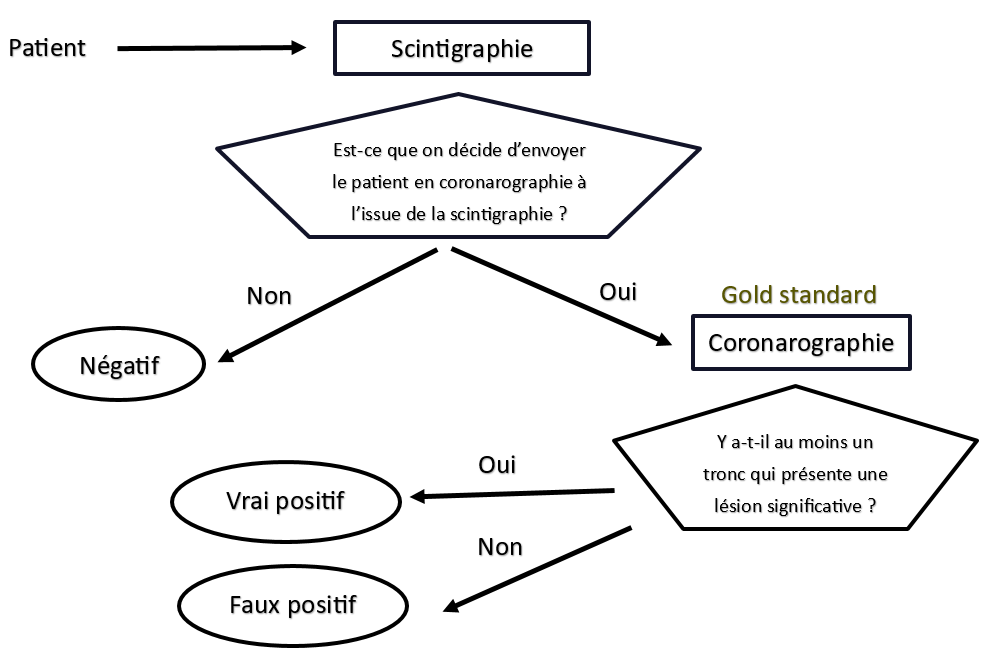


Figure : Schéma explicatif de la comparaison des performances diagnostiques de la scintigraphie myocardique de perfusion

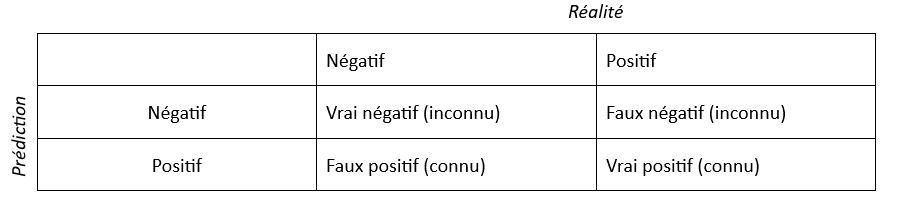
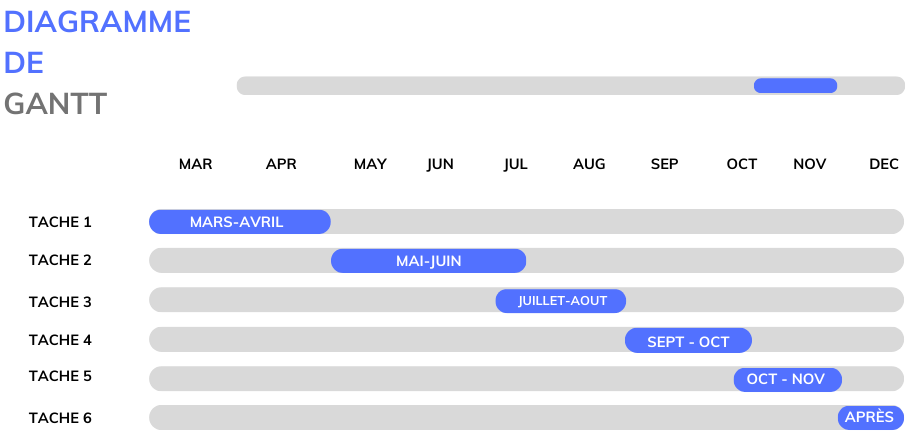


Figure 4 : Matrice de confusion

Comme nous le voyons sur la figure 3, la coronarographie est le « gold standard » : c’est par convention l’examen qui vérifie la présence de la maladie chez le patient.

Les patients prédits négatifs (à la scintigraphie) ne sont jamais envoyés en coronarographie (examen invasif). On ne peut pas savoir si on a fait une erreur de prédiction dans ce cas-ci.

### Macro-planning initial



Tache 1 : Recherche sur les réglementations des projets de recherche

Tache 2 : Veille technique, documentation sur les métriques de performance utilisées.

Tache 3 : Développement d’un code permettant de comparer différents modèles de la librairie « Scikit-learn ».

Tache 4 : Organisation du code, développement des fonctions finales de prédiction.

Tache 5 : Mise en production des modèles avec une application qui extrait les informations sur les comptes rendus de scintigraphies.

Tache 6 : Mise au propre du code, amélioration des modèles.

Figure : Diagramme de Gantt du projet

### Etat de l’art de l’extraction de données sur du texte

L’extraction de données sur du texte se fait majoritairement avec des expressions régulières dans les articles scientifiques. C’est une méthode relativement simple à déployer, cependant, elle se limite rapidement avec la diversité du texte et la complexité des éléments à extraire. Quant aux algorithmes de deep learning, ils nécessitent d’annoter une grande quantité de données.

Jusqu’à maintenant, il n’y a pas vraiment de méthodologie développée pour solutionner ce type de problèmes. On remarque une absence de tâches communes qui permettraient de comparer différentes méthodes de développement, ainsi qu’une absence de métriques communes.

De nouvelles pipelines basées sur des modèles pré-entraînés comme GPT-3 pourraient résoudre en partie le problème du manque de données d’entrainement (1).

## Elements de conception technique

## Choix techniques liés au projet

## Réponse finale apporté (ce qui a été réalisé)

# Partie 2 (mise en œuvre)

## L’organisation technique et l’environnement de développement tout au long de la production

## La gestion de projet

## Retours d’expérience sur les outils, techniques et compétences à l'oeuvre tout au long du projet

# Partie 3 bilan de projet et améliorations envisageables

# Conclusion

# Références

1. **Process Extraction from Text: Benchmarking the State of the Art and Paving the Way for Future Challenges.** [En ligne] https://arxiv.org/pdf/2110.03754.pdf.