## Motivations pour la création du jeu de données

**Pourquoi le jeu de données a-t-il été initialement créé ?**

Ces dossiers médicaux figurent des notes fictives de médecins généralistes dans leur activité de premier recours au cours de consultations successives. Ces informations constituent l’essentiel de la traçabilité constitué par le professionnel de santé dans son dossier médical.

Ce jeu de données a été créé pour le défi 1A du Défi iDoc Santé, qui vise à identifier automatiquement les recommandations de bonnes pratiques pertinentes dans les situations cliniques décrites dans ces dossiers médicaux.

**Pour quelles autres tâches le jeu de données pourrait-il être utilisé ?**

**Des dossiers médicaux réels pourraient être utilisés pour effectuer des recherches, études ou évaluations sur la qualité des soins et les pratiques des professionnels mais ce n’est pas le cas pour les données présentées ici qui sont fictives.**

**Quelles sont les utilisations trompeuses du jeu de données ?** Le format de ces données n’est proposé que pour le défi 1A du Défi iDoc Santé. Il est conforme à aucune norme d’interopérabilité, la HAS et ses partenaires dans ce concours ne promeuvent en aucun cas d’autre usage de ce format.

**Qui a financé ou soutenu la création du jeu de données ?** La HAS est à l’origine de la rédaction de ces dossiers médicaux fictifs. Ils s’inspirent de véritables dossiers médicaux extraits de son logiciel de gestion de cabinet par Philippe Szidon, médecin généraliste à Paris. De ces dossiers médicaux, les épisodes d’histoires pathologiques emblématiques ont été extraits, les dates décalées aléatoirement et les détails changés. Ces « histoires pathologiques » ont ensuite été regroupées de façon aléatoire pour fabriques des histoires possibles de patients de médecine générale.  
Des balises ont ensuite été insérées pour produire un document json. Ces traitements ont été réalisés par Pierre Liot, HAS.

## Composition du jeu de données

**Que contient le jeu de données principalement ?** Les principales informations de ces données sont des informations médicales et des prescriptions.

**Dispose-t-on d’un schéma décrivant les variables du jeu de données ?**

Oui. Le schéma json du document est fourni. Il ne correspond à aucun standard.

**Que contient chaque champ du jeu de données ?**

Le jeu de données est une liste de patients, qui comprend chacun les informations de sex, date de naissance, et une liste de prescription.

Les balises json sont explicites et en français :

* **Sex** : sexe
* **DDN** : date de naissance
* **Consultations** : liste de consultations contenu d’une consultation
* **Date\_consultation** : date de la consultation
* **Resultat\_consultation** : diagnostics et problèmes principaux justifiant la consultation. A noter que ces contenus comportent de nombreux libellés de CISP2
* **Biometrie** : constantes cliniques
* **Biologie** : résultats d’examen biologique
* **Accident\_travail** : contenu du formulaire « Accident du travail »
* **Text** : autres notes prises lors de la consultation

**Est-ce que le contenu du jeu de données dépend de ressources externes ? Non.**

**De quelles garanties dispose-t-on concernant la pérennité de ces ressources ? Le contenu médical ne dépend pas de sources externes.**

## Processus de collecte des données

**Comment les données ont été collectées (avec des capteurs, manuellement par des outils informatiques…) ?** Les dossiers médicaux d’origine, servant d’inspiration, ont été saisies manuellement dans le logiciel de gestion de cabinet d’un médecin généraliste.

**Qui a assuré le processus de collecte de données (des agents, des bénévoles, des étudiants…) ?** Les dossiers initiaux ont été saisies par Philippe Szidon, médecin généraliste à Paris. Les dossiers inventés ont été saisis par Pierre Liot (HAS, neurologue) dans des outils bureautiques en s’inspirant des dossiers réels affichés via un script spécifique.

**Quelle a été la période de collecte des données ?** De janvier 1998 à mai 2022 pour les dossiers initiaux.

**Les données ont-elles été collectées directement ou inférées à partir d’autres données ?** Les données ont été créées en s’inspirant de données collectées directement.

**Les données ont-elles été collectées sur un échantillon ? Selon quelles méthodes ?** Ces données représentent un échantillon très restreint, défini heuristiquement par un professionnel de santé pour illustrer une variété de situations cliniques.

**Quelles sont les erreurs connues, les limites, les sources de bruit ou de redondances associées à ces données ?** Ces données ne sont ni exhaustives ni représentatives. Elles ne sont qu’une illustration de situations cliniques emblématiques et fréquentes telles qu’on peut les trouver dans les dossiers médicaux d’un médecin de premier recours. Elles ne sont pas publiées selon un format standard.

## Pré-traitement des données

**Comment les données ont-elles nettoyées ou préparées ?** Ces dossiers médicaux ne comportent pas de nom, aucune date (naissance, consultation, arrêt ou accident de travail, …) n’est réelle. Aucune des associations pathologiques des fiches initiales n’a été conservée dès lors qu’elle n’avait pas de relation médicale connue. Ces dossiers médicaux publiés sont donc inventés et sans relation avec de réels patients.

**Les données « brutes » ont-elles été conservées ? Sont-elles diffusées ?** Les données sources, les dossiers médicaux ayant servi d’inspiration, ne sont pas diffusées.

**L’outil de prétraitement des données est-il disponible ?** Non.

## Diffusion du jeu de données

**Les données sont-elles diffusées en ligne ? Selon quelles modalités (sur un portail open data, un site web, une API…) ?** Oui, sur cette page [data.gouv.fr](http://data.gouv.fr).

**Selon quelle licence les données sont-elles diffusées ?** [License Ouverte version 2.0](https://www.etalab.gouv.fr/wp-content/uploads/2017/04/ETALAB-Licence-Ouverte-v2.0.pdf)

**Des redevances ou des restrictions sont-elles appliquées dans l’accès aux données ?** Non.

## Maintenance du jeu de données

**Qui assure la maintenance du jeu de données ? Comment peut-on contacter cette personne ? Quel est le service responsable du jeu de données ?** Ce jeu de données n’est pas maintenu. La publication est réalisée par la HAS.

**Est-ce que les rôles sont distincts entre la production des données, leur éditorialisation et leur diffusion ?** Non. La production de ces dossiers médicaux fictifs a été réalisée par la mission Data de la HAS qui en assure la diffusion.

**Le jeu de données sera-t-il mis à jour ? Si oui, à quelle fréquence ?** Non

**Si les données deviennent obsolètes, comment cette information sera-t-elle communiquée ?** Sur cette page.

**Est-il possible de contribuer à l’amélioration des données ? Selon quelles modalités ?** La mission data de la HAS est à l’écoute de toute proposition constructive.

## Considérations légales et éthiques

**Si le jeu de données concerne des individus, ont-ils exprimé leur consentement de manière claire ?** Les dossiers médicaux fictifs ne concernent pas des individus.

**Le jeu de données peut-il exposer de manière directe ou indirecte des individus ?** Non. Les dossiers publiés sont fictifs.

**Ces données sont-elles conformes au RGPD ?** Oui

**Les données peuvent-elles avantager ou désavantager des groupes sociaux ?** Non.

**Le jeu de données contient-il des informations pouvant être considérées comme inappropriées ou offensantes ?** Non.