Epidémiologie et Biostatistique sur un Entrepôt de Données hospitalier

02 juin 2025
Présentation d'introduction

Luca THIEBAUD & Adam REMAKI
adam.remaki@centralesupelec.fr
luca.thiebaud@centralesupelec.fr

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Plan

- **L'AP-HP en chiffres**
- **Présentation de l'EDS et de ses enjeux**
  - L'EDS en trois mots
  - Techniquement, qu'est-ce que l'EDS?
  - A quoi peuvent servir les données de l'EDS?
  - Comment les acteurs de l'EDS interagissent-ils?
- **Présentation de l'Enseignement d'Intégration**
  - Objectifs
  - Organisation de la semaine
  - Contexte scientifique
- **Questions**

1.

L’AP-HP en chiffres

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

L'AP-HP en chiffres

CHU d'Ile de France

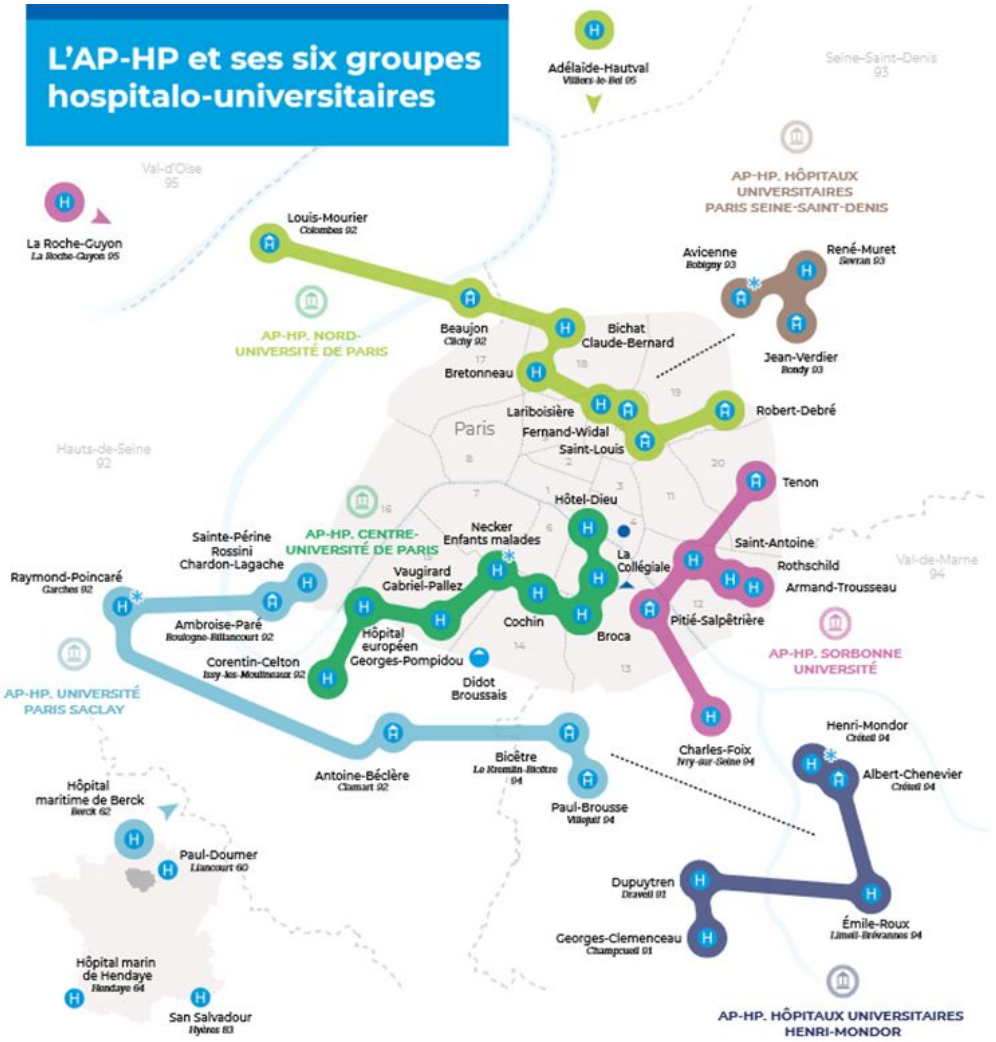
- Directeur général: M. Nicolas Revel
- Présidente du conseil de surveillance : Mme. Anne Hidalgo

39 hôpitaux
20 098 lits
54 blocs chirurgicaux (315 salles d'opération)

100 000 professionnels

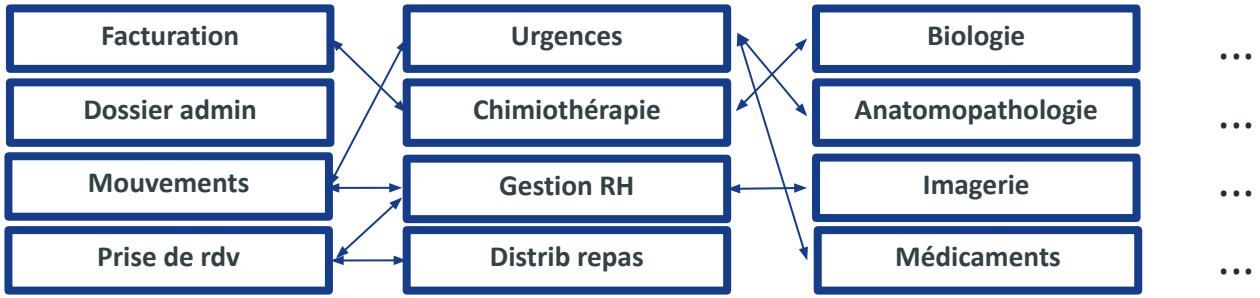
- 13 220 médecins
- 2000 bénévoles auprès des patients et des familles

9,6Md€ de budget



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

DSI de L'AP-HP



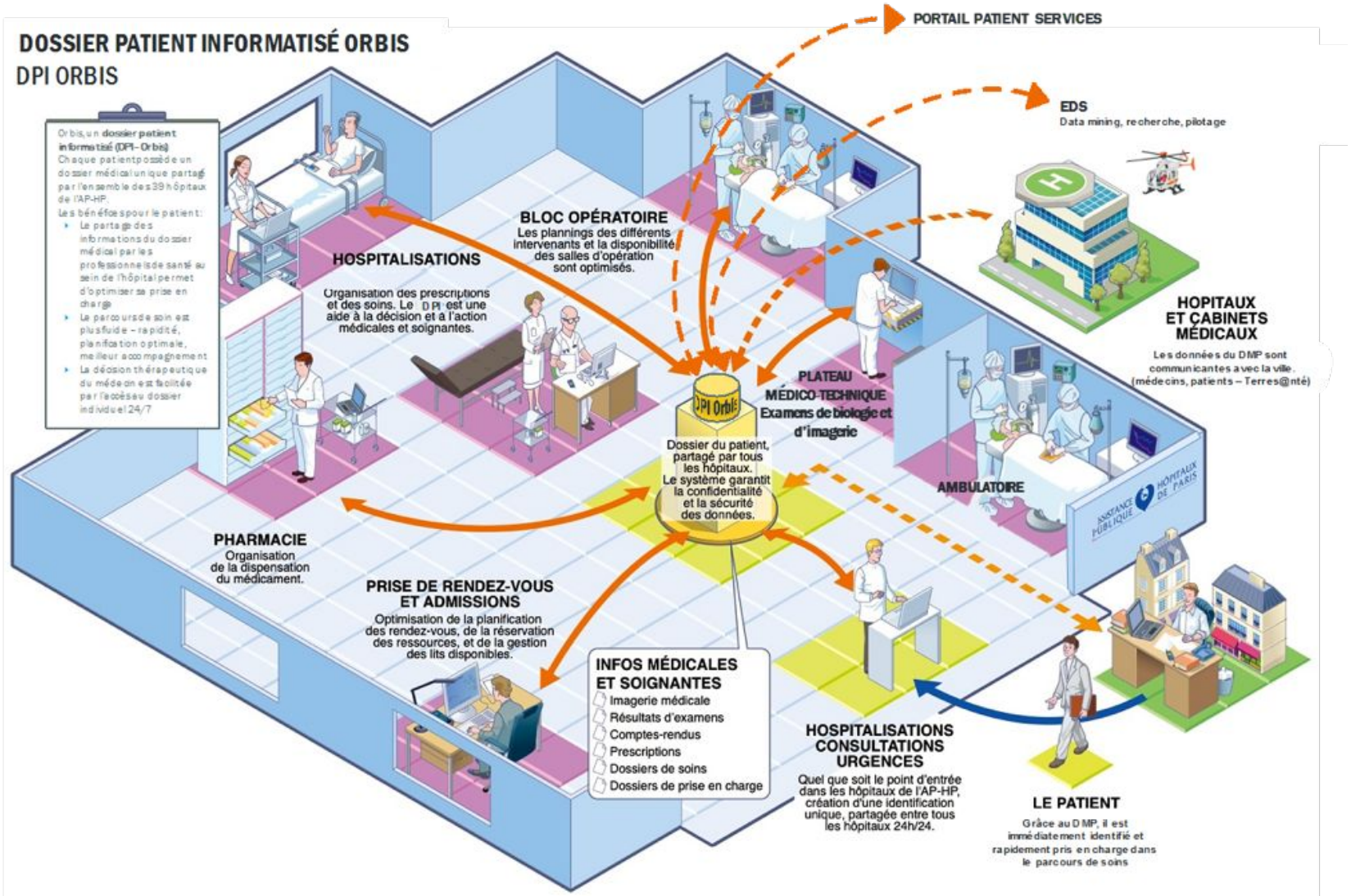
- **Direction des Systèmes d'Information**
  - 800 professionnels
  - 71M€ budget
  - 800 applications répertoriées

- **Adaptation continuelle du système d'information aux besoins métiers**
  - Définitions de « Dossiers de spécialités »
  - Intégration en cours d'algorithmes avancés (ML, ..)

- **Depuis 2012, déploiement d'un Dossier Patient Informatisé commun aux 39 hôpitaux**

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Le DPI



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

2.

Présentation de l’EDS

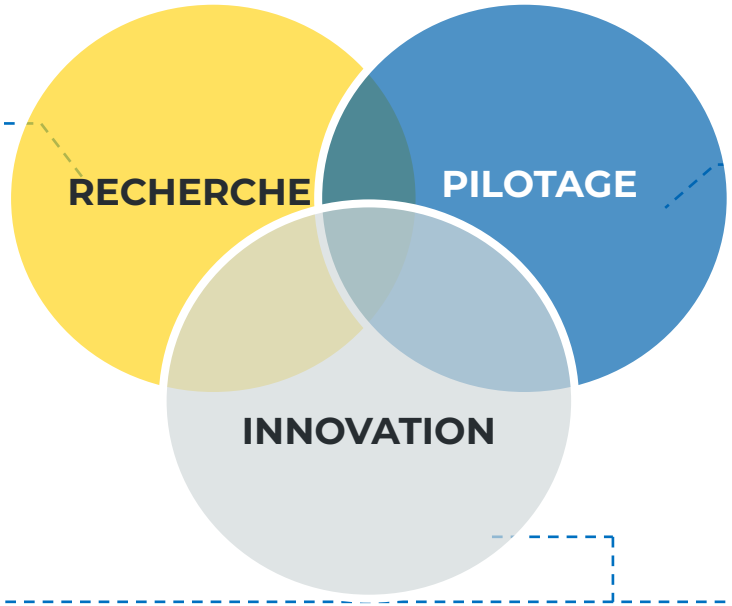
ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

L'EDS en trois mots

Soutenir la recherche

> Développer les **recherches basées sur la réutilisation des données de santé**

> Développer des **technologies d'optimisation de la recherche clinique** (repérage automatique des patients, transfert de données)



Faciliter le pilotage de l'activité hospitalière et l'organisation des soins

> Démarche *Value Based Health Care*

Soutenir l'innovation

> **Évaluer et valider des technologies/algorithmes** d'**aide à la décision médicale**

> Construire et fiabiliser des **jeux de données.**

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Les données de l'EDS

Données administratives

19M de patients

30M de dossiers administratifs

Données cliniques non structurées

175M de CR textuels

40M d'examens d'imagerie

Données médico-économiques (PMSI)

30M d'actes CCAM

30M de diagnostics CIM-10

Données cliniques structurées

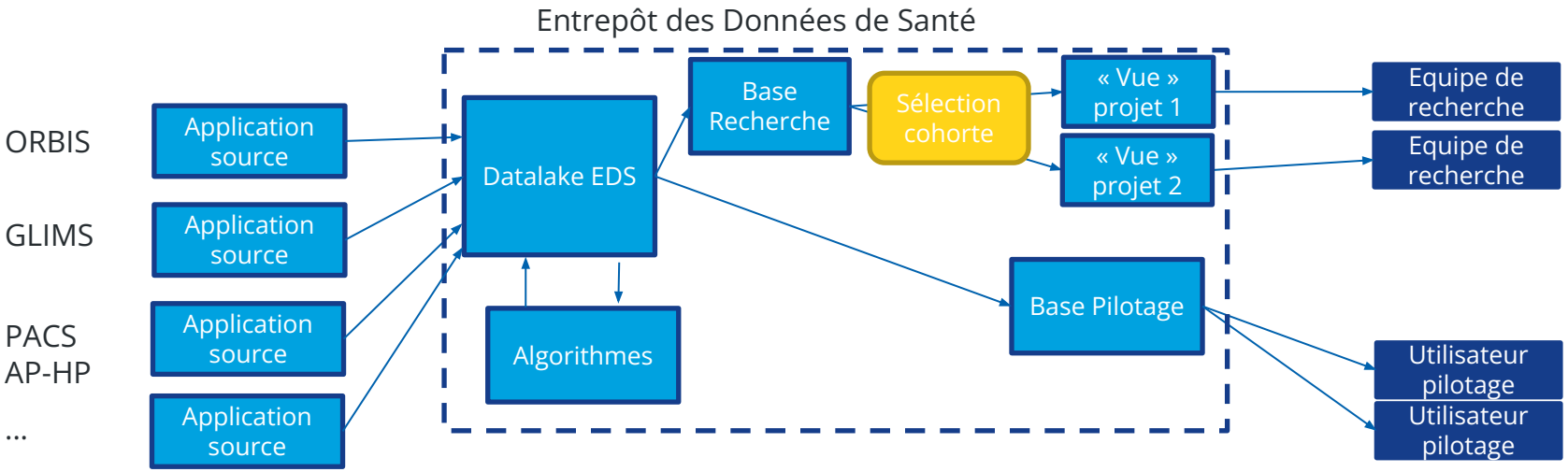
1 200M d'analyses biologiques

100M de médicaments administrés

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Techniquement, qu'est-ce que l'EDS ?

- **L'EDS** est un **agrégateur** de données de vie réelle et une **plateforme** de mise à disposition de ces données
  - *« On désigne sous le terme 'données de vie réelle', ou 'données de vraie vie', des données qui sont sans intervention sur les modalités usuelles de prise en charge des malades et ne sont pas collectées dans un cadre expérimental (le cadre notamment des essais randomisés contrôlés, ECR), mais qui sont générées à l'occasion des soins réalisés en routine pour un patient, et qui reflètent donc a priori la pratique courante. » Rapport Bégaud, Polton, von Lennep 2017*



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Techniquement, qu'est-ce que l'EDS ?

Forces et faiblesses des données de l'EDS

Forces:

- ***Données variées***
- ***Données massives***
- ***Données mises à jour régulièrement***
- ***Données représentatives du contexte de soin***

Faiblesses:

- *Certaines **données manquantes** (ex: déploiement d'ORBIS inhomogène)*
- *Certaines **données erronées** en raison du circuit de données (ex: documents datés de l'an 1…)*
- *Certaines **données biaisées** (ex: grève du codage, diagnostics PMSI sous-codés)*
- *De nombreuses données présentes sous forme **non-structurée** (ex: CRH)*
- *Une **temporalité complexe** (documents mis à jour quotidiennement, PMSI mensuellement avec correction, etc.)*

A quoi peuvent servir les données de l'EDS?

- **Etude observationnelle rétrospective**

ORIGINAL ARTICLE

WILEY

Care pathway for patients hospitalized with venous thromboembolism

Isabelle Mahé[1,2] | Yara Skaff[1] | Hélène Helfer[1] | Samuel Benarroch[1] | Florent Happe[1] | Adam Remaki[3] | Kankoe Sallah[4]

JMIR Med Inform. 2025 Jan 25. doi: 10.2196/68704. [Epub ahead of print]

Improving Phenotyping of Patients With Immune-Mediated Inflammatory Diseases Through Automated Processing of Discharge Summaries: Multicenter Cohort Study.

Remaki A[1], Ung J[2], Pages P[2], Wajsburt P[2], Liu E[3], Faure G[1], Petit-Jean T[2], Tannier X[1], Gérardin C[1,4].

Adjusting for the progressive digitization of health records: working examples on a multi-hospital clinical data warehouse

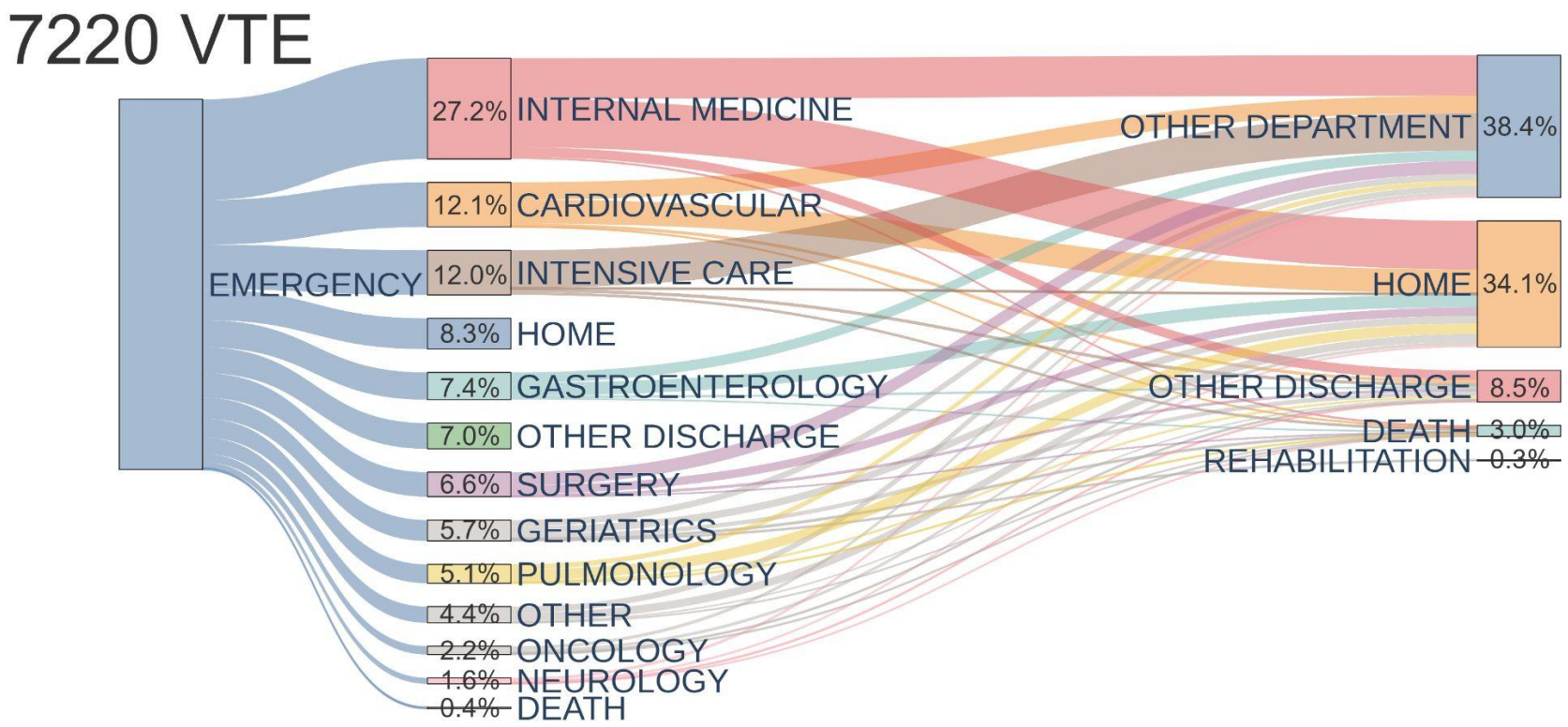
Adam Remaki, Benoît Playe, Paul Bernard, Simon Vittoz, Matthieu Doutreligne, Gilles Chatelier, Etienne Audureau, Emmanuelle Kempf, Raphaël Porcher, Romain Bey

doi: https://doi.org/10.1101/2023.08.17.23294220

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

A quoi peuvent servir les données de l'EDS?

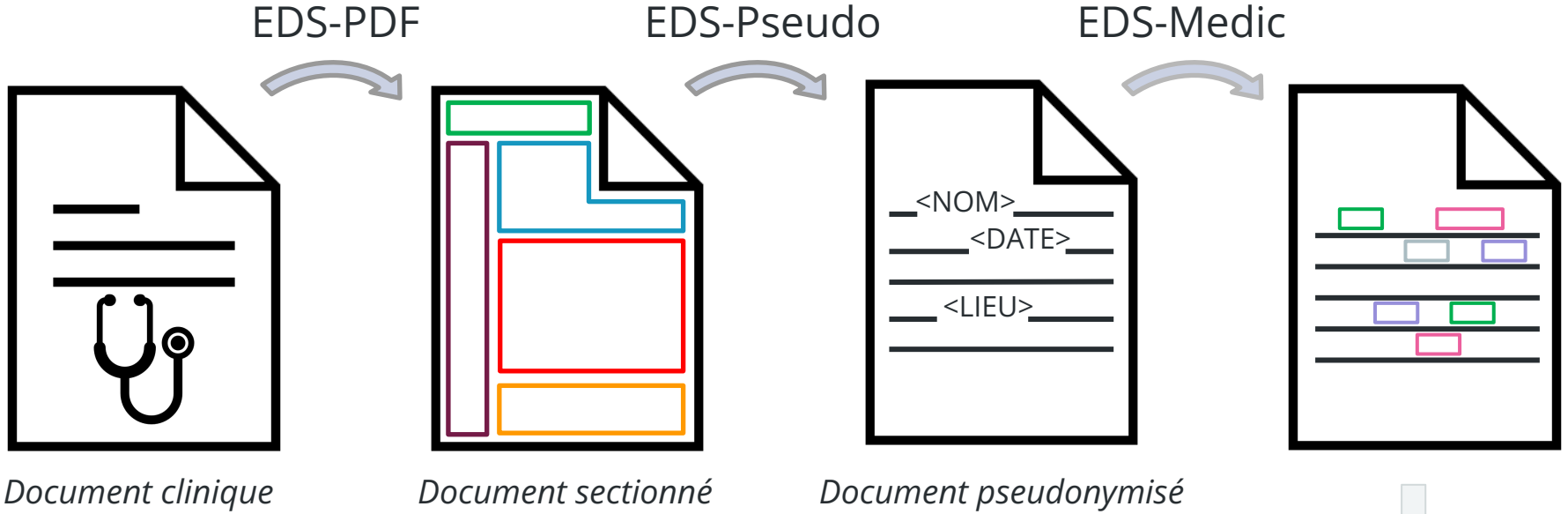
- Parcours de soins pour des patients pour thrombose veineuse
  - Description des parcours de patients arrivés aux urgences



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

A quoi peuvent servir les données de l'EDS?

- **Algorithme d'extraction d'entités médicamenteuses par NLP**



person_id	note_id	drug	dose	freq
1	1	MedA	3mg	1 fois par jour
1	2	MedB	1 comp	tous les matins
…	…	…	…	…

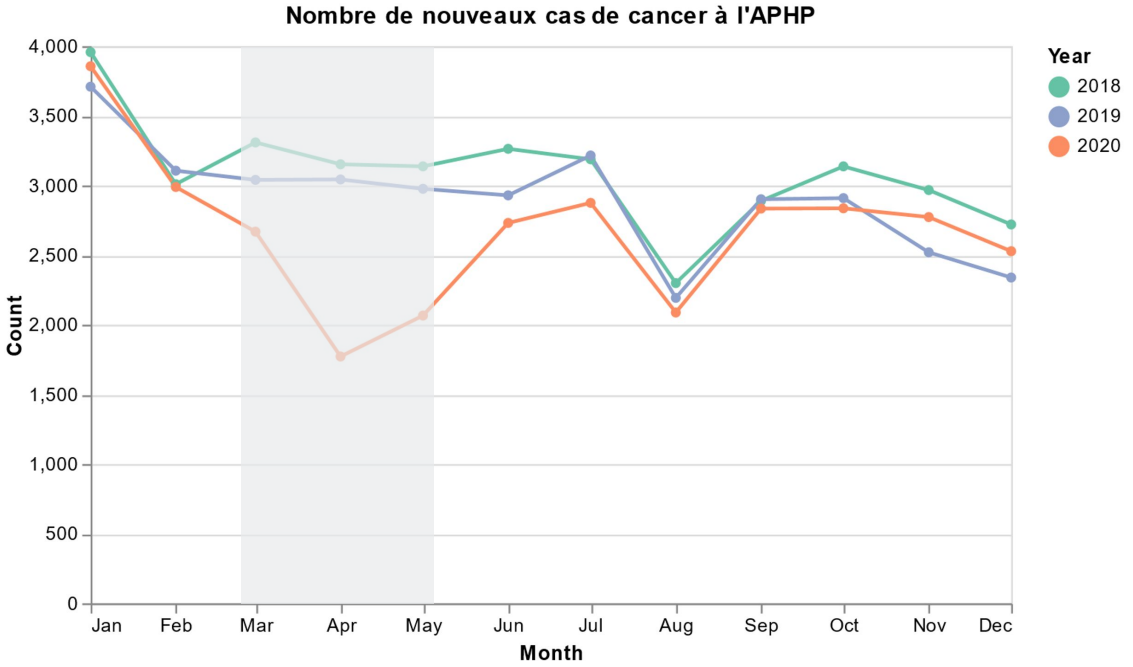
Table finale

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

A quoi peuvent servir les données de l'EDS?

Impact des confinements sur les diagnostics de cancer

- Evaluation du nombre de nouvelles prises en charge de cancer à l'AP-HP
  - *Détection des nouveaux cancers*
  - *Analyse de la prise en charge des patients pendant la pandémie de Covid-19*



*Kempf, E., Lamé, G., Layese, R., Priou, S., Chatellier, G., Chaieb, H., ... & de Paris Cancer, A. P. H. (2021). New cancer cases at the time of SARS-Cov2 pandemic and related public health policies: A persistent and concerning decrease long after the end of the national lockdown. *European Journal of Cancer*, *150*, 260-267.

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

A quoi peuvent servir les données de l'EDS?

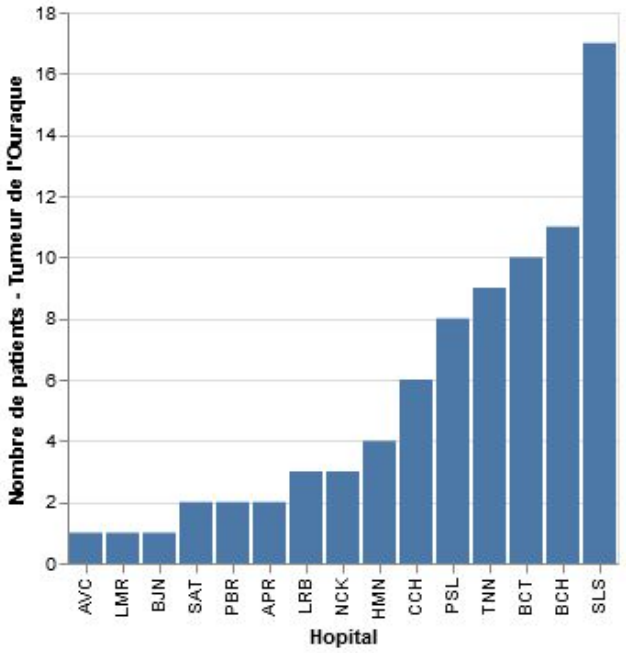
Détection de patients atteints de maladies rares

- Exemple: détection de patients atteints d'une tumeur de l'ouraque
  - *Environ 80 patients pris en charge à l'AP-HP*
  - ***Données massives:*** *couvrir une grande population, de nombreux hôpitaux*
  - ***Données variées:*** *utiliser plusieurs critères de détection*

- Données PMSI:
  - *Codes CIM10 peu renseignés*
  - *Peu de FP, mais nombreux FN*

- Données textuelles:
  - *CRH mieux renseignés*
  - *Recherche (tumeur & Ouraque) | (adenocarcinome & Ouraque)*
  - *Peu de FN, beaucoup de FP*

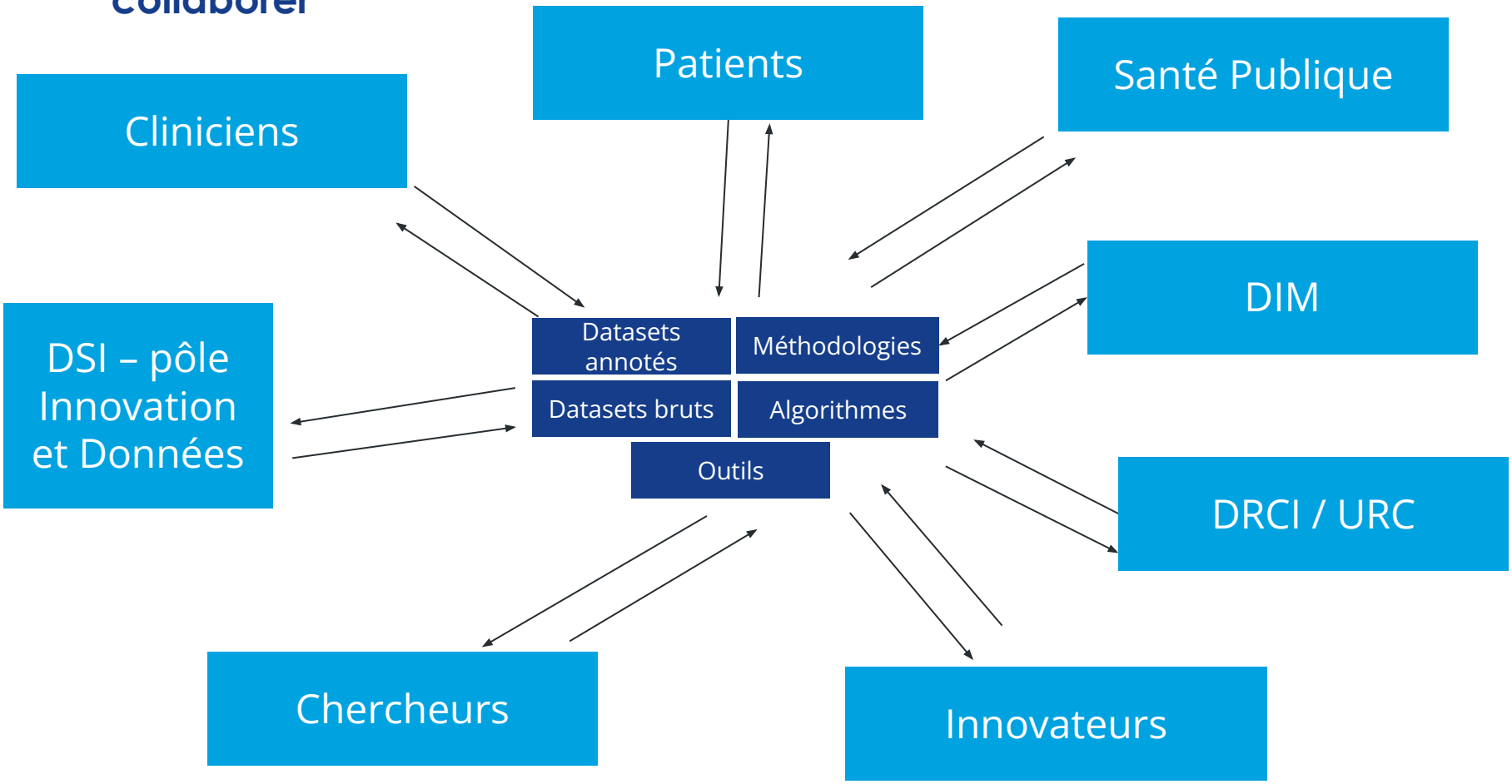
	Doc+	Doc-	Total
PMSI+	13	6	19
PMSI-	59	XX	
Total	72		



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Comment les acteurs de l'EDS interagissent-ils?

- **L'EDS n'est pas qu'une solution technique, c'est avant tout une plateforme permettant à une communauté de se structurer et de collaborer**



ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Objectifs de l'équipe

- Développer des **outils** (Python), **algorithmes** (NLP-phénotypage) et **méthodologies** (gestion de biais) pour faciliter l'exploitation des bases du pôle I&D (en particulier l'EDS)
  - *Livrables: algo NLP, algo appariement probabiliste, tools, librairie, rapports données*
- Faciliter **l'échange de méthodologies/code** entre projets (recherche/pilotage). Structurer le réseau de data scientists. Améliorer la reproductibilité des recherches réalisées à l'EDS
  - *Animation technique et scientifique à destination des data scientists (recherche et pilotage)*
  - *Coordination et co-développement de librairies de pre-processing*
  - *Partage de notebooks sur Gitlab et/ou Github*
  - *Mise en production de certains algorithmes d'intérêt développés en dehors de la DSI*
- **Piloter les partenariats IA/ML**
- Réaliser les **études ad-hoc** demandées par la direction

2.

Présentation de

l’Enseignement d’Intégration

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Objectifs :

- Initiation aux traitements statistiques usuels dans le cadre de projets de recherche clinique et épidémiologique
- Sensibilisation aux biais techniques et cliniques inhérents aux données de vie réelle
- Formalisation des résultats sous forme d'un court article scientifique + un notebook reproductible

Organisation de la semaine

Programme

- Lundi 02/06 : Introduction, exploration des données, analyse de survie
- Mardi 03/06 : Déduplication des identités
- Mercredi 04/06 : Traitement Automatique du Langage
- Jeudi 05/06 : Analyse biostatistique (1/2)
- Vendredi 06/06 : Analyse biostatistique (2/2) & Finalisation du projet

Organisation pratique

- Matin : présentation des enjeux + exercices associés / Après-midi : projet
- Support sur GitHub, code en Python

Livrables :

- Court article scientifique (~1500 mots) – 50% de la note
- Code (notebook) permettant de reproduire les résultats présentés – 50% de la note

Usage des IA génératives

Ce Projet est à visée pédagogique et dans ce cadre, nous jugeons qu'il est préférable d'éviter l'utilisation des IA génératives (Copilot, ChatGPT…etc)

- **Pour les exercices: À éviter totalement.** Nous tenons à ce que vous cherchiez les réponses par vous-même, en passant par la documentation des librairies, ou en cas de gros blocage en sollicitant les autres personnes de votre groupe ou en nous sollicitant.

- **Pour le Projet: Utilisation tolérée pour la partie notebook, à condition d'être capable de d'expliquer chaque ligne du code. À ne surtout pas utiliser pour la partie rédaction (cela se voit).** Une utilisation pour la partie notebook, nous paraissant inappropriée et confuse sera pénalisée. Idem, pour la rédaction de l'article.

3.

Exercice 1 – Exploration des données

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Application 1 – Exploration des données

- **Installer un IDE (VS Code)**
  - VS Code : https://code.visualstudio.com/download

- **Cloner le repo GitHub**
  - Lien URL : https://github.com/Aremaki/edstuto_2025

- **Ouvrir le notebook *exercises/exercise-1***
  - Contexte : Analyse épidémiologique pour comparer deux médicaments traitant la grippe – DrugA et DrugB
  - Objectif : construire des courbes de survie pour comparer l'effet des médicaments sur la population
  - Plan d'action :
    - *Prise en main des tables*
    - *Pré-traitement pour nettoyer les tables*
    - *Tracé des courbes de survie*

Estimation de Kaplan-Meier

- Taux de survie : probabilité d'être vivant à un instant t
- Objectif : représenter graphiquement la survie d'une ou de plusieurs populations

Exemple courbe de survie

Durée d'observation (en mois)	Statut vital
2	1
3	0
6	1
6	1
7	1
10	0
15	1
15	1
16	1
27	1
30	1
32	1

Patient *décédé*

Patient *perdu de vue*

Quelle est la probabilité de survie à *n* mois?

*Les données sont fictives

Exemple courbe de survie

On veut estimer la probabilité P(T>t) qu'un membre d'une population N donnée, ait une durée de vie supérieure à t. T la variable aléatoire qui mesure le temps de survie

Durée	Statut vital
2	1
3	0
6	1
6	1
7	1
10	0
15	1
15	1
16	1
27	1
30	1
32	1

t n d P(T>t)

Temps (en mois)	# de patients à risque	# de patients décédés	Proba de survie > t
0	12	0	100%
2	12	1	$\frac{12-1}{12}$ 92%

Etat initial

Le 1er décès est à 2 mois

Nombre de patients vivants juste avant la période *2 mois*

Nombre de patients décédé à *2 mois*

Pourcentage de patients survivant au delà de *2 mois*

*Les données sont fictives

Exemple courbe de survie

Durée	Statut vital
2	1
3	0
6	1
6	1
7	1
10	0
15	1
15	1
16	1
27	1
30	1
32	1

A *3 mois*, pas de décès

t n d P(T>t)

Temps (en mois	# de patients à risque	# de patients décédés	Proba de survie > t
0	12	0	100%
2	12	1	92 %

*Les données sont fictives

Exemple courbe de survie

Durée	Statut vital
2	1
3	0
6	1
6	1
7	1
10	0
15	1
15	1
16	1
27	1
30	1
32	1

t	n	d	P(T>t)
Temps (en mois	**# de patients à risque**	**# de patients décédés**	**Proba de survie > t**
0	12	0	100%
2	12	1	92 %
6	10	2	

Décès suivant

Nombre de patients vivants juste avant la période 6 *mois*

Nombre de patients décédé à 6 *mois*

*Les données sont fictives

Exemple courbe de survie

Durée	Statut vital
2	1
3	0
6	1
6	1
7	1
10	0
15	1
15	1
16	1
27	1
30	1
32	1

t	n	d	P(T>t)
Temps (en mois	**# de patients à risque**	**# de patients décédés**	**Proba de survie > t**
0	12	0	100%
2	12	1	92 %
6	10	2	74%

Probabilité de survivre au-delà de *6 mois* = Probabilité de survivre au-delà de *2 mois* x Probabilité de survivre à *6 mois*

$$= 0.92 \times \frac{10 - 2}{10}$$

*Les données sont fictives

Exemple courbe de survie

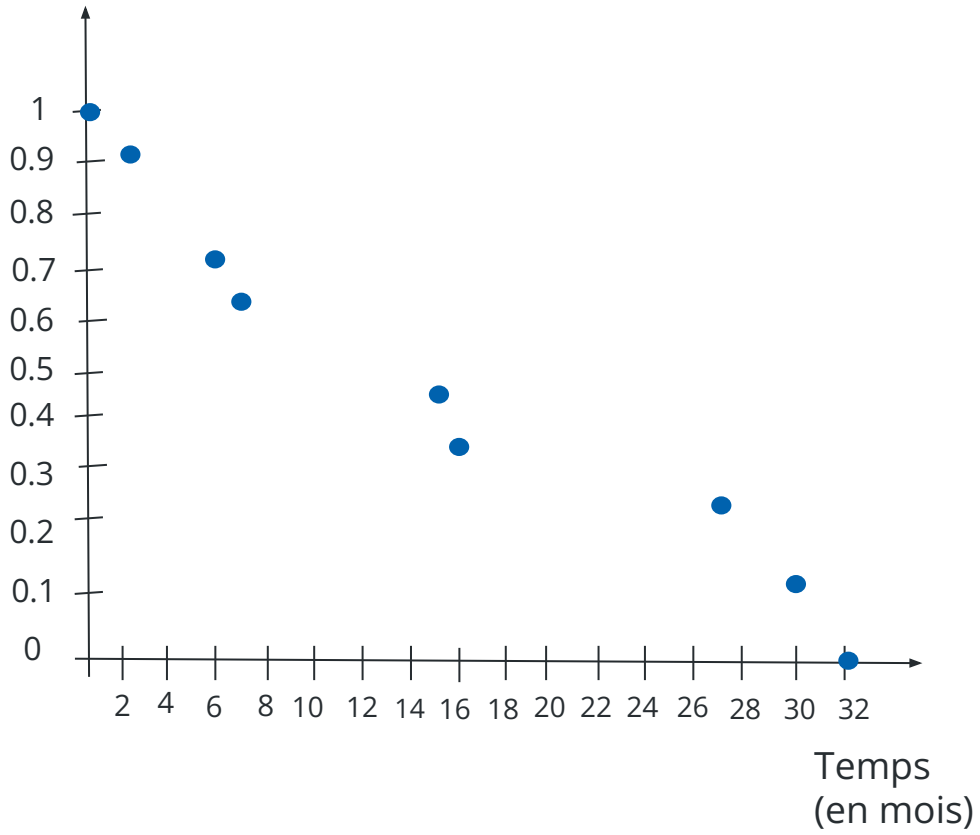
Durée	Statut vital
2	1
3	0
6	1
6	1
7	1
10	0
15	1
15	1
16	1
27	1
30	1
32	1

t	n	d	P(T>t)
Temps (en mois	**# de patients à risque**	**# de patients décédés**	**Proba de survie > t**
0	12	0	100%
2	12	1	92 %
6	10	2	74%
7	8	1	64%
15	6	2	43%
16	4	1	32%
27	3	1	21%
30	2	1	10%
32	1	1	0%

*Les données sont fictives

Exemple courbe de survie

Temps (en mois	# de patients à risque	# de patients décédés	Proba de survie > t
0	12	0	100%
2	12	1	92 %
6	10	2	74%
7	8	1	64%
15	6	2	43%
16	4	1	32%
27	3	1	21%
30	2	1	10%
32	1	1	0%

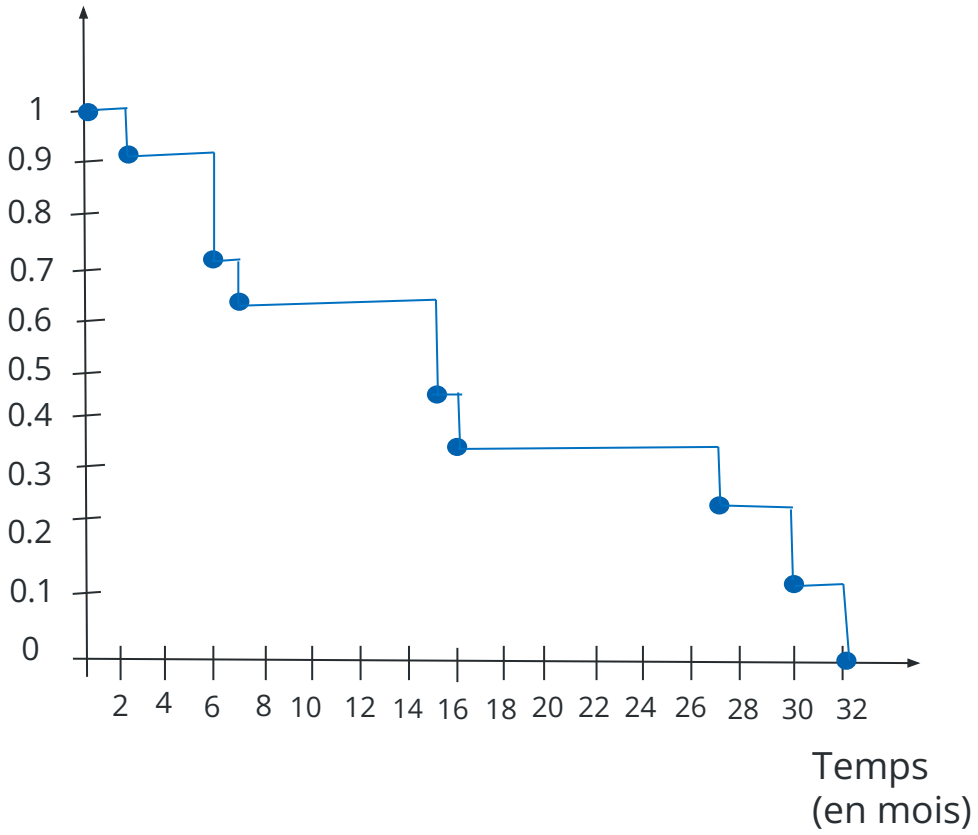


*Les données sont fictives

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

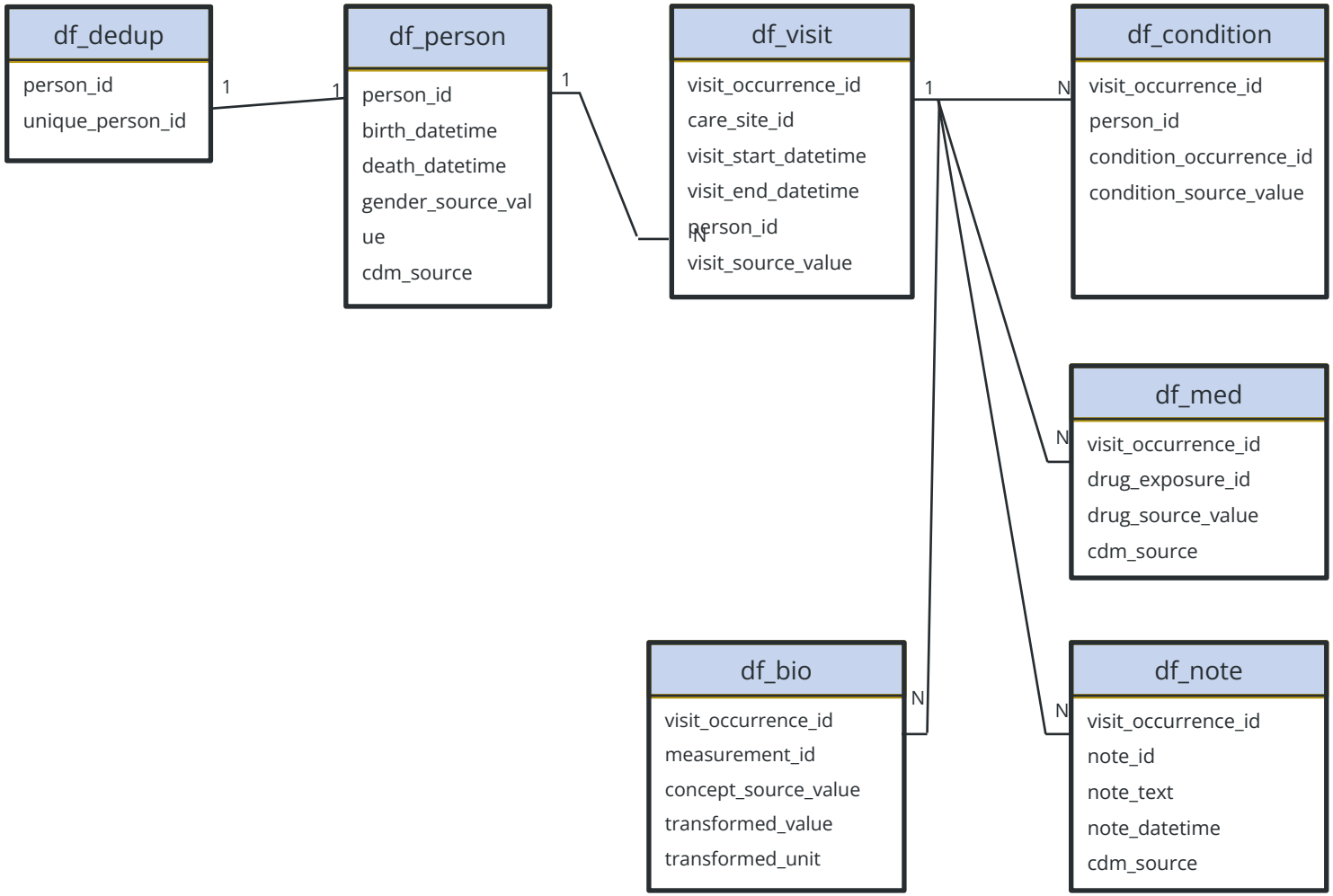
Exemple courbe de survie

Temps (en mois	# de patients à risque	# de patients décédés	Proba de survie > t
0	12	0	100%
2	12	1	92 %
6	10	2	74%
7	8	1	64%
15	6	2	43%
16	4	1	32%
27	3	1	21%
30	2	1	10%
32	1	1	0%



*Les données sont fictives

Schéma de la base mise à disposition



4.

Projet – Exploration des données & revue scientifique

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Projet – Exploration des données

Question scientifique

- Identification des facteurs de risque associés au cancer du sein
- Analyse rétrospective à partir des données de l'EDS

Objectifs globaux

- Synthèse de la littérature sur le sujet
- Construction des objectifs techniques pour répondre à la question scientifique
- Nettoyage/Traitement de la BDD mise à disposition
- Analyses statistiques
- Restitution
  - *Court article scientifique (~1500 mots)*
  - *Notebook reproductible*

Projet – Exploration des données

Objectifs du jour

- Lecture de la base
- Exploration & mapping des données
- Revue documentaire (introduction de l'article scientifique)

Données

- Base de données disponible dans le dossier *final_project/data*

Bibliographie

- *Kohane, Isaac S, Bruce J Aronow, Paul Avillach, Brett K Beaulieu-Jones, Riccardo Bellazzi, Robert L Bradford, Gabriel A Brat, et al. « What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask ». Journal of Medical Internet Research 23, n° 3 (2 mars 2021): e22219. https://doi.org/10.2196/22219.*
- *Agniel, Denis, Isaac S Kohane, et Griffin M Weber. « Biases in Electronic Health Record Data Due to Processes within the Healthcare System: Retrospective Observational Study ». BMJ, 30 avril 2018, k1479. https://doi.org/10.1136/bmj.k1479.*
- *Wilson, Greg, D. A. Aruliah, C. Titus Brown, Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, et al. « Best Practices for Scientific Computing ». Édité par Jonathan A. Eisen. PLoS Biology 12, n° 1 (7 janvier 2014): e1001745. https://doi.org/10.1371/journal.pbio.1001745.*
- *Benchimol, Eric I., Liam Smeeth, Astrid Guttmann, Katie Harron, David Moher, Irene Petersen, Henrik T. Sørensen, Erik von Elm, Sinéad M. Langan, et RECORD Working Committee. « The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement ». PLOS Medicine 12, n° 10 (6 octobre 2015): e1001885. https://doi.org/10.1371/journal.pmed.1001885.*