

**ComPaRe**

**Utilisation de l'IA générative pour classer les commentaires  
en texte libre à la fin des questionnaires en ligne**

Responsable principal du projet

Nom	JEANTON-SEYBEL
Prénom	Capucine
Affiliation(s)	Université Paris Cité

Comité scientifique

Nom	Grade/Titre	Affiliation(s)
Viet-Thi TRAN	PU-PH	AP-HP
Céline BEJI	CCA	AP-HP

## 1. Contexte spécifique

En général, dans les questionnaires, une dernière question en texte libre sert à capturer les commentaires de participants. Ces commentaires peuvent avoir trait au questionnaire (idée d'amélioration, problème rencontré, etc.), être un avis sur la recherche, ou exprimer un problème non directement lié à la recherche.

Les commentaires ne sont habituellement pas lus en temps réel, ce qui limite l'identification en temps réel de problèmes dans le questionnaire.

## 2. ComPaRe : Communauté de Patients pour la Recherche

Le présent projet sera niché dans la *Communauté de Patients pour la Recherche (ComPaRe)* ([www.compare.aphp.fr](http://www.compare.aphp.fr)). ComPaRe est un projet porté par l'Assistance Publique – Hôpitaux de Paris et l'Université de Paris visant à accélérer la recherche sur les maladies chroniques. ComPaRe s'appuie sur une e-cohorte de patients adultes, souffrant d'une maladie chronique. L'objectif est de recruter 100 000 patients en 10 ans et de les suivre pendant 10 ans.

Dans ComPaRe, les participants sont suivis via des questionnaires en ligne permettant de collecter des critères de jugement rapportés par les patients (en anglais « Patient Reported Outcome Measures » [PROMS]) ou des expériences rapportées par les patients (Patient Reported Experience Measures [PREMS]).

Les données rapportées directement par les patients ont pour objectif d'être enrichies par d'autres sources de données telles que :

- Données provenant de bases médico-administratives (e.g. SNDS)
- Données hospitalières (entrepôts de données hospitaliers)
- Données provenant des outils connectés des patients (mesure de l'activité physique, du sommeil, glycémie)

### 2.1. ComPaRe : une plateforme de recherche

ComPaRe est une infrastructure de recherche permettant à toute équipe de recherche publique (hospitalière, universitaire ou institutionnelle) d'utiliser des données collectées dans ComPaRe ou de demander la collecte de nouvelles données (e.g., en ajoutant de nouveaux questionnaires en ligne) pour répondre à des questions de recherche sur les maladies chroniques.

Le modèle permet d'accélérer la recherche dans les maladies chroniques en:

- *Simplifiant le recrutement des participants dans la recherche clinique.* Le recrutement des patients est réalisé une seule fois. Ensuite, les patients peuvent participer aux études qui les concernent et qui les intéressent.
- *Simplifier la réalisation de recherches nécessitant l'accès à des données externes (entrepôts de données hospitaliers, SNIIRAM, objets connectés etc.).*
- *Améliorant le partage de données.* Toutes les informations recueillies pourront être réutilisées. Ceci allègera le fardeau des participants qui ne devront pas répondre plusieurs fois à la même question, posée par des chercheurs différents ; limitant un « gâchis de la recherche »
- *Réduisant les coûts liés à la recherche avec une plateforme commune.* La plateforme ne sera construite qu'une seule fois et servira de multiples fois. Toutes les améliorations bénéficieront à l'ensemble des recherches qui y sont (et seront) menées.

### 3. Objectifs

Evaluer la performance de l'utilisation d'un modèle de langage (large language model) pour classer automatiquement les commentaires en texte libre à la fin de questionnaires en ligne.

## 4. Design

Cette étude comporte trois étapes :

- 1) Développement d'un « prompt » pour expliciter la tâche au modèle de langage. Ce prompt est construit à partir de 100 commentaires en textes libres tirés au sort parmi l'ensemble des commentaires entrés dans ComPaRe, classés manuellement en 6 catégories, et d'une explication expliquant la catégorisation. Les 6 catégories sont les suivantes :
  - **Complément** : complément d'information sur la maladie, le traitement, le parcours de soin ou tout autre item du questionnaire, permettant d'affiner ou de contextualiser les réponses du participant
  - **Avis** : avis sur l'étude ou le questionnaire de l'étude
  - **Amélioration** : questionnements et problématiques en lien avec le questionnaire ou l'étude, pouvant constituer un axe d'amélioration
  - **Personnel** : opinion et ressentis personnels non graves
  - **Problème** : problèmes et ressentis personnels graves constituant une alerte
  - **Inclassable** : pour les commentaires non compréhensibles et/ou qui ne rentrent dans aucune des catégories précédentes
- 2) Amélioration du « prompt » via des itérations successives testant différentes formulations et comparaison entre la sortie du modèle de langage et une classification manuelle des commentaires dans les 6 catégories précédemment citées. L'échantillon d'amélioration consistera en 1000 commentaires, tirés au sort parmi les commentaires entrés dans 10 questionnaires ComPaRe (200 commentaires tirés au sort par questionnaire)
- 3) Validation du « prompt » dans un échantillon de validation de 1000 commentaires, tirés au sort parmi les commentaires entrés dans 10 questionnaires ComPaRe (200 commentaires tirés au sort par questionnaire)

### 4.1. Patients participants

Les critères d'éligibilité sont :

Tous les patients sont éligibles à l'étude. Les participants reçoivent un mail d'information que leurs commentaires en texte libre sont utilisés dans l'étude et peuvent choisir de ne pas participer.

### 4.2. Modalités de recrutement

Les patients seront recrutés de deux manières différentes et complémentaires :

## 5. Données recueillies

### 5.1. Recueil spécifique dans le cadre du projet

Donnée collectée ou instrument utilisé	Moment de la collecte
Commentaires en texte libre entrés par des participants à la fin des questionnaires ComPaRe	

### 5.2. Données ComPaRe utilisées

En plus des données spécifiques au projet, des données recueillies pour tous les patients, dans le cadre du suivi général de ComPaRe seront utilisées :

X Données démographiques (Age, Sexe, statut marital, lieu d'habitat)

X Niveau d'étude

Données socioprofessionnelle

Origine géographique des parents du participant

Statut tabagique

Consommation d'alcool

Données anthropométriques

Handicaps

Santé des femmes (nombre de grossesses/parité, contraception, ménopause, etc.)

X Maladies chroniques

Traitements pris par le patient

Fardeau du traitement (TBQ)

Observance aux traitements médicamenteux

Précarité (EPICES, mesure écologique de précarité, auto évaluation de situation financière)

- Qualité de vie (EQ-5D)
- Sévérité des symptômes (MYMOP2)
- Qualité de sommeil (PSQI)
- Activité physique (IPAQ)
- Symptômes dépressifs (PHQ-9) / Symptômes d'anxiété (GAD-7)

## 6. Analyses statistiques

L'analyse des données de la cohorte sera réalisée par

CEC Hôtel Dieu

### 6.1. Variables à but descriptif

Caractéristiques des participants

Nombre de commentaires dans ComPaRe

Longueur des commentaires

### 6.2. Critères d'évaluation

- 1) Concordance entre la classification manuelle et la classification automatique exprimée par un coefficient Kappa.
- 2) Nombre de commentaires classés dans chaque catégorie

## 7. Gestion du projet

Cette cohorte spécifique sera pilotée par un **Comité scientifique spécifique (Comité de pilotage)** composé impérativement :

- d'un chercheur responsable du projet
- de médecins et/ou chercheurs associés
- **d'au moins un patient**

Le rôle de ce comité est de :

- Définir le projet et veiller à sa rigueur scientifique,
- Valider les outils de recueil de données,
- Proposer les éléments d'animation scientifique
- Superviser l'analyse des données
- Rapporter de manière transparente l'ensemble des résultats du projet.

## **8. Aspects éthiques et réglementaires**

Le projet ComPaRe a reçu un avis favorable par le Comité de protection des personnes – Ile de France 1 (IRB : 0008367).

Le projet ComPaRe a reçu un avis favorable du Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé (CCTIRS) sous le numéro 16-395bis.

Le projet a reçu un avis favorable de la CNIL, sous le numéro de dossier n°916397 (DR-2016-459) en date du 25/11/2016.

### **8.1. Gestionnaire de la recherche**

L'Assistance Publique – Hôpitaux de Paris sera le gestionnaire de la recherche. Le Département de la Recherche Clinique et du Développement (DRCD) en sera son représentant.

L'Assistance Publique – Hôpitaux de Paris se réserve le droit d'interrompre la recherche à tout moment pour des raisons administratives.

Le responsable scientifique, également investigateur coordonnateur, s'engage à fournir au gestionnaire les informations relatives aux inclusions des participants dans la recherche.

Toute modification du protocole de recherche devra être soumise au gestionnaire.

### **8.2. Encadrement réglementaire de la recherche**

Conformément à la loi n°78-17 du 6 janvier 1978 modifiée par la loi n° 2004-801 du 6 août 2004 relative à la protection des personnes physiques à l'égard des traitements de données à caractère personnel et son décret d'application n° 2005-1309 du 20 oct. 2005, ComPaRe a obtenu l'autorisation du CCTIRS N° 16.395 et de la CNIL N° 916397.

### **8.3. Qualification des intervenants**

Les membres de l'équipe porteuses du présent projet sont habilités à mener les tâches qui leur incombent, c'est-à-dire qu'ils disposent des compétences appropriées et documentées.

### **8.4. Protection des données**

La Direction des Services Informatiques (DSI) de l'AP-HP assure le développement, l'hébergement, le maintien en conditions opérationnelles et l'exploitation de la plateforme

ainsi que de l'application ComPaRe. Elle met en œuvre tous ses moyens techniques afin de garantir une sécurisation adéquate des données.

### **8.5. Saisie des données**

Les patients saisiront directement leurs questionnaires dans un espace personnel sécurisé, protégé par un identifiant et un mot de passe, sur un serveur accessible, sécurisé par un certificat, et placé derrière un « firewall ».

### **8.6. Confidentialité et sécurisation des données**

Toute information concernant les participants sera confidentielle et l'identité du participant sera strictement préservée.

Toutes les données à caractère personnel recueillies auprès des patients seront sécurisées et encryptées en cas d'échanges de données ; les communications entre les participants ou les chercheurs et la plateforme seront cryptés par SSL (https).

- Un nombre restreint de personnes habilitées aura accès aux données directement nominatives, à des fins exclusivement logistiques et scientifiques. Elles prendront toutes les précautions nécessaires en vue d'assurer la confidentialité de ces données.
- Conservation des données
- Toutes les données seront conservées sur deux serveurs sécurisés, et traitées de façon strictement confidentielle.
- Les données complétées sur le site seront stockées sur des serveurs agréés pour l'hébergement de données médicales.
- Les deux serveurs seront garantis d'avoir un haut niveau de sécurité.

### **8.7. Accès aux données**

Seuls les chercheurs habilités et les intervenants en charge de la coordination de la recherche auront accès aux données, et cela uniquement pendant la durée de celle-ci.

### **8.8. Destruction des données**

Les données des participants seront conservées pendant 10 ans après la fin de la recherche sur les personnes (i.e. fin du suivi du dernier participant recruté). Elles seront ensuite détruites, conformément à la loi Informatique, Fichiers et Libertés.

### **8.9. Information et consentement des participants**

### **8.9.1. Information**

Lors de leur inscription en ligne, les participants devront consulter la note d'information expliquant le cadre et les objectifs de ComPaRe. Ils seront informés du respect de la stricte confidentialité de leurs données ainsi que de leurs droits relatifs à la loi Informatique, Fichiers et Libertés. Il leur sera notamment notifié qu'à tout moment ils pourront exercer leur droit d'accès et de rectification des données, conformément à la loi 78-17 du 06 janvier 1978 relative à l'Informatique, aux Fichiers et aux Libertés, modifiée par la loi n°94-548 du 1er juillet 1994, relative au traitement des données nominatives ayant pour fin la recherche dans le domaine de la santé.

### **8.9.2. Consentement**

Les patients déjà inscrits dans ComPaRe auront préalablement signés le formulaire de consentement général et pourront participer à ce projet.

**Le consentement général couvre la participation à cette cohorte spécifique.**

Un consentement spécifique sera demandé pour donner accès à son dossier médical hospitalier (compte-rendu d'hospitalisation, compte-rendu de consultation, imagerie médicale, résultats d'examen, ...);

## **8.10. Publication et communications**

Les résultats des travaux réalisés à partir des données obtenues sur ComPaRe doivent être rendus publics. Dans le cas où ces travaux donnent lieu à une publication scientifique, un résumé « grand public » sera publié sur le site de ComPaRe

**La mention que la recherche a été réalisée dans le cadre de la cohorte ComPaRe doit être faite dans toute les publications ou communications. Le mot « ComPaRe » doit figurer autant que possible dans le titre de toute publication.**

Le responsable d'un projet de recherche est tenu de communiquer à l'équipe ComPaRe une copie des manuscrits avant soumission pour publication.

Les signataires des papiers utilisant les données ComPaRe doivent être déterminés en accord avec les équipes impliquées l'organisation de la recherche et le recueil des données.

## 9. Bibliographie