HEALTH DATA HUB

Health Data Hub

Journée de l'open science en santé

PariSanté Campus
04/06/2025

Journée de l’open science en santé - Programme

09h30 - 09h35	**Message de bienvenue**
09h35 - 09h50	**Ouverture de la journée**, Clara Locher (CHU de Rennes)
09h50 - 10h20	**INSERM et science ouverte : renforcer l’impact et l’efficacité de la recherche en santé**, Michel Pohl (Inserm)
10h20 - 10h40	**L’AMDAC Solidarités-Santé : encourager et valoriser l’ouverture des données et des codes**, Claude Gissot (DREES)
10h40 - 10h55	**Pause**
10h55 - 11h15	**La Cartographie des pathologies : un outil désormais ouvert au service de la recherche et de la santé publique**, Antoine Rachas (CNAM)
11h15 - 11h45	**Ouvrir l’accès aux ressources scientifiques : les initiatives de l’INRAE en faveur de l’ouverture de la science**, Odile Hologne (INRAE)
11h45 - 12h30	**Annonce des lauréats de la 8e vague de l’AMI de la Bibliothèque Ouverte d’Algorithmes en Santé (BOAS)**, Maxime Caillet (HDH)
12h30 - 14h00	**Pause**

Journée de l’open science en santé - Programme

14h00 - 14h25	**The publicly-available Medical Information Mart for Intensive Care (MIMIC) Database: An Open Data Success Story**, Leo Anthony Celi (MIT)
14h25 - 14h50	**Les initiatives du Health Data Hub en faveur de l'ouverture de la science**, Laurie Alla (HDH)
14h50 - 15h05	**Le programme Data Challenges en santé, catalyseur d’innovations ouvertes**, Lauriane Armand (HDH)
15h05 - 15h20	**Le Data Challenge CytologIA - Améliorer le diagnostic en hématologie biologique grâce à l’IA**, Dr Samy Dahmani (Algoscope)
15h20 - 15h40	**Données ouvertes et IA en hématologie biologique : quelles perspectives après le Data Challenge CytologIA ?**, Dr Thomas Boyer (GFHC)
15h40 - 16h00	**Pause**
16h00 - 17h00	**Présentation des résultats et remise des prix du Data Challenge CytologIA**
17h00 - 17h20	**L'open data, un catalyseur de l'engagement citoyen au service de la santé**, Augustin Courtier (Latitudes)
17h20 - 17h30	**Conclusion et remerciements**
17h30 - 19h00	**Cocktail**

Ouverture de la journée

HEALTH DATA HUB

Ouverture de la journée

09h35 - 09h50

Clara Locher

Praticien Hospitalier en Pharmacologie Clinique

Clara Locher est Praticien Hospitalier en Pharmacologie Clinique. Elle exerce comme méthodologiste au CHU de Rennes et est membre de la Commission de la Transparence de la Haute Autorité de Santé. Son activité en méta-recherche porte sur les conflits entre éditeurs et auteurs, la qualité et la transparence des essais d'escalade de dose en oncologie ainsi que sur l'impact du partage des données des essais cliniques en oncologie.

Science ouverte appliquée aux spécificités de la recherche clinique

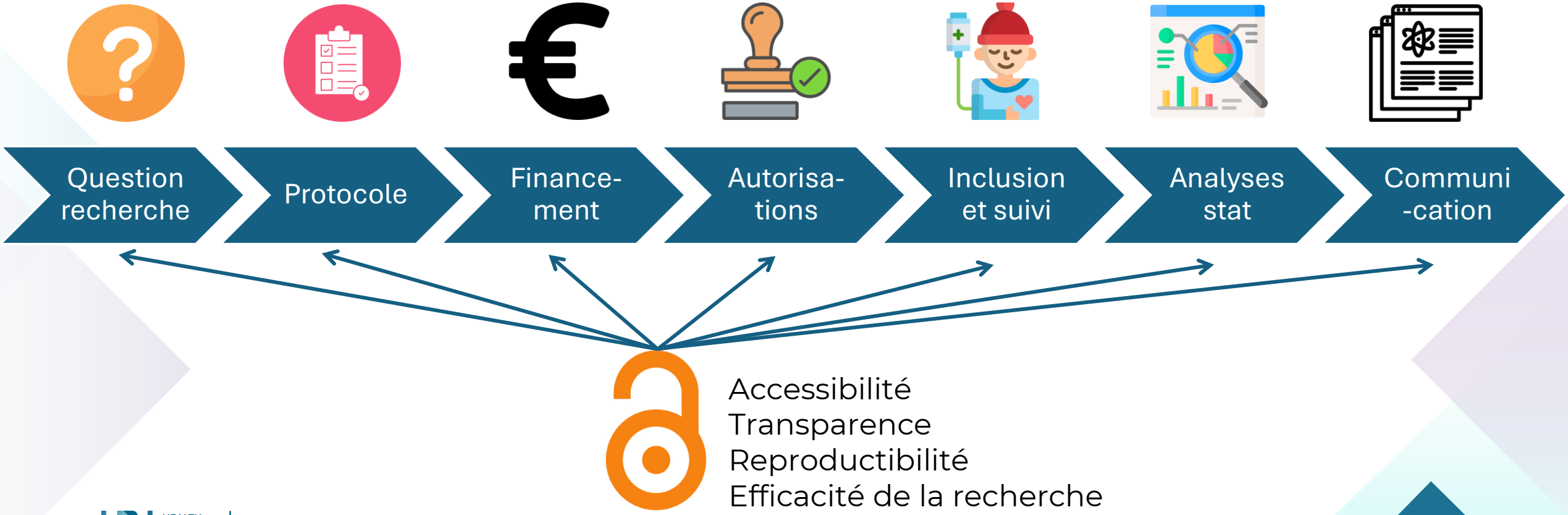
Déclaration de liens d'intérêt

- Aucun avec organisme privé
- Plusieurs avec organismes publics
  - CHU de Rennes : employeur
  - Haute Autorité de Santé – Commission de la Transparence

  - Agence Nationale de la Recherche – RESTORES
  - Union Européenne – SHARE-CTD

Science ouverte

- La science ouverte est la diffusion **sans entrave** des résultats, des méthodes, et des produits de la recherche

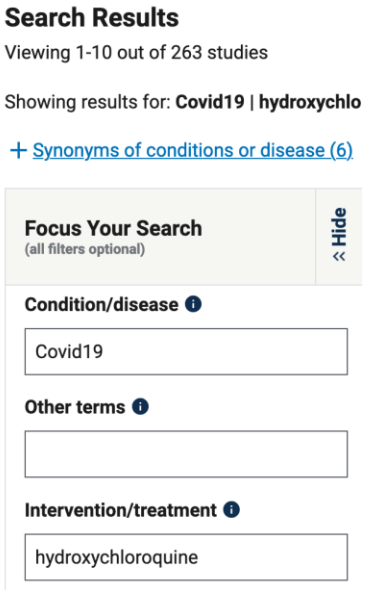


Science ouverte et protocole

- Fin 90 : registres d'essais cliniques

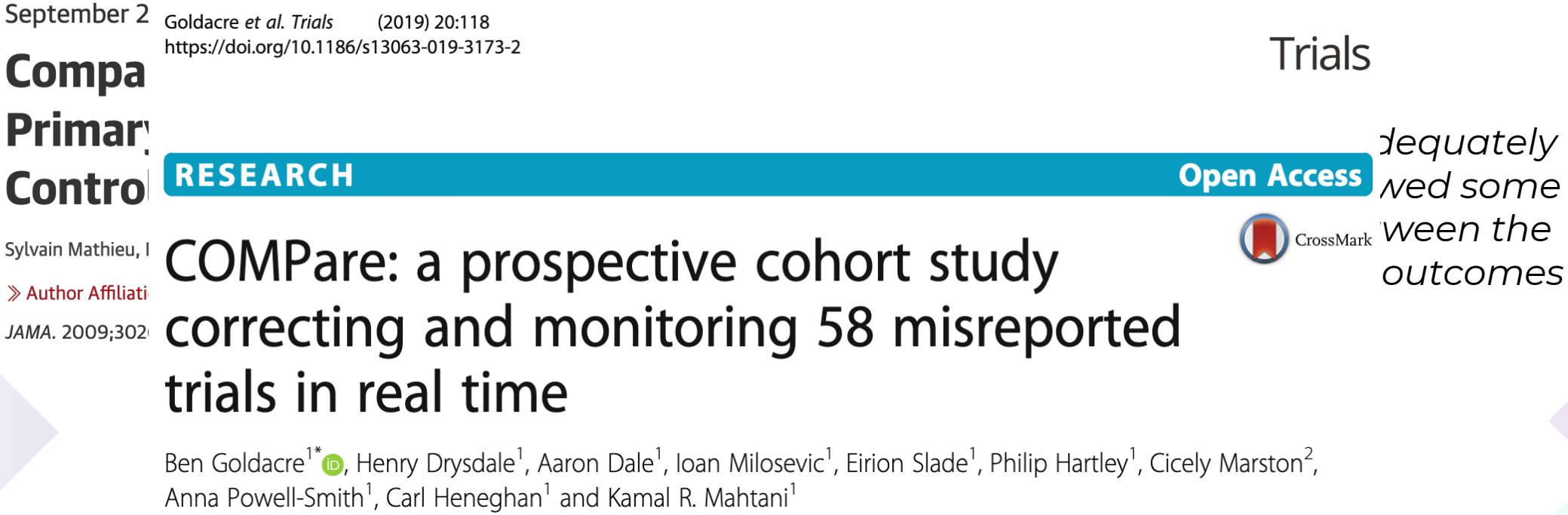
  Clinical Trials ClinicalTrials.gov

- 2004 : obligation d'enregistrement **a priori**
  - Rôle majeur des journaux
  - ICMJE : *International Committee of Medical Journal Editors*
- Objectifs
  - Aider patients & public à connaitre les essais en cours
  - Prévenir la duplication d'essais
  - Prévenir les biais de *reporting*
  - Prévenir les biais de publication



Prévenir les biais de *reporting*

- Enregistrement sur un registre d'essais cliniques
  - Améliore transparence mais… insuffisant
  - Intérêt de l'enregistrement si personne ne vérifie l'adéquation ?



September 2

Compa
Primar
Contro

Sylvain Mathieu, I

» Author Affiliati

JAMA. 2009;302

Goldacre *et al. Trials* (2019) 20:118
https://doi.org/10.1186/s13063-019-3173-2

Trials

RESEARCH Open Access

CrossMark

COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time

Ben Goldacre[1*], Henry Drysdale[1], Aaron Dale[1], Ioan Milosevic[1], Eirion Slade[1], Philip Hartley[1], Cicely Marston[2], Anna Powell-Smith[1], Carl Heneghan[1] and Kamal R. Mahtani[1]

dequately
wed some
ween the
outcomes

Prévenir les biais de publication

WORLD MEDICAL ASSOCIATION

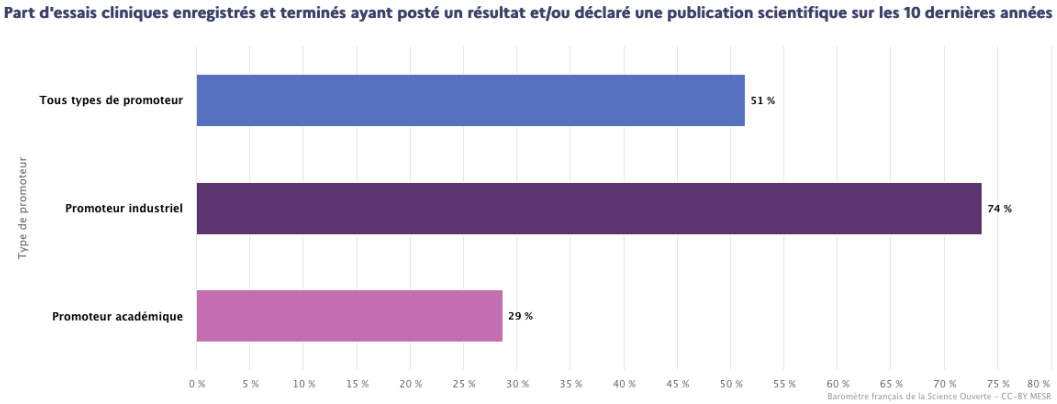
WMA DECLARATION OF HELSINKI – ETHICAL PRINCIPLES FOR MEDICAL RESEARCH INVOLVING HUMAN PARTICIPANTS

- Les résultats doivent être diffusés
  - quelque soit l'orientation des résultats
  - sans entrave
    - Registres d'essais cliniques
    - Serveur de préprint
    - *Open access*
- Baromètre français de la Science Ouverte

Results Posted

medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES

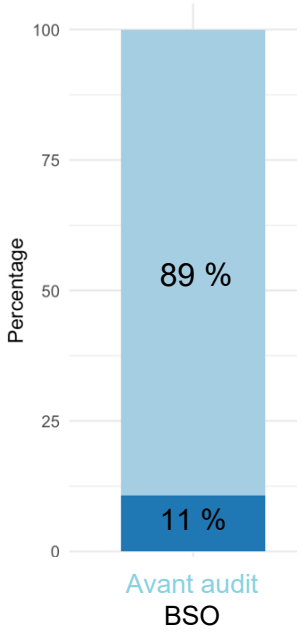


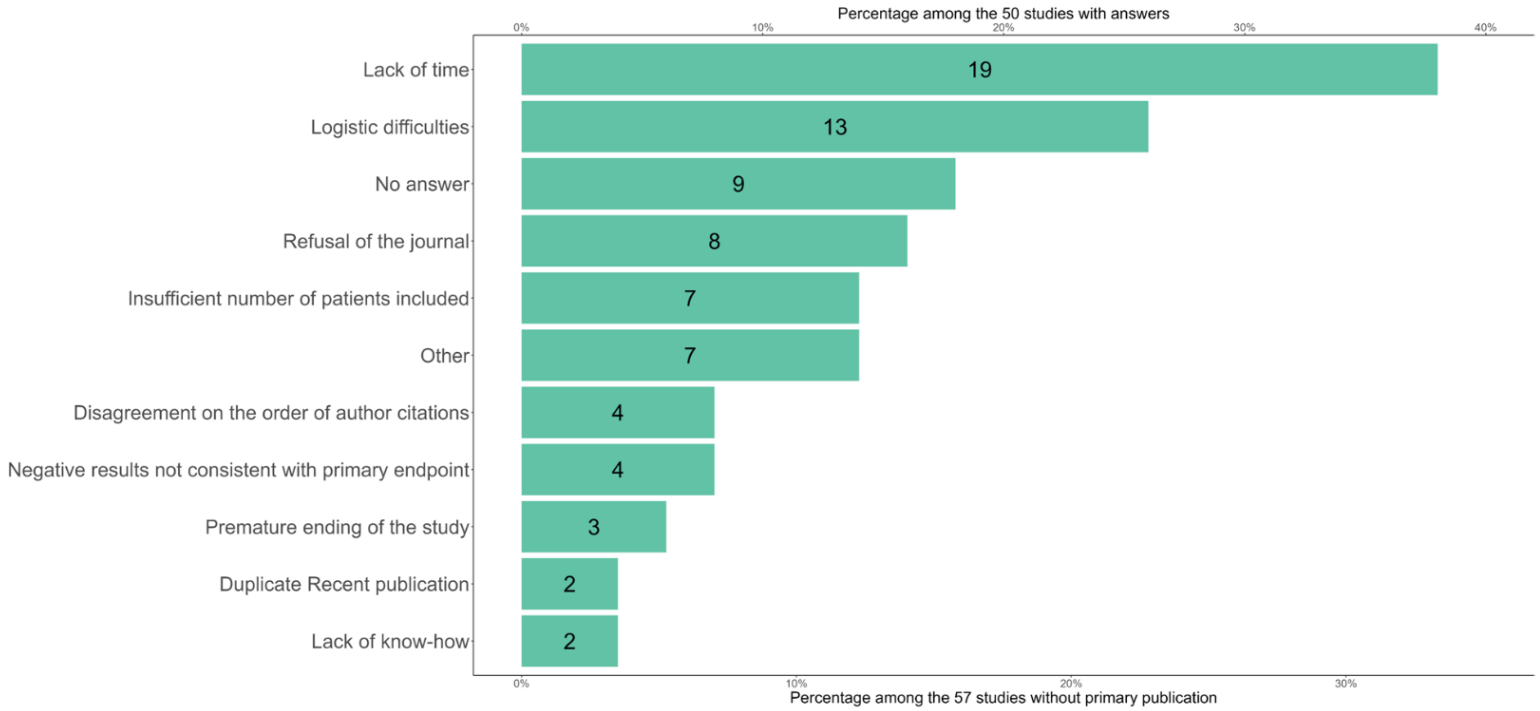
Prévenir les biais de publication

RESEARCH Open Access

Reporting of interventional clinical trial results in an academic center: a survey of completed studies

Anne Sophie Alix-Doucet[1], Constant Vinatier[2], Loïc Fin[1], Hervé Léna[3], Hélène Rangé[4], Clara Locher[2] and Florian Naudet[2,5*]





Rôle majeur des financeurs

- Conditionner le financement au respect des principes de science ouverte
  - Exiger l'enregistrement *a priori*
  - Exiger un plan de gestion des données
  - Exiger la communication des résultats
    - dernière tranche conditionnée par…
    - ~~la soumission d'un article~~ le *posting* des résultats
- Financer le temps RH nécessaire au respect des principes de science ouverte

Données individuelles des patients [IPD] : partage

- Barrières nombreuses +++ [sans être insurmontables !]

Barrières juridiques	Protection de la vie privée | du secret médical
Barrières motivationnelles	Pas d'incitations, peur des critiques, désaccord | compétition sur la réutilisation des données
Barrières financières	Manque de ressources
Barrières techniques	Manque de normalisation, données manquantes, barrière linguistique, absence de métadonnées

- Partages des IPD = condition nécessaire (mais pas suffisante) à leur **réutilisation**
  - **Findable** (Facile à trouver)
  - **Accessible** (Accessible)
  - **Interoperable** (Interopérable)
  - **Reusable** (Réutilisable)

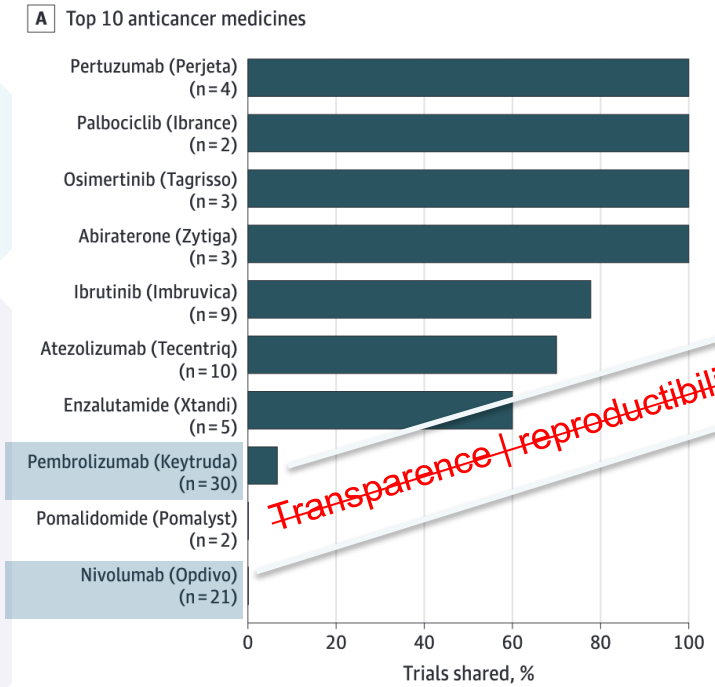
En pratique, les IPD sont-elles partagées ?

JAMA Oncology | Original Investigation

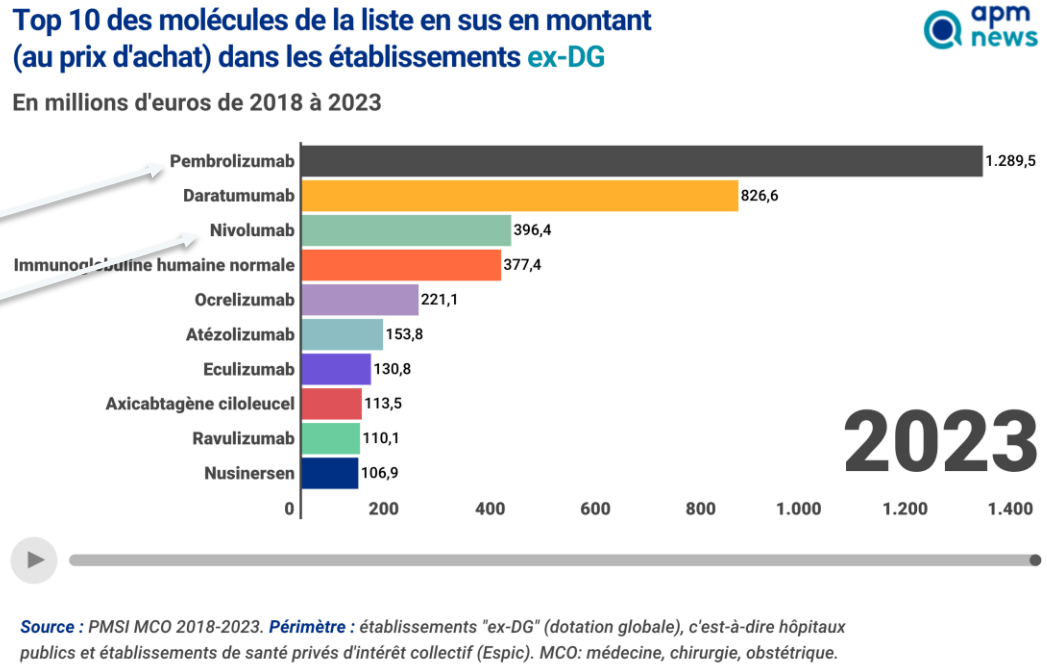
Audit of Data Sharing by Pharmaceutical Companies for Anticancer Medicines Approved by the US Food and Drug Administration

Natansh D. Modi, BPharm (Hons); Ahmad Y. Abuhelwa, PhD; Ross A. McKinnon, PhD; Alan V. Boddy, PhD; Mark Haseloff; Michael D. Wiese, PhD; Tammy C. Hoffmann, PhD; Eric D. Perakslis, PhD; Andrew Rowland, PhD; Michael J. Sorich, PhD; Ashley M. Hopkins, PhD

RESULTS During the 10-year period examined, 115 anticancer medicines were approved by the FDA on the basis of evidence from 304 pharmaceutical industry–sponsored trials. Of these trials, 136 (45%) were eligible for IPD sharing and 168 (55%) were not. Data sharing rates differed substantially among industry sponsors, with the most common reason for not sharing trial IPD being that the collection of long-term follow-up data was still ongoing (89 of 168 trials [53%]). Of the top 10 anticancer medicines by global sales, nivolumab, pembrolizumab, and pomalidomide had the lowest eligibility rates for data sharing (<10% of trials).



~~Transparence | reproductibilité | analyses secondaires~~



Comment augmenter le partage des IPD ?

- Informer | Former
  - Au partage et à la réutilisation
- Tenir compte du partage dans l'évaluation des chercheurs
  - Recommandation de la déclaration de DORA
- Rendre la déclaration de partage obligatoire
  - ICMJE
- Rendre le partage obligatoire ?
  - Toutes les données non pas la même valeur…
  - Pas nécessairement adapté
- Mesurer l'impact du partage (et de la réutilisation)

Illustration by David Parkins

IPD et réutilisation responsable

- Former sur les aspects **réglementaires**
  - Réglementation complexe (et variable selon les pays)
- Former sur les aspects de **sécurité**
  - Minimiser les risques via des plateformes adaptées
- Former sur les aspects **scientifiques**
  - Promouvoir des pratiques de recherche reproductible
  - Minimiser les « faux positifs », les biais de publication, le p-harking, le p-hacking
- **LORIER** L'Organisation pour une Recherche Inserm Ethique et Responsable
- **RRFR** réseau de la recherche reproductible

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Inserm
La science pour la santé
From science to health

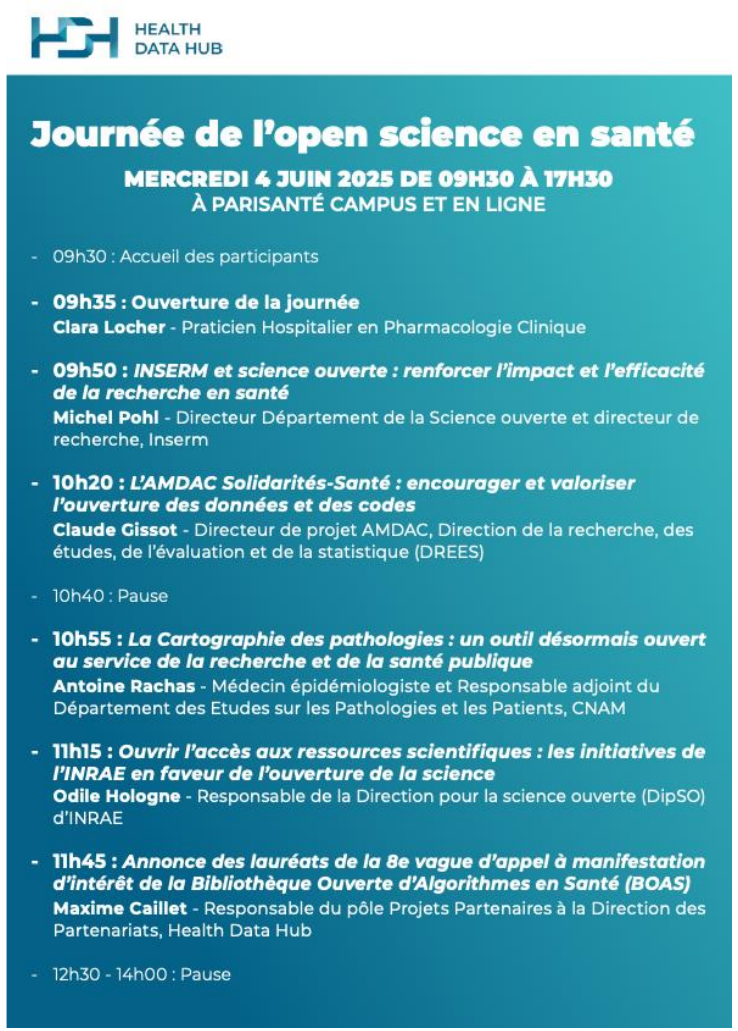
lorier

Conclusion

- Toutes les étapes de la recherche clinique concernées
- Tous les acteurs de la recherche clinique concernés
  - Chercheurs, promoteurs, patients,
  - Financeurs, éditeurs, évaluateurs : incitatives SO
- Doit être anticipé dès la phase de conception de l'étude
- Formation | sensibilisation aux enjeux de la SO indispensables

Programme de la journée

- Promouvoir l'ouverture de la science auprès de l'écosystème santé
- Valoriser les travaux partagés en open source et en open data



HEALTH DATA HUB

Journée de l'open science en santé

MERCREDI 4 JUIN 2025 DE 09H30 À 17H30
À PARISANTÉ CAMPUS ET EN LIGNE

- 09h30 : Accueil des participants
- **09h35 : Ouverture de la journée**
  Clara Locher - Praticien Hospitalier en Pharmacologie Clinique
- **09h50 : *INSERM et science ouverte : renforcer l'impact et l'efficacité de la recherche en santé***
  Michel Pohl - Directeur Département de la Science ouverte et directeur de recherche, Inserm
- **10h20 : *L'AMDAC Solidarités-Santé : encourager et valoriser l'ouverture des données et des codes***
  Claude Gissot - Directeur de projet AMDAC, Direction de la recherche, des études, de l'évaluation et de la statistique (DREES)
- 10h40 : Pause
- **10h55 : *La Cartographie des pathologies : un outil désormais ouvert au service de la recherche et de la santé publique***
  Antoine Rachas - Médecin épidémiologiste et Responsable adjoint du Département des Etudes sur les Pathologies et les Patients, CNAM
- **11h15 : *Ouvrir l'accès aux ressources scientifiques : les initiatives de l'INRAE en faveur de l'ouverture de la science***
  Odile Hologne - Responsable de la Direction pour la science ouverte (DipSO) d'INRAE
- **11h45 : *Annonce des lauréats de la 8e vague d'appel à manifestation d'intérêt de la Bibliothèque Ouverte d'Algorithmes en Santé (BOAS)***
  Maxime Caillet - Responsable du pôle Projets Partenaires à la Direction des Partenariats, Health Data Hub
- 12h30 - 14h00 : Pause

HEALTH DATA HUB

- **14h00 : *The publicly-available MIMIC Database: An Open Data Success Story***
  Leo Anthony Celi - MD, Research director and principal research scientist at the MIT Laboratory for Computational Physiology (LCP)
- **14h25 : *Les initiatives du Health Data Hub en faveur de l'ouverture de la science***
  Laurie Alla - Cheffe de projet Open Science, Health Data Hub
- **14h50 : *Le programme Data Challenges en santé, catalyseur d'innovations ouvertes***
  Lauriane Armand - Cheffe de projet, Health Data Hub, PharmD
- **15h05 : *Le Data Challenge CytologIA - Améliorer le diagnostic en hématologie biologique grâce à l'IA***
  Dr Samy Dahmani - Biologiste médical et co-fondateur d'Algoscope
- **15h20 : *Données ouvertes et IA en hématologie biologique : quelles perspectives après le Data Challenge CytologIA ?***
  Dr Thomas Boyer - Hématologue et biologiste au CHU d'Amiens, Secrétaire adjoint du Groupe Francophone d'Hématologie Cellulaire (GFHC)
- 15h40 : Pause
- **16h00 : Présentation des résultats et remise des prix du Data Challenge CytologIA :**
  - 3ème place : **Simon Thomine** - Ingénieur de recherche, VitaDX
  - 2ème place : **Xueer Chen** - Senior Scientist, Bristol Myers Squibb
  - 1ère place : **Eric Ben Hamou** - Senior Software Engineer et Data Scientist
- **17h00 : *L'open data, un catalyseur de l'engagement citoyen au service de la santé***
  Augustin Courtier - Co-fondateur de l'association Latitudes
- **17h20 : Conclusion et remerciements**
- 17h30 - 19h00 : Cocktail

INSERM et science ouverte : renforcer l'impact et l'efficacité de la recherche en santé

HEALTH DATA HUB

INSERM et science ouverte : renforcer l'impact et l'efficacité de la recherche en santé

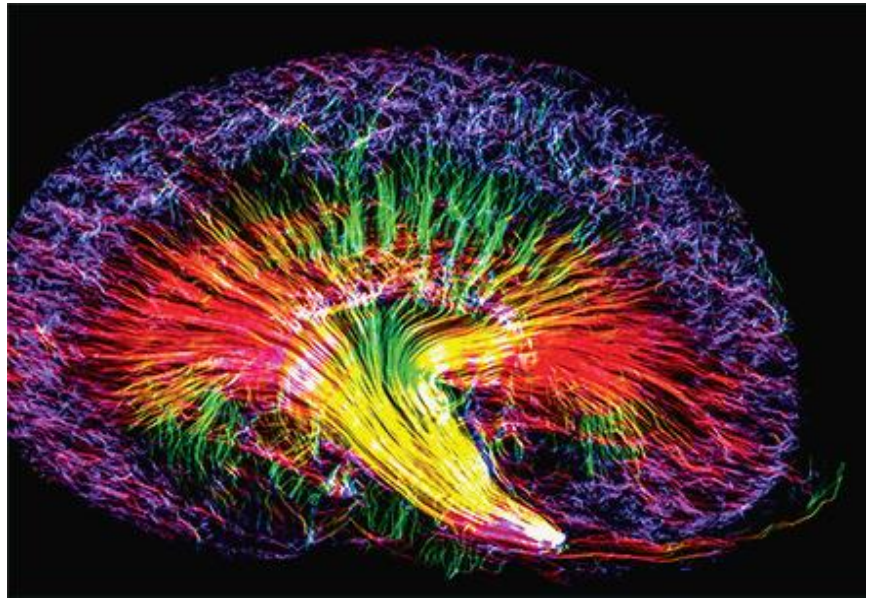
09h50 - 10h20

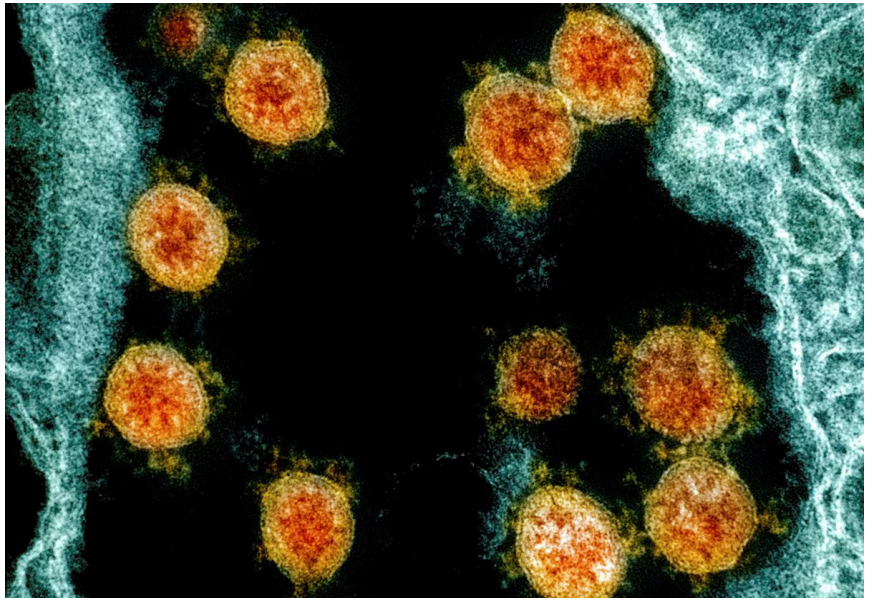
Michel Pohl

Directeur Département de la Science ouverte et directeur de recherche Inserm

Neurobiologiste, directeur de recherches à l'Inserm, j'ai animé pendant plus de 15 ans une équipe travaillant sur les mécanismes moléculaires des douleurs pathologiques et sur le développement des thérapies expérimentales innovantes fondées sur le contrôle ciblé de l'expression de gènes. En 2016 j'ai rejoint la direction générale de l'Inserm en tant que directeur adjoint du Département de l'Information Scientifique et Communication en charge du pôle IST. J'y ai notamment participé à définir la politique de l'Institut en matière de science ouverte que je continue de porter aujourd'hui en tant que directeur du Département de la science ouverte de l'Inserm.

Renforcer l’impact de la recherche en santé





HDH – journée de l’open science en santé

Open Acces – Open Science

Initié dans les années 1990 avec la notion de l'accès libre aux publications scientifiques

Open Access

* 2003 – Déclaration de Berlin; Inserm signataire
* 2006 – HAL; portail institutionnel – HAL-Inserm 

* 2012 – San Francisco declaration on Research assessment (Dora); Inserm signataire (2018)
* 2016 – Loi pour une République numérique
* 2018; 2021 – Plan national pour la SO Inserm 

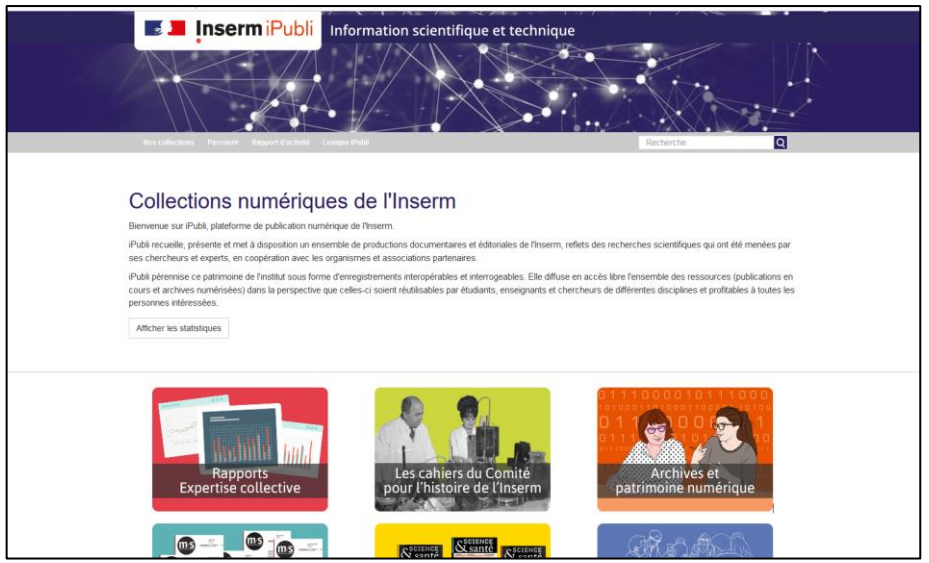
* 2022 – CoARA; Inserm signataire ; Chapitre Français ; Inserm copilote

Open Science

Science Ouverte

Plateformes à disposition du grand public, journalistes, enseignants, chercheurs …

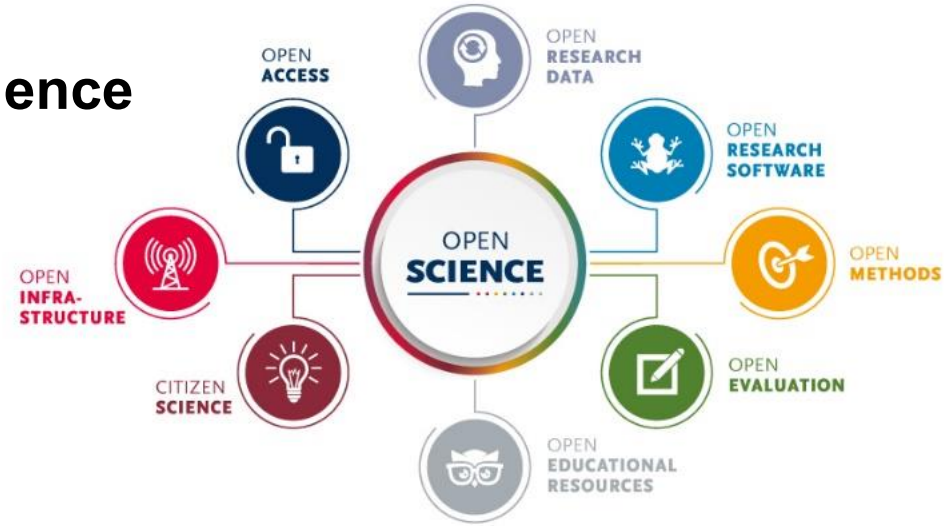
portail institutionnel HAL

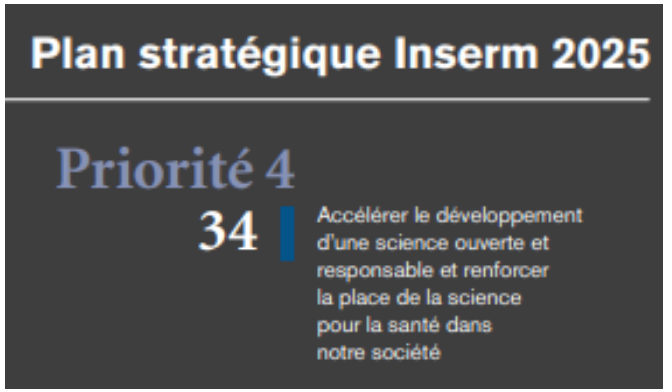


collections éditoriales Inserm

Science Ouverte

Open Acces ➡ Open Science





Science Ouverte

Pourquoi ?

La SO améliore-t-elle l'impact, l'application, les retombées de la recherche ?

Comment ?

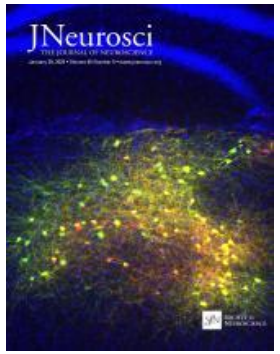
La SO impose une nouvelle façon de concevoir et de pratiquer la recherche

(plan de gestion des données [*PGD*]; exploitation de données existantes; publications et partage)

Partage de l'information scientifique

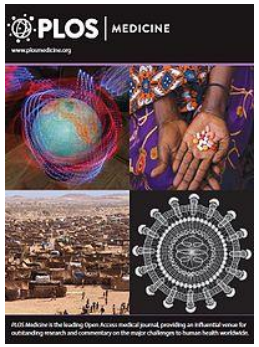
Différents modèles "économiques" de la publication scientifique coexistent :

Modèle classique



- abonnement (lecture)
- copyright transfer

Modèle Gold

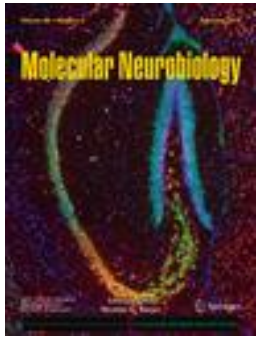


eLIFE

nature COMMUNICATIONS

- auteur payeur (APC) - AO
- licence ouverte CC-BY

Modèle Hybride



SAGE choice

Springer

- abonnement
- APC pour AO ; CC-BY

X

Partage de l'information scientifique

L'information scientifique peut être partagée en dehors de la publication "classique" :

* archives ouvertes « ***Green*** »

  MAA (Loi sur le numérique, 2016)

* preprints :

Cornell University Library arXiv.org

bioRχiv beta THE PREPRINT SERVER FOR BIOLOGY

MedArXiv

VeriXiv

* plateformes: **FC3R Short Notes**

FC·3R centre français

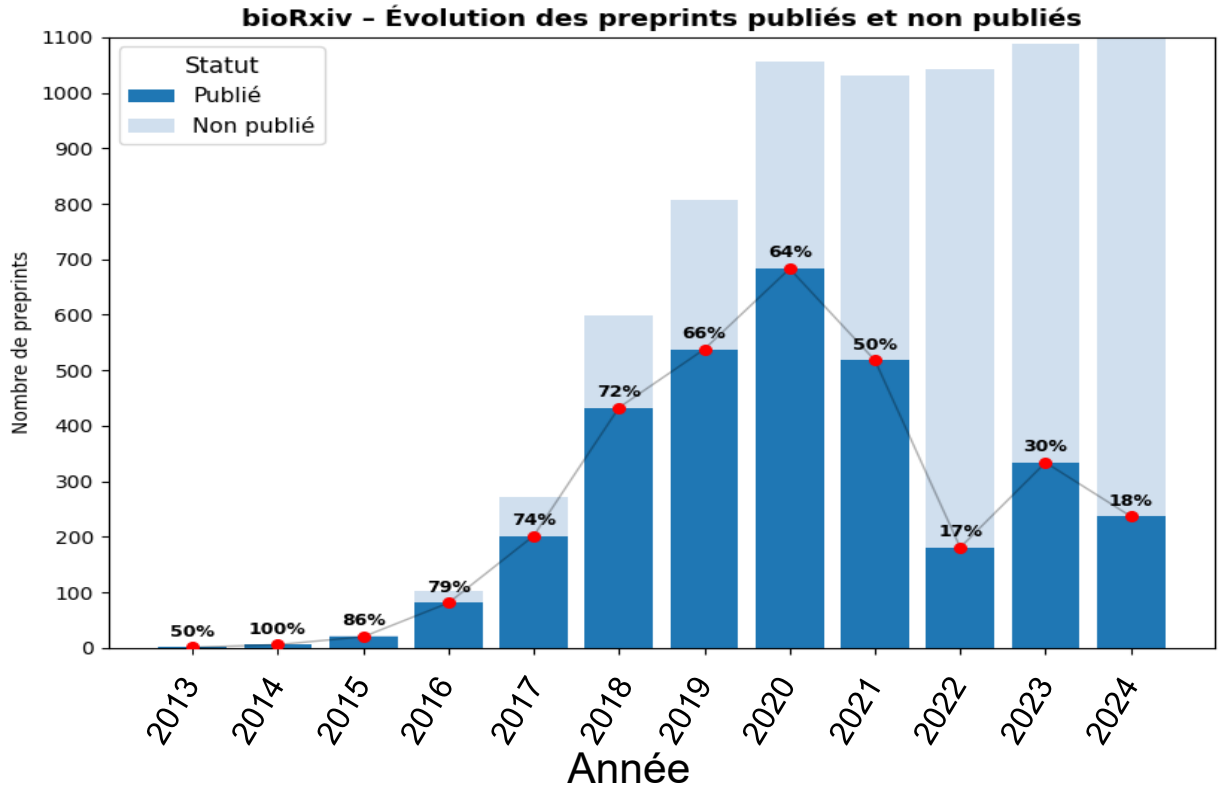
In & Sight

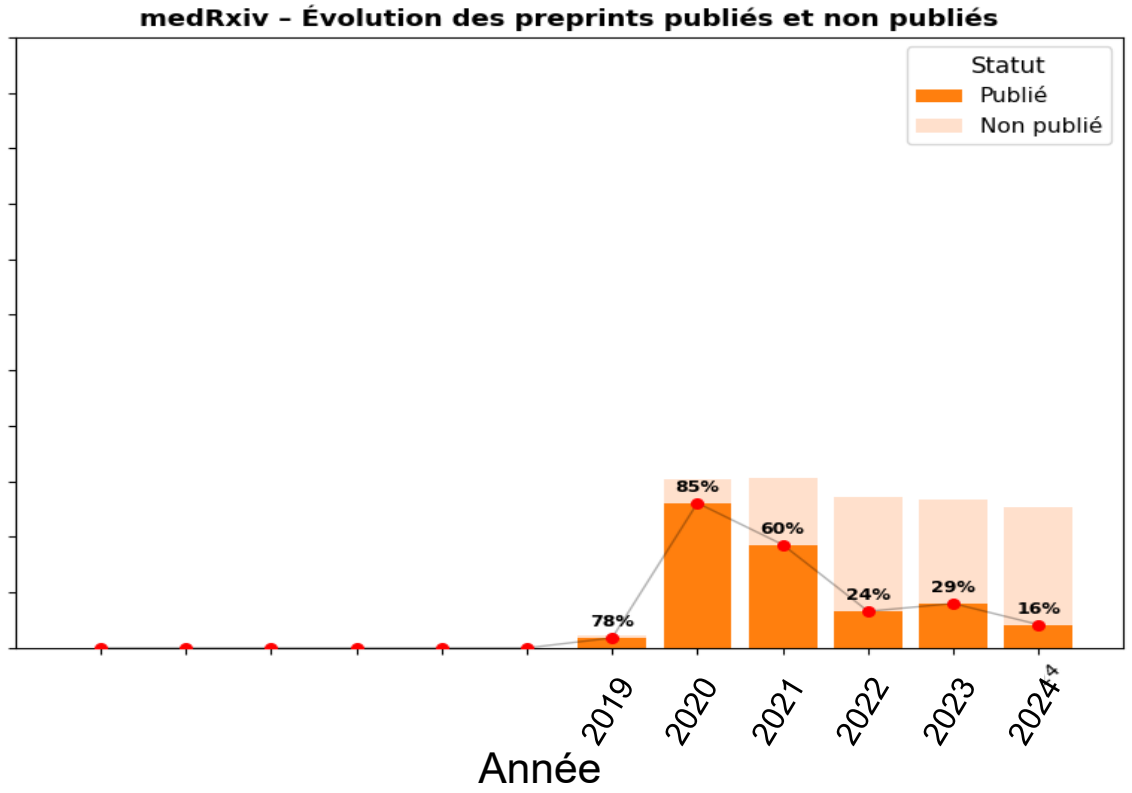
Evolution de dépôts sur plateformes « *preprint* »

* preprints

Gates

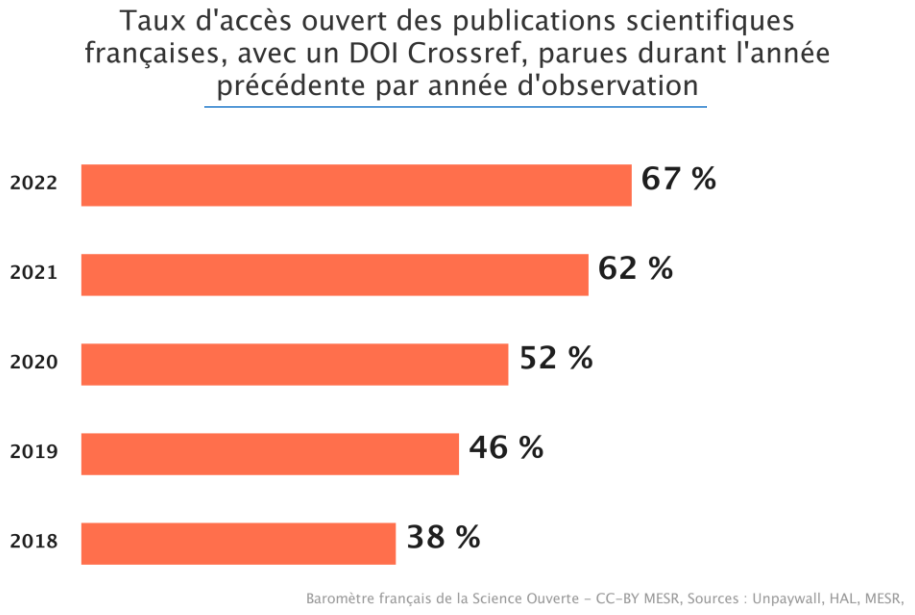
F1000



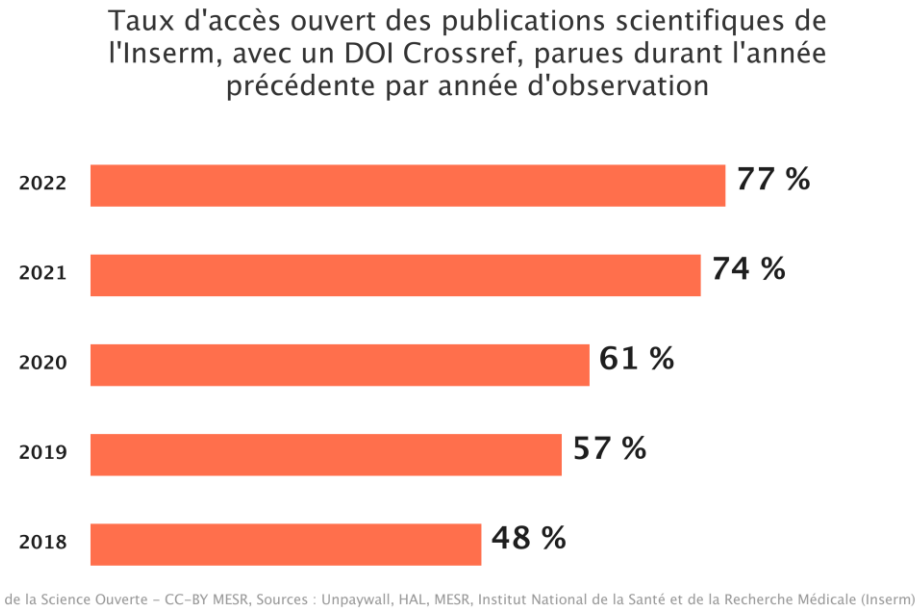


Baromètre science ouverte Inserm (BSO II)

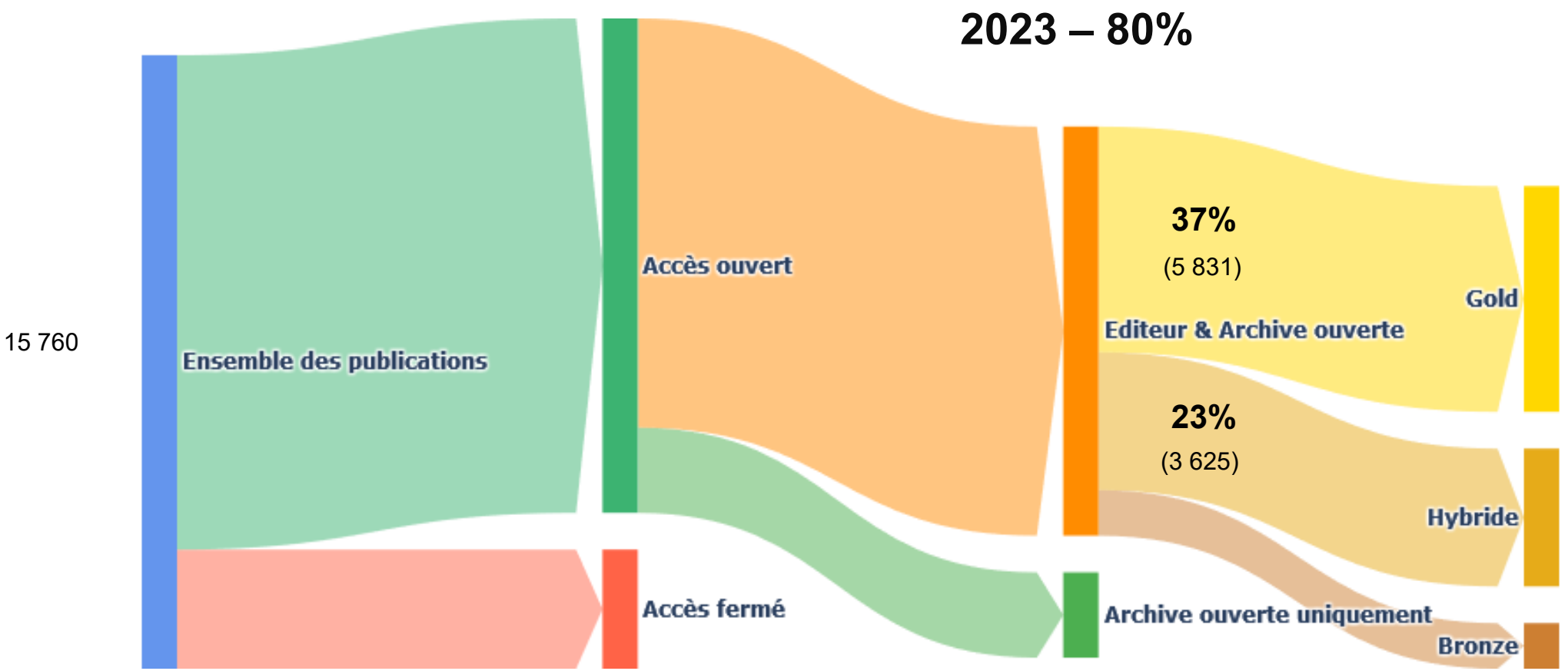
National



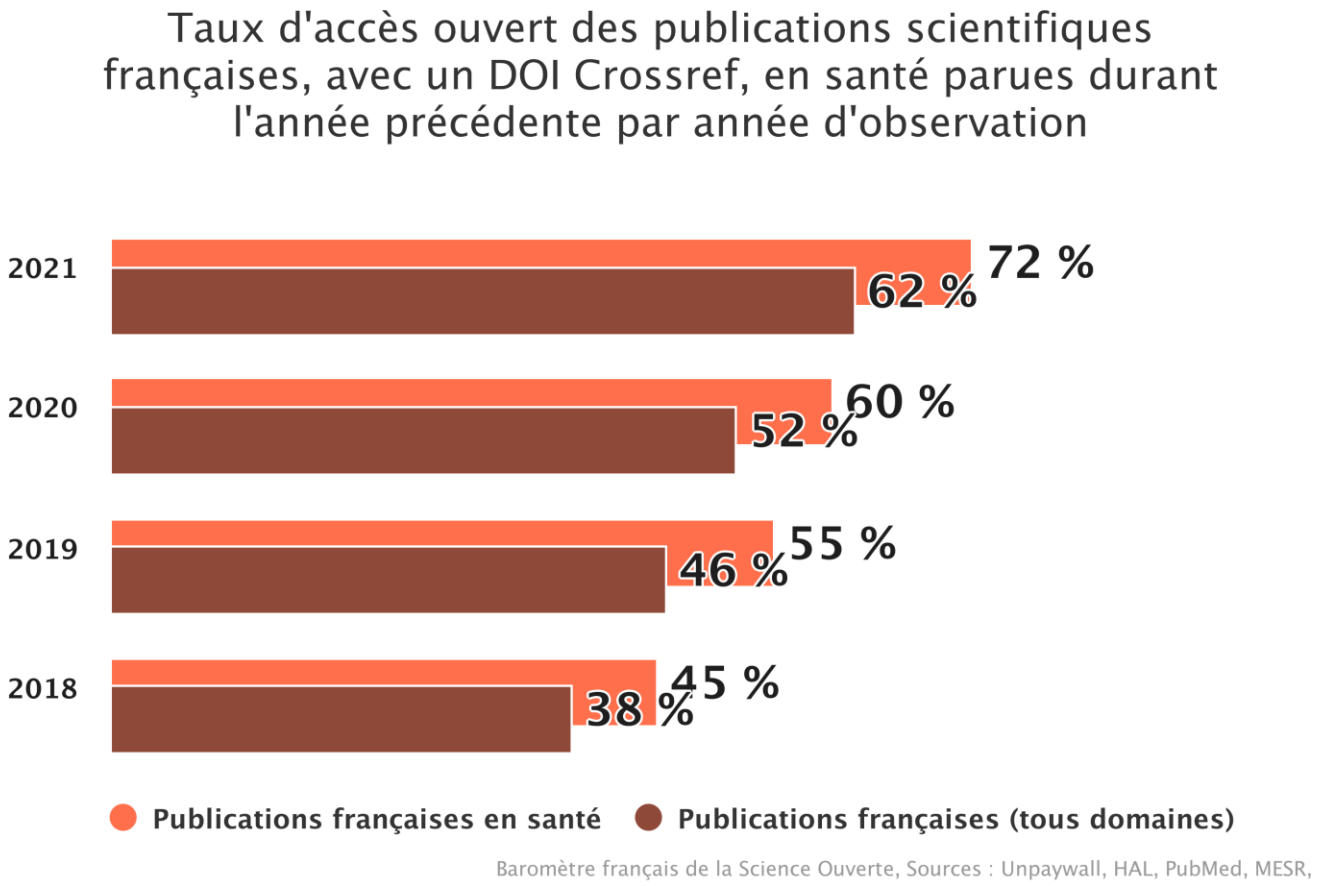
Inserm



Baromètre science ouverte Inserm (BSO II)



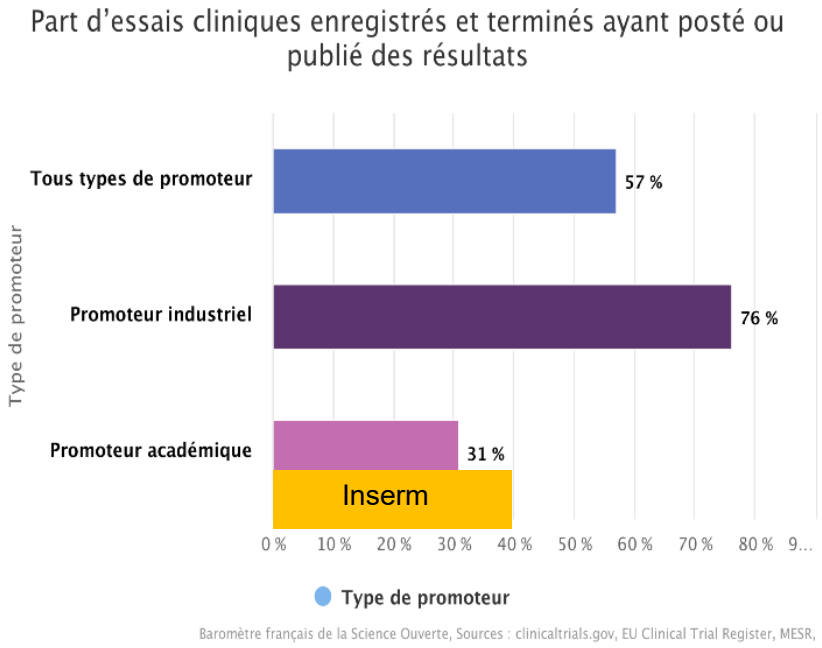
BSO national – publications en santé

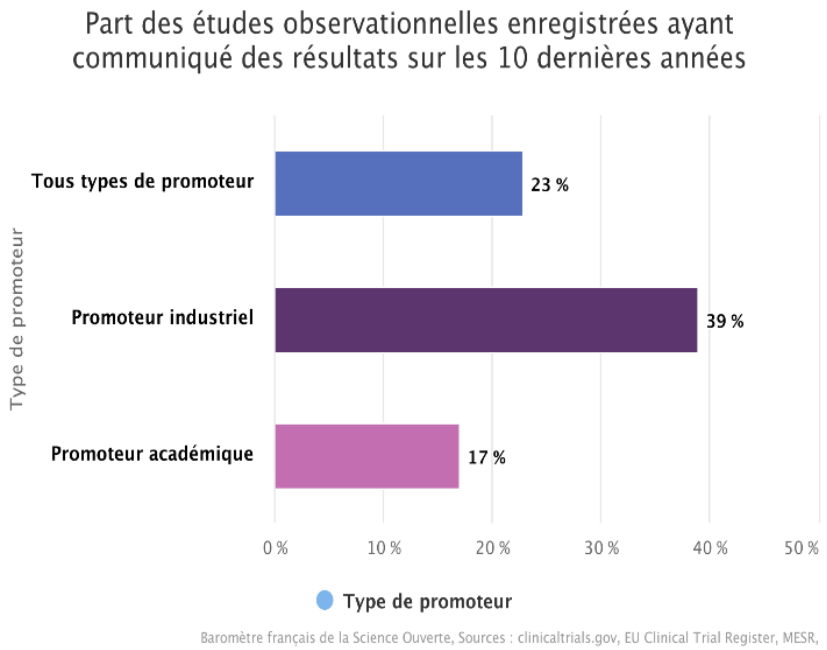


Gestion des données / partage de l'information scientifique

Essais cliniques et études observationnelles

Baromètre national Science Ouverte (BSO II)





GT "Transparence de la recherché en santé"; Philippe Ravaud - mai 2025 (https://www.enseignementsup-recherche.gouv.fr/fr/media/36810)

Recherche clinique – quelles solutions ?

* enregistrer les essais cliniques (**Inserm** ~ 250 essais cliniques, si promoteur 100% enregistrés)

EU Clinical Trials Register

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

~ 44 000

NIH U.S. National Library of Medicine

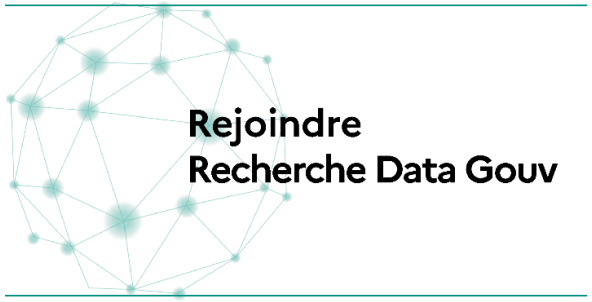
ClinicalTrials.gov

~ 450 000

* publier/poster les résultats des essais cliniques et d'études observationnelles

Recherche clinique – quelles solutions ?

* **L’entrepôt de données Inserm (EDI)**
  Pour préserver, partager et ouvrir les données, les équipes de recherche sont invitées à déposer les données de leurs travaux dans l’Entrepôt de données Inserm (EDI) sur la plateforme ministérielle Recherche Data Gouv.



* **Le « cloud » Inserm (certifié HDS)**

* **Entrepôt de données de Santé (EDS** APHP, HDH, EEDS**)**

Covid-19 : prise de conscience ouverture et partage des données ?

- janvier 2020

Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak

~ 160 organisations

« We call on researchers, journals and funders to ensure that research findings and data relevant to this outbreak are shared rapidly and openly to inform the public health response and help save lives. »

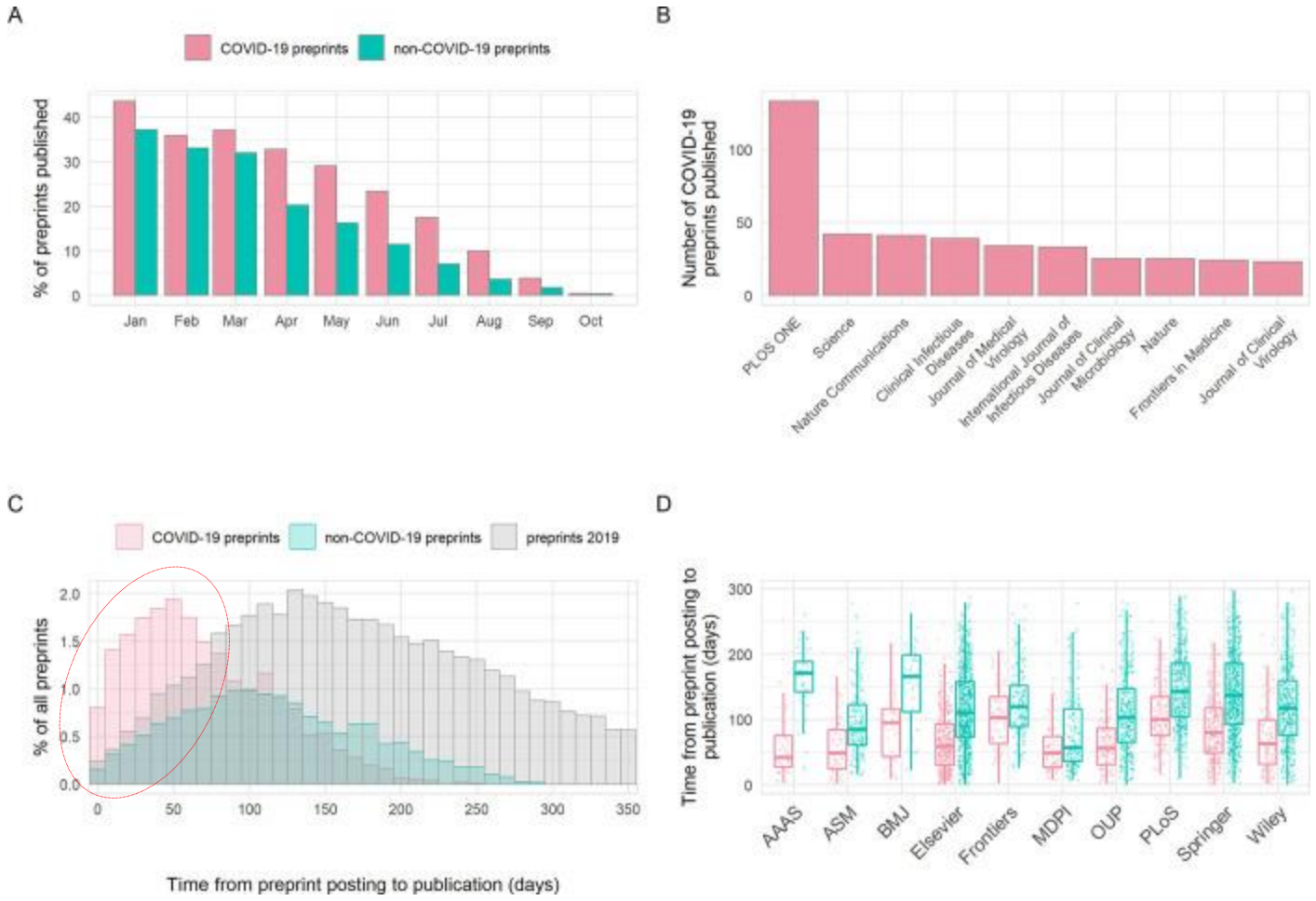
* open or free access to publications
* preprinting
* data sharing

- janvier / mars 2020 « free access » (≠ open access ; durant la pandémie)

~ 30 éditeurs

* facilitating peer review of Covid articles
* speeding up publication times of Covid articles

Covid-19 : prise de conscience ouverture et partage des données ?



Fraser N. et al. (2021) The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. PLOS Biology 19(4): e3000959.

Quelques mots pour la fin…

Scholarly communication in times of crisis:
The response of the scholarly communication system to the COVID-19 pandemic

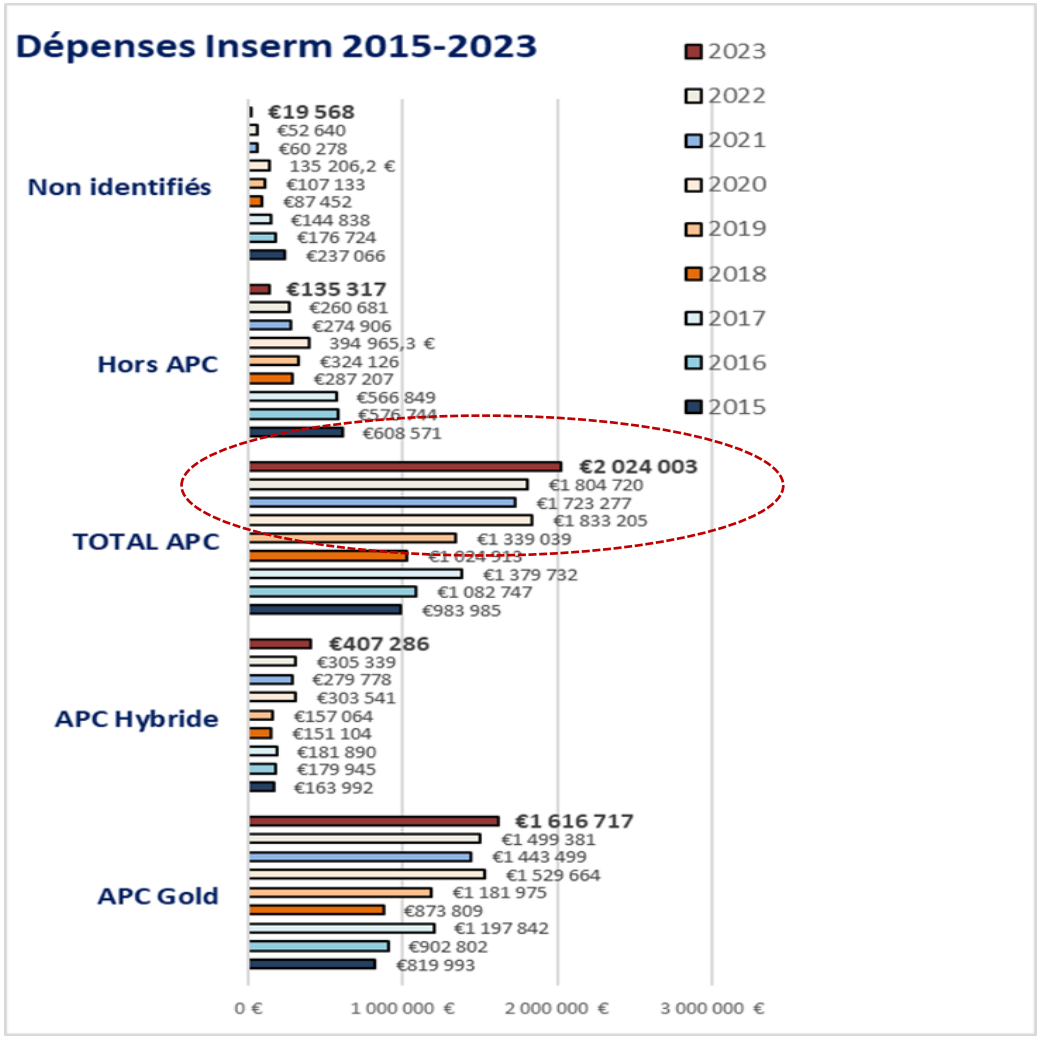
→ ~ 88% articles Covid (review) ouverts et gratuits **++**

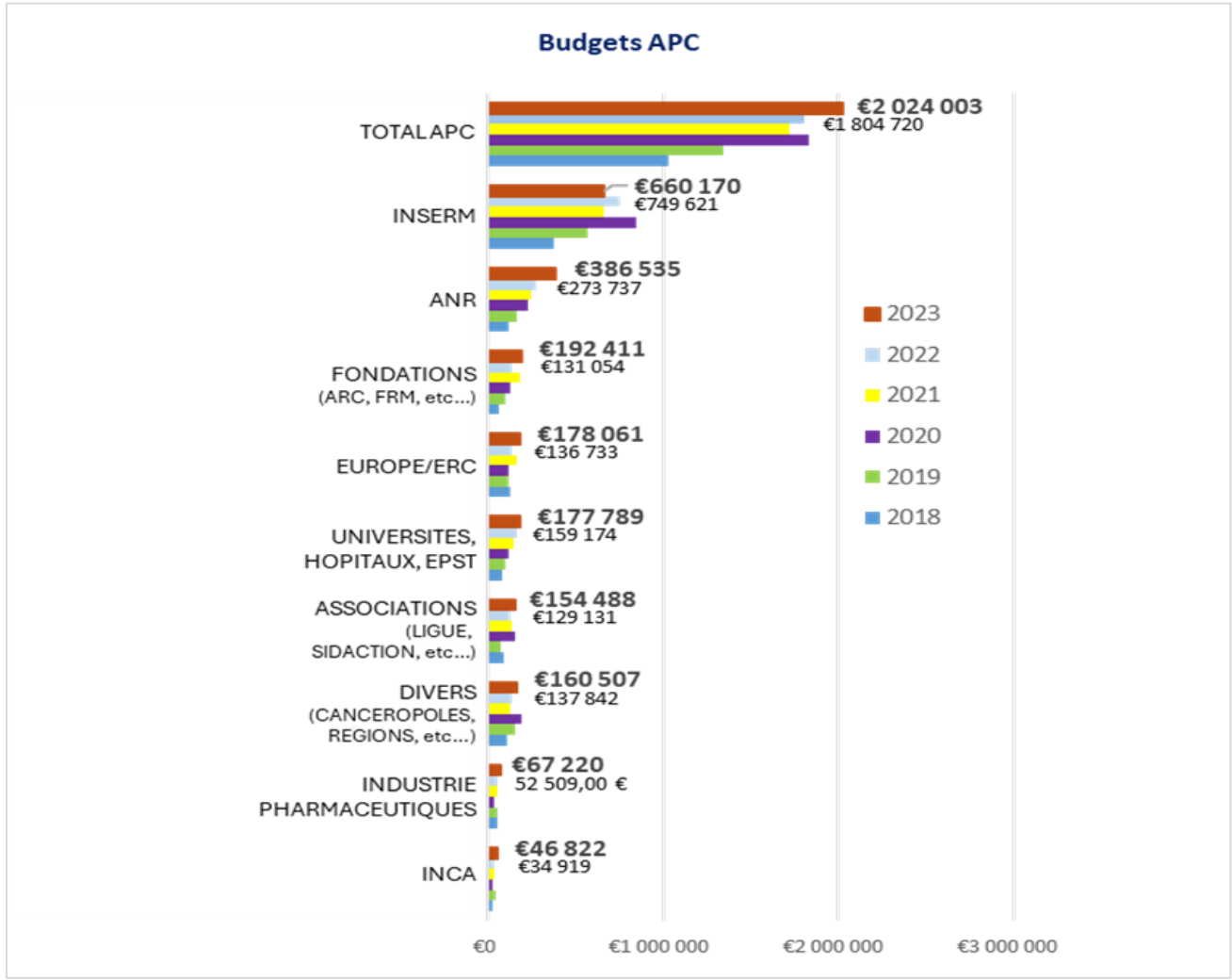
→ ↗ preprints; usage massif et durable **±**

→ → partage des données lent & partiel **Ø**

. . . . keep calm, science hard & **open**

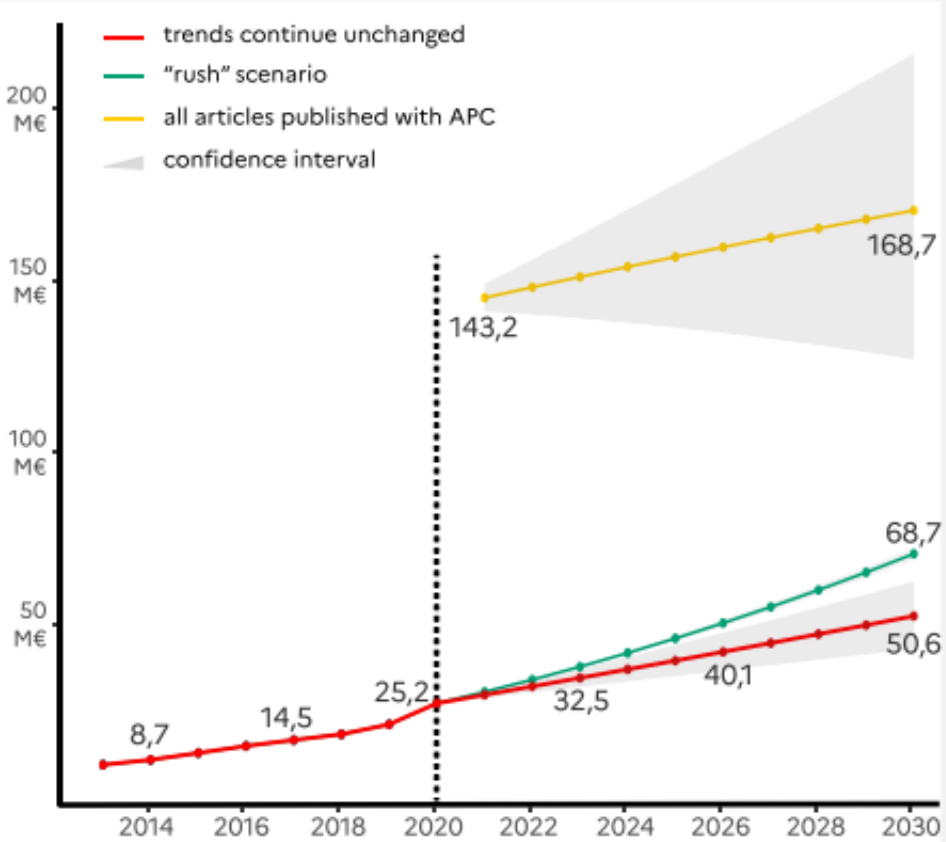
Coûts publications & financements





Etude prospective nationale – évolution coûts APC

2021-2030 During the decade, APC cost will increase again



50 M€ in 2030...
current trend of APC
paid by French institutions

...up to 68 M€...
paid by French institutions
in an alternative scenario

...at most 168 M€
paid by French institutions if all articles
(except 10% diamond) had APC

Subscription expenditures: 87,5 M€ in 2020, going up to 97,5 M€ in 2030

APC: Article processing charges
BSO: Baromètre de la science ouverte
QOAM: Quality of Open Access Marker
HybridOA: open access article in subscription journal
Gold: article in fully open access journal with APC
Diamond: article in fully OA journal without APC

L’AMDAC Solidarités-Santé : encourager et valoriser l’ouverture des données et des codes

HEALTH DATA HUB

L'AMDAC Solidarités-Santé : encourager et valoriser l'ouverture des données et des codes

10h20 - 10h40

Claude Gissot

Directeur de projet AMDAC, Direction de la recherche, des études, de l'évaluation et de la statistique (DREES)

Claude GISSOT est Directeur de projet Administration ministérielle des données, des algorithmes et des codes (AMDAC) à la Direction de la recherche, des études, de l'évaluation et des statistiques du ministère de la Santé sur le champ santé/solidarités. Auparavant, Il a occupé le poste de Directeur de la Stratégie des Études et des Statistiques à la Caisse nationale de l'assurance maladie.

RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

Drees

DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES

L'AMDAC Solidarités-Santé : encourager et valoriser l'ouverture des données et des codes

04/06/2025

1. L'AMDAC : administrateur ministère des données, algorithmes et codes sources

Une politique de la donnée pour la sphère solidarités-santé

Une gouvernance de la donnée renouvelée et enrichie en 2021

2020 : Rapport Bothorel pour une politique publique de la donnée

2021 : Circulaire Premier ministre du 27 avril relative à la politique publique de la donnée

Mai 2021 : nomination du DREES, administrateur ministériel des données, algorithmes et codes sources (AMDAC)

Elaboration de la feuille de route en collaboration avec l'ensemble des directions du ministère et ses principaux opérateurs et établissements

Septembre 2021 : publication de la 1ere feuille de route AMDAC MSS, dans une vision transversale et « complémentaire »

Animation

Fédérer les acteurs de la sphère Solidarités-Santé autour d'actions communes

Communication et sensibilisation

Communiquer sur les données, sensibiliser les agents aux enjeux et encourager la mise en place de bonnes pratiques

Gouvernance

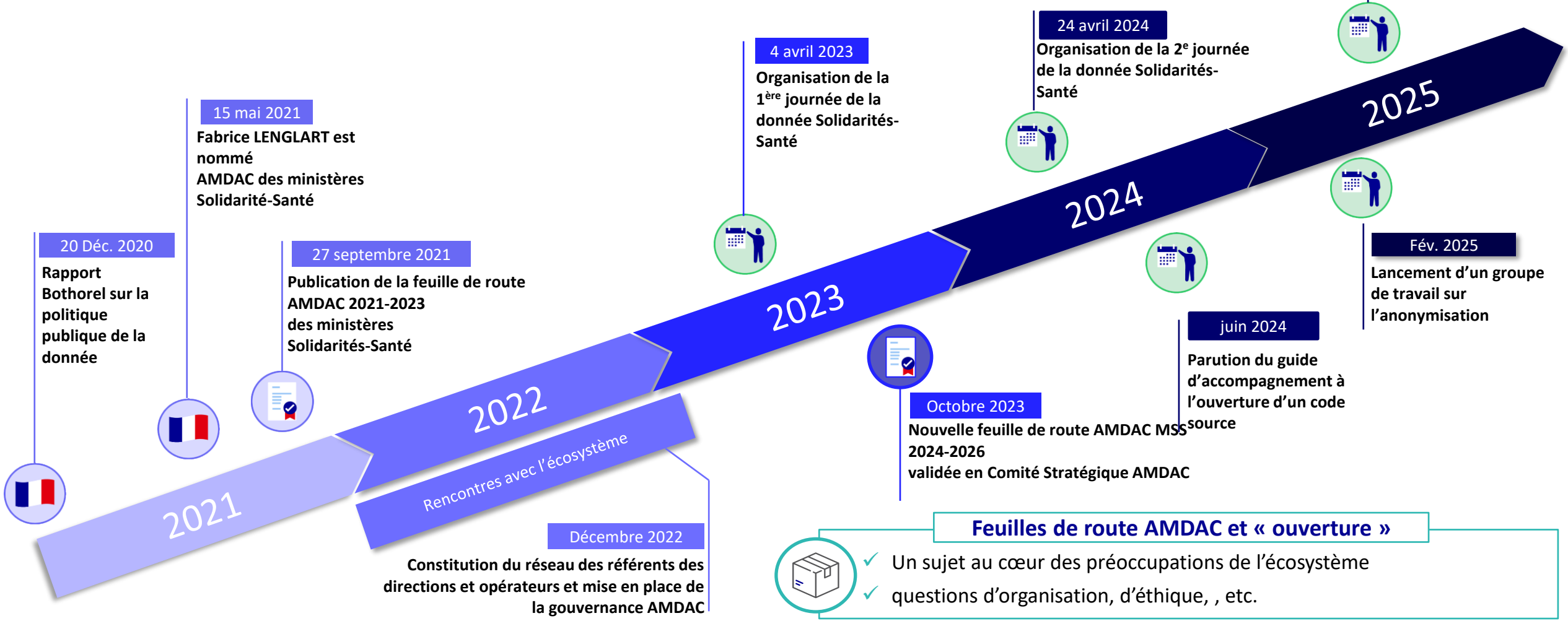
Développer la connaissance du patrimoine et la gouvernance autour des sujets données et codes

Circulation des données

Faciliter l'exploitation, le partage et l'ouverture des données et des codes

La démarche AMDAC : dates clés

Gouvernance, Animation, Sensibilisation, Circulation des données



Feuilles de route AMDAC et « ouverture »

- ✓ Un sujet au cœur des préoccupations de l'écosystème
- ✓ questions d'organisation, d'éthique, , etc.

Une nouvelle feuille de route 2024-2026

Gouvernance, Animation, Sensibilisation, Circulation des données

Les enjeux :

❑ Élaboration avec le réseau des référents, travail collectif

❑ Continuité des travaux dans une logique de transversalité et de partage d'expériences :

- Groupes de travail
- Dynamique du plan d'ouverture
- Communication et événements orientés « données »
- Renforcement du rôle des référents et de la gouvernance

❑ Ancrage métier plus fort, matérialisé par des projets emblématiques

❑ Développement du réseau AMDAC : DGS, DGOS, DSS, DGCS, DNS, DNUM, HAS, CNAM, CNAF, CNAV, ATIH, SpF, ANSM, EFS, ANS, ABM, etc.

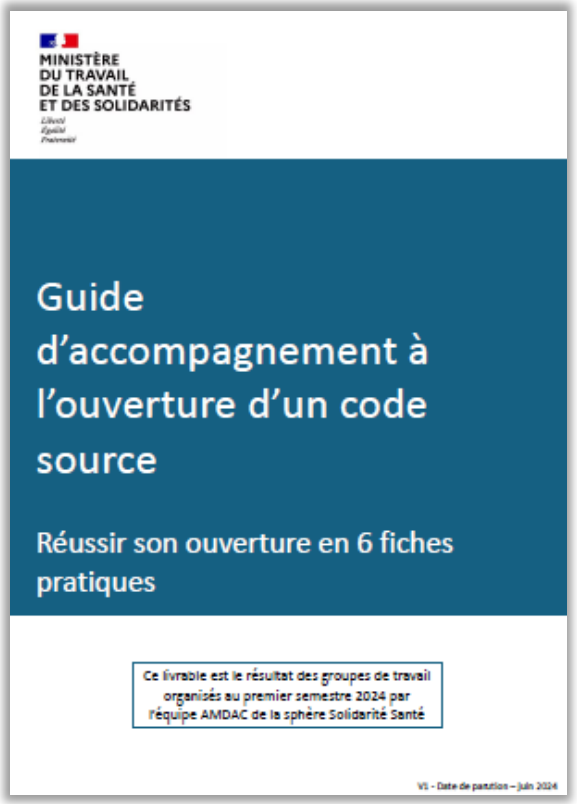
Projets emblématiques

- Réduire le non-recours aux prestations sociales en utilisant la donnée comme levier
- Créer un espace commun de données de "solidarités"
- Créer un partenariat avec Ecolab (MTES) et identifier des cas d'usages d'échange de données pour la santé et pour les solidarités
- Simplifier la collecte des données de santé (notamment ETS)
- Mieux maitriser le capital de données, d'algorithmes, de codes sources et d'applicatifs pour favoriser la mise en place de la gouvernance de la donnée

2. Les actions de l'AMDAC en faveur de l'ouverture

Les productions du réseau AMDAC

Un guide pratique pour accompagner les projets d'ouvertures de codes sources



Télécharger le guide

La méthode

- Contributeurs volontaires venant du **HDH** (copilote du GT), de la **DINUM, l'ATIH, SpF, la CNAM, la CNAF, la DREES et la CNAV**
- Une démarche condensée en 6 mois : **1 atelier de cadrage et 3 ateliers de travail** autour des enjeux, des **licences**, des **plateformes** et des relations avec la **communauté**
- Livrable **construit au fil des ateliers** et **validé le 4 juin 2024**

Le livrable

- Destiné aux équipes confrontées à leur première opération d'ouverture de code source
- **Guide synthétique de 30 pages fondé sur 6 fiches pratiques encadrant la démarche d'ouverture de bout en bout**
  - Préparer l'ouverture (objectifs, prérequis juridiques, licence, plateforme, travail sur le code) et les suites de l'ouverture
- Une section « boîte à outils » et compléments

Les productions du réseau AMDAC

Un groupe de travail expert pour faciliter les démarches de pseudonymisation de données

Livrable à paraitre en septembre 2025

La méthode

- Contributeurs volontaires
- Une démarche condensée en 6 mois : **1 atelier de cadrage et 3 ateliers de travail** autour des enjeux
- Livrable en cours de construction **au fil des ateliers**

Objectif du livrable

- **Destiné aux équipes confrontées à des opérations d'anonymisation de données, par exemple en vue de leur ouverture en open data**
- Guide synthétique, clarifiant la différence entre anonymisation et pseudonymisation
- Approche pratiques mettant en avant des méthodes employées d'anonymisation

Les productions du réseau AMDAC

Un livret référence l'ensemble des productions AMDAC et de son réseau

Télécharger le livret

L'approche

- Un document qui offre un panorama des productions de l'AMDAC et de son réseau
- Une mise à jour régulière au fil des nouvelles productions
- Un document pivot pour découvrir les données et codes sources des acteurs du domaine santé solidarité

Le livrable

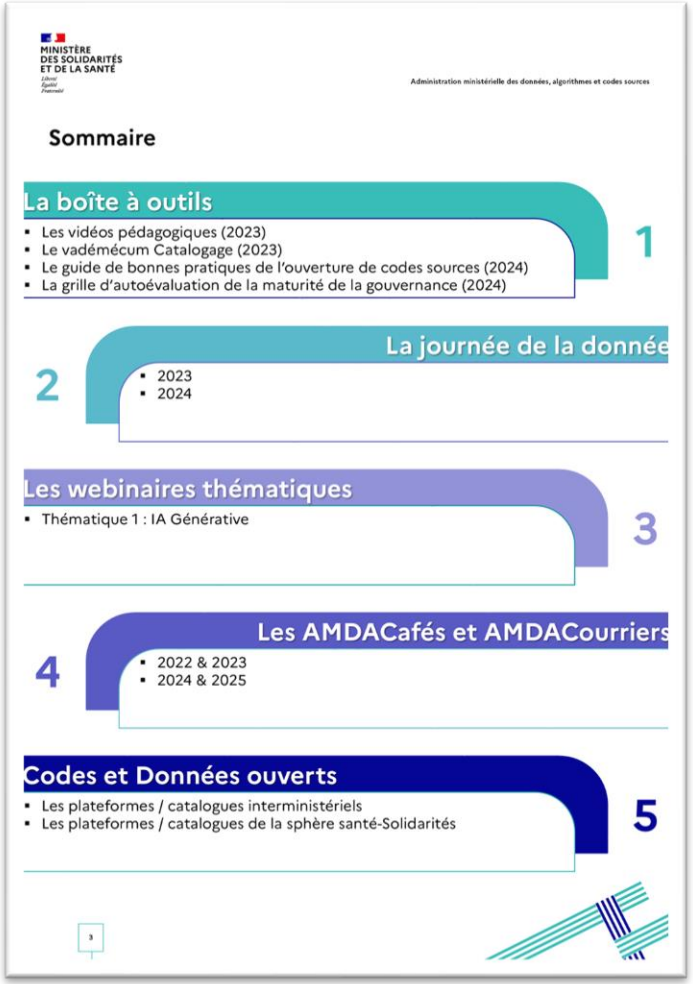
Liens vers des ressources utiles

- Boite à outils (vidéos pédagogiques, guides, vadémécum, etc.)
- Replays (journées de la donnée, webinaires thématiques, amdacafés, amdacourriers, etc.)

Liens vers les plateformes et catalogues de la sphère santé-solidarités

- Plateformes open data, catalogues de données, plateformes de partage de code, etc.

Réalisation d’un livret centralisant les productions du réseau AMDAC MSS



- Valoriser les productions du réseau AMDAC (livrables des groupes de travail, actions d’acculturation, ouvertures des données et codes)
- Diffusion publique
- Mise à jour mensuelle et partage lors de l’AMDACourrier

Réalisation d'un livret centralisant les productions du réseau AMDAC MSS

Focus sur les ouvertures : une approche par les plateformes

Codes et données ouverts

Les ouvertures de données de la sphère ministérielle Santé et Solidarités

Annuaire santé : accès au référentiel RPPS
Plateforme | Données | Cartographie

Le référentiel FINESS
Catalogue | Données

Caf Data
Plateforme | Données | Visualisation | Cartographie

Data.drees
Plateforme | Données

Odin : outil de visualisation
Plateforme | Données | Visualisation | Cartographie

Répertoire des spécialités pharmaceutiques
Liste | Données

Data.ameli
Plateforme | Données | Visualisation

Liste des données ouvertes
Liste

SCORE-Santé
Plateforme | Données | Visualisation

Atlasanté
Catalogue | Données | Cartographie

Datalogue
Catalogue

Data.cnav
Catalogue | Données

Plateforme open data
Plateforme | Données | Visualisation

Liste des données ouvertes
Liste

ScanSanté
Plateforme | Visualisation | Cartographie | Restreint

Les chiffres clés de l'hospitalisation
Plateforme | Données | Visualisation

Portrait des territoires
Plateforme | Données | Visualisation

Geodes
Plateforme | Données | Visualisation | Cartographie

Conduite d'un "groupe de réflexion" sur la gouvernance des données

Un lieu d'échange sur les problématiques au cœur des données

L'équipe AMDAC à lancé en 2025 un groupe de réflexion et d'échanges abordant différents thèmes de la gouvernance des données :

PARTICIPANTS

- Responsables data des opérateurs (sur la base du volontariat)
- Référents
- Autres participants à définir

ORGANISATION

- Organisé et animé par l'équipe AMDAC
- Format webinaire
- Distanciel
- Calendrier des thèmes et durée à définir

OBJECTIFS

- Partager les pratiques et les difficultés
- Identifier des leviers d'action
- Aider à la déclinaison de la gouvernance
- Appuyer la réflexion stratégique

CONTENU PROPOSÉ

- RETEX et témoignages, y compris externes
- Ateliers
- Thèmes abordés issus des axes de progrès identifiés

FRÉQUENCE

- Trimestrielle

LIVRABLES

- Supports des présentations
- Compte-rendus (ad hoc)

Les actions de communication/sensibilisation menées

Des dispositifs variés à destination des acteurs de la sphère Santé-Solidarité

AMDACafés

Présentation

- Webinaire mensuel
- 30 minutes en distanciel
- Actualités AMDAC et interministérielle
- Présentation d'un projet de la sphère MSS

Chiffres clés

- 20~50 participants réguliers
- +30 sessions réalisées depuis 2022

Journée de la donnée

Présentation

- Évènement annuel
- Journée complète (hybride)
- Conférences, tables rondes, et webconf thématiques
- Public ciblé : agents de la sphère MSS

Chiffres clés

- 3ème édition le 03/06/2024
- 100~150 participants

Webinaire thématique

Présentation

- En fonction des opportunités et thématiques
- ½ journée en distanciel
- Retours d'expérience
- Atelier collaboratif

Chiffres clés

- Première thématique : IAGén
- 50+ participants
- 3 sessions menées en 2024 et 2025

AMDACourrier

Présentation

- Synthèse des actualités de l'AMDAC (avancement des projets, actions à venir, etc.)
- Format courriel (plutôt que pdf)
- 1 occurrence à la rentrée septembre 2024
- Envoyé au réseau de référent
- Format allégé

RÉPUBLIQUE FRANÇAISE

Liberté
Égalité
Fraternité

Drees

DIRECTION DE LA RECHERCHE, DES ÉTUDES, DE L'ÉVALUATION ET DES STATISTIQUES

Contacts AMDAC MSS :
amd@sante.gouv.fr

Direction de la Recherche, des Études, de l'Évaluation et des Statistiques

HEALTH DATA HUB

PAUSE

10h40 - 10h55

Journée de l’open science en santé - Programme

10h40 - 10h55
Pause

10h55 - 11h15 **La Cartographie des pathologies : un outil désormais ouvert au service de la recherche et de la santé publique**, Antoine Rachas (CNAM)

11h15 - 11h45 **Ouvrir l’accès aux ressources scientifiques : les initiatives de l’INRAE en faveur de l’ouverture de la science**, Odile Hologne (INRAE)

11h45 - 12h30 **Annonce des lauréats de la 8e vague de l’AMI de la Bibliothèque Ouverte d’Algorithmes en Santé (BOAS)**, Maxime Caillet (HDH)

12h30 - 14h00
Pause

La Cartographie des pathologies : un outil désormais ouvert au service de la recherche et de la santé publique

HEALTH DATA HUB

La Cartographie des pathologies : un outil désormais ouvert au service de la recherche et de la santé publique

10h55 - 11h15

Antoine Rachas

Médecin épidémiologiste et Responsable adjoint du Département des Etudes sur les Pathologies et les Patients à la CNAM

Responsable adjoint du Département des Etudes sur les Pathologies et les Patients à la Cnam, Antoine Rachas (MD, PhD) est spécialisé dans la méthodologie des études en santé publique, notamment l'épidémiologie et les recherches basées sur les bases de données médico-administratives. Il est l'un des médecins référents de la Cartographie des pathologies et des dépenses, un outil analytique qui identifie et évalue les dépenses de santé associées à diverses pathologies en France. Il est par ailleurs membre du réseau ReDSiam et des comités scientifiques et éthiques de la plateforme des données de cancer et de l'entrepôt des données de santé de l'APHP.

La Cartographie des pathologies : un outil désormais ouvert au service de la recherche et de la santé publique

Sécurité sociale
L'Assurance Maladie
Caisse Nationale

Journée Open Science en Santé – 4 juin 2025

Equipe cartographie (Cnam – DSES) :
Pauline BARTHELEMY, Victor BRET, Panayotis CONSTANTINOU, Gonzague DEBEUGNY, Pierre DENIS, Dimitri LASTIER, Thomas LESUFFLEUR, Corinne METTE, Muriel NICOLAS, Antoine RACHAS, Martine THOMAS

Pour mieux comprendre les dépenses de santé, il faut les aborder sous une logique médicale.

Résilience du système de santé *OCDE Health at a Glance: Europe 2022*

- Identifier les priorités de santé publique
- Améliorer son efficience
- **Comprendre les dépenses de santé**

Logique médicale ➔ *Economic burden of disease*

La Cartographie décrit les dépenses par pathologie.

General cost-of-illness study (Rosen 2016)

Logique comptable

Consultations

Hospitalisations

Médicaments

…

Logique médicale

Diabète

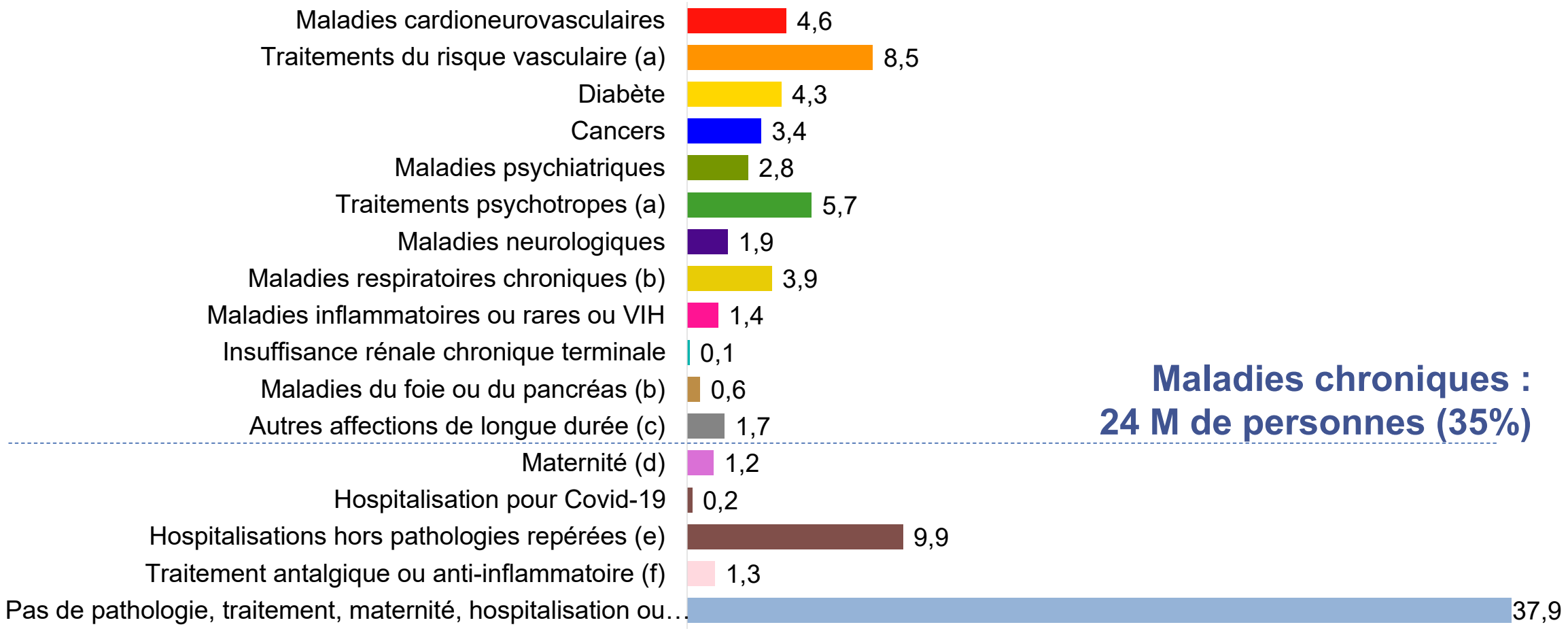
Cancer colorectal

Maternité

…

En 2022, 24 millions de personnes ont consommé des soins en lien avec une pathologie chronique.

Effectifs de personnes par pathologie, traitement chronique et épisode de soins en 2022 (**millions**) (N = 68,7 millions)



Maladies chroniques : 24 M de personnes (35%)

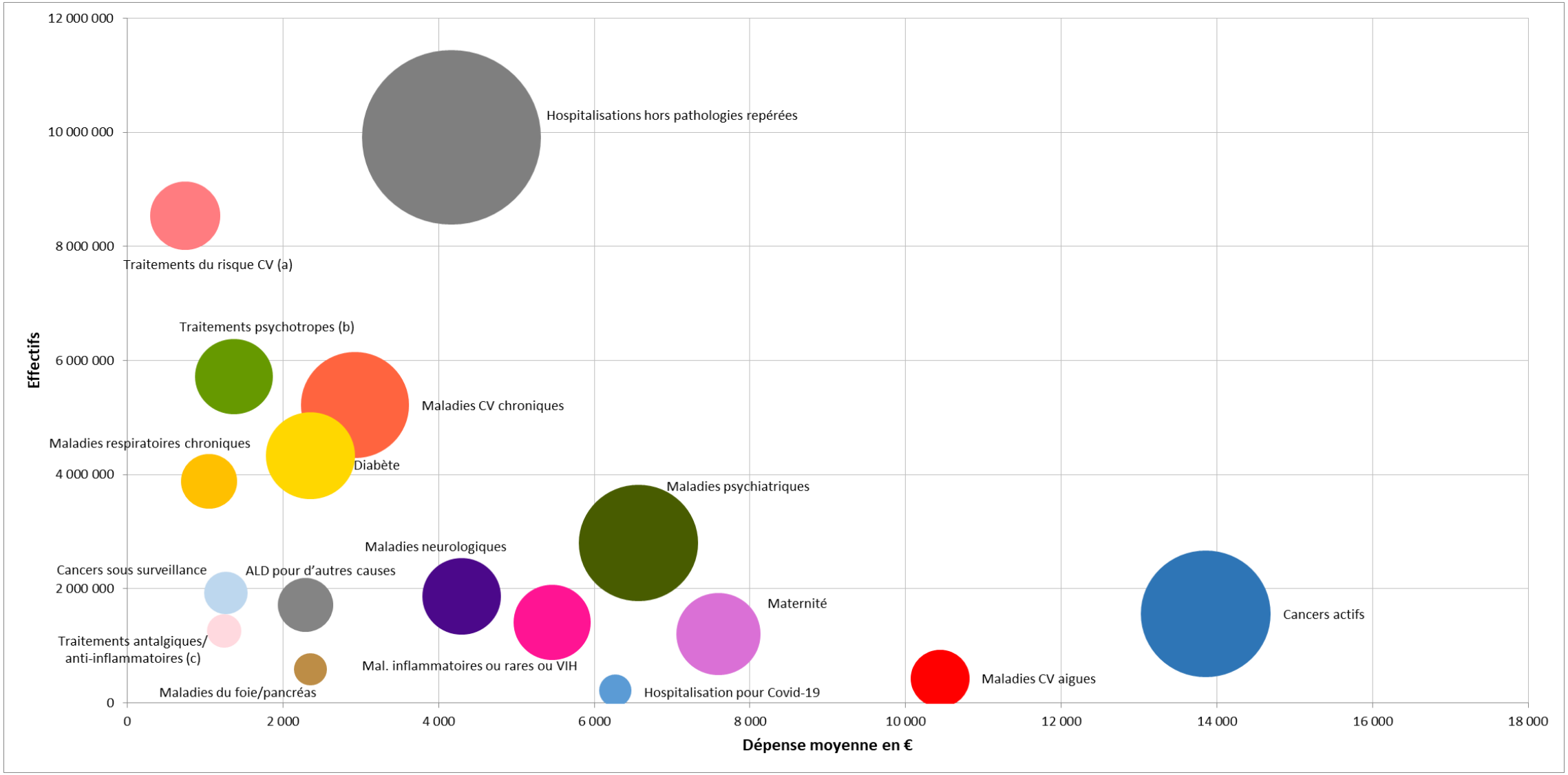
(a) hors pathologies
(b) hors mucoviscidose
(c) dont 31 et 32
(d) avec ou sans pathologie
(e) avec ou sans pathologie, traitements ou maternité
(f) hors pathologies, traitements, maternité ou hospitalisations

Champ : tous régimes - France entière
Source : Cnam (cartographie - version de juillet 2024)

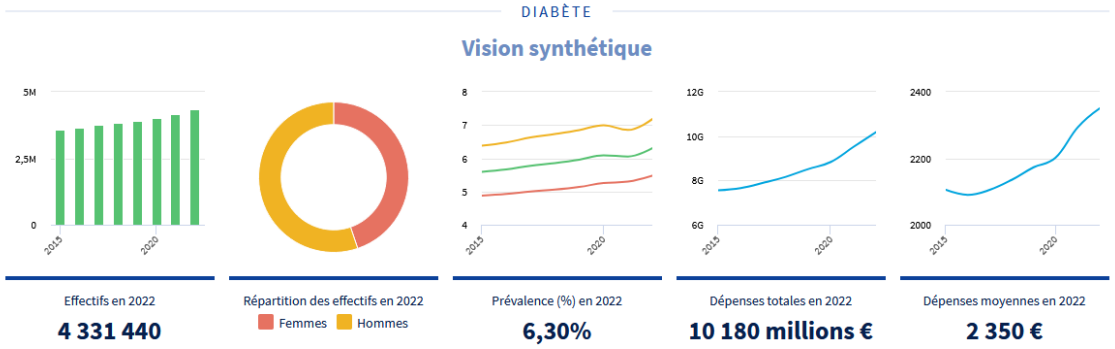
l'Assurance Maladie
Agir ensemble, protéger chacun

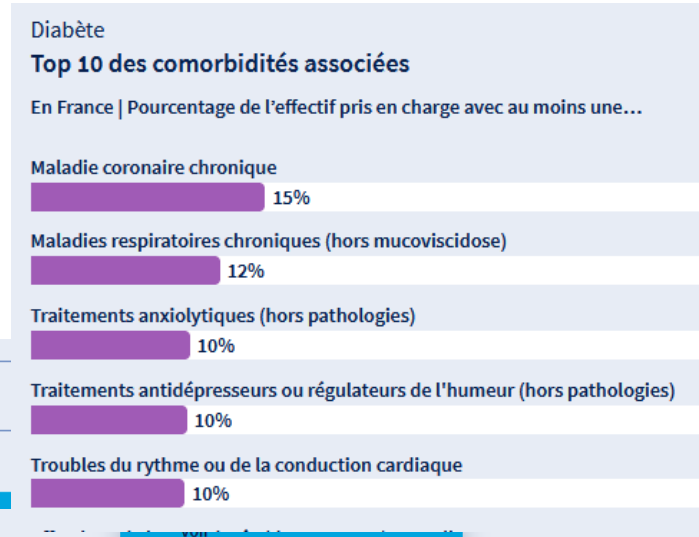
Selon les pathologies, les dépenses s'expliquent plutôt par le nombre de patients ou par la dépense moyenne par patient.

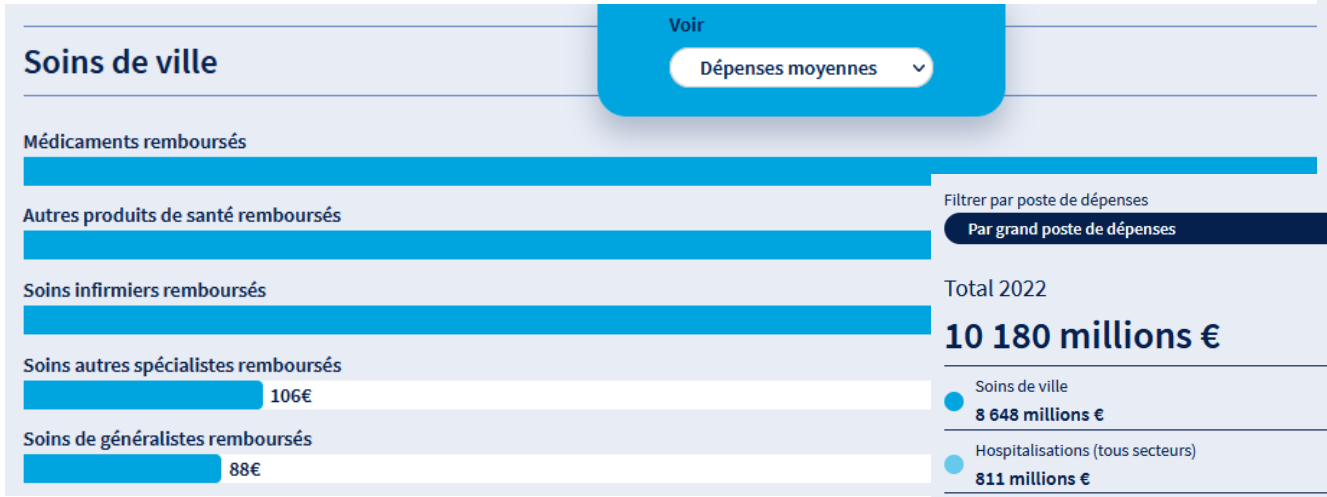
Effectifs, dépenses moyennes par patient et dépenses totales en 2022 pour chaque catégorie de pathologies, traitements chroniques ou épisode de soins

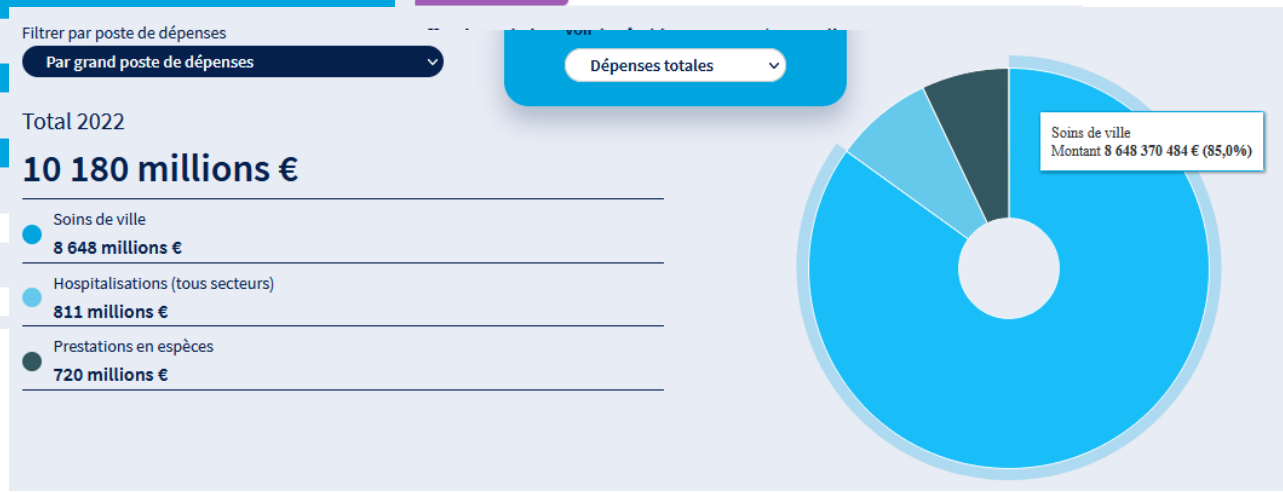


Pour en savoir plus, rendez-vous sur Data pathologies !









https://data.ameli.fr/pages/data-pathologies/

l'Assurance Maladie
Agir ensemble, protéger chacun

Comment ça marche ?

Authentic picture. Designed by Freepik

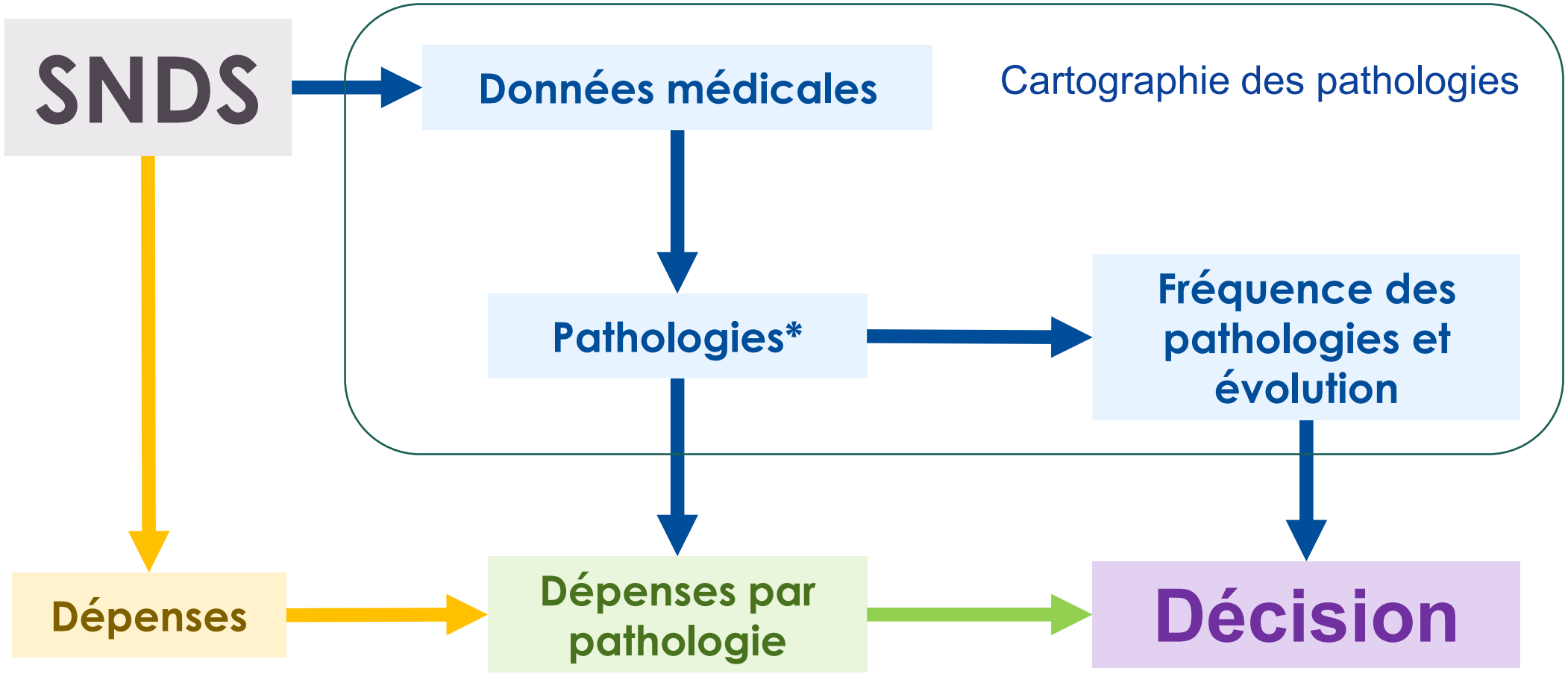
Qui est concerné et quelles dépenses sont analysées ?

- ≥ 1 soin remboursé dans l'année
- Tous les régimes d'assurance maladie
- Exclusions spécifiques (aide médicale d'Etat, incohérences, etc.)

- Dépenses remboursées individualisables
- Soins de ville, hospitalisations, prestations en espèces
- 26 postes de dépenses

Version G12 : Années 2015-2023

3 grandes étapes : repérage des pathologies, extraction des dépenses et affectation des dépenses aux pathologies.



*pathologies, traitements chroniques ou épisodes de soins

275 algorithmes médicaux, dont 61 pour analyser les dépenses

Maladies cardioneurovasculaires
Traitements du risque vasculaire
Diabète
Cancers
Maladies psychiatriques
Traitements psychotropes
Maladies neurologiques
Maladies respiratoires chroniques
Maladies inflammatoires ou rares ou VIH
Insuffisance rénale chronique terminale
Maladies du foie ou du pancréas
Autres affections de longue durée
Hospitalisation pour COVID-19
Maternité
Hospitalisations hors pathologies repérées
Traitement antalgique/anti-inflammatoire
Pas de pathologie repérée

17 catégories

275 algorithmes médicaux, dont 61 pour analyser les dépenses

- Maladies cardioneurovasculaires
- Traitements du risque vasculaire
- Diabète
- **Cancers**
  - Cancer du sein actif
  - Cancer du sein en surveillance
  - Cancer colorectal actif
  - Cancer colorectal en surveillance
  - Cancer bronchopulm. actif
  - Cancer bronchopulm. en surveillance
  - Cancer de la prostate actif
  - Cancer de la prostate en surveillance
  - Autres cancers actifs
  - Autres cancers en surveillance
- Maladies psychiatriques
- Traitements psychotropes
- Maladies neurologiques
- Maladies respiratoires chroniques
- Maladies inflammatoires ou rares ou VIH
- Insuffisance rénale chronique terminale
- Maladies du foie ou du pancréas
- Autres affections de longue durée
- Hospitalisation pour COVID-19
- Maternité
- Hospitalisations hors pathologies repérées
- Traitement antalgique/anti-inflammatoire
- Pas de pathologie repérée

10 algorithmes « cancers » (5 localisations) pour analyser les dépenses

275 algorithmes médicaux, dont 61 pour analyser les dépenses

- Maladies cardioneurovasculaires
- Traitements du risque vasculaire
- Diabète
- Cancers
- Maladies psychiatriques
- Traitements psychotropes
- Maladies neurologiques
- Maladies respiratoires chroniques
- Maladies inflammatoires ou rares ou VIH
- Insuffisance rénale chronique terminale
- Maladies du foie ou du pancréas
- Autres affections de longue durée
- Hospitalisation pour COVID-19
- Maternité
- Hospitalisations hors pathologies repérées
- Traitement antalgique/anti-inflammatoire
- Pas de pathologie repérée

- Cancer du sein actif
- Cancer du sein en surveillance
- Cancer colorectal actif
- Cancer colorectal en surveillance
- Cancer bronchopulm. actif
- Cancer bronchopulm. en surveillance
- Cancer de la prostate actif
- Cancer de la prostate en surveillance
- Autres cancers actifs
- Autres cancers en surveillance

- Col de l'utérus
- Estomac, œsophage
- Voies aérodig. supérieures
- Lymphome hodgkinien
- Lymphome non hodgkinien
- Mélanome de la peau
- Ovaire
- Pancréas
- Rein
- Os
- Thyroïde
- Foie et voies biliaires
- Vessie
- Système nerveux
- Tissus mous
- …

114 algorithmes « cancers »
(30 localisations) au total

Plusieurs sources de données médicales, avec un recul jusqu'à 5 ans

Nombre d'années utilisées pour chaque algorithme, selon la source de données (extrait)

Pathologies (états de santé et traitements)	ALD	Médicaments	PMSI MCO				PMSI RIM-P		PMSI SSR		
			DP	DR	DP des RUM	DA, DP/DR des RUM	DP	DA	MMP	AE	DA
Syndrome coronaire aigu	-	-	-	-	1	-	-	-	-	-	-
Maladie coronaire chronique	1	1	-	5	5	1	-	-	1	1	1
Insuffisance cardiaque aiguë	-	-	-	1	1	1	-	-	-	-	-
Insuffisance cardiaque chronique	1	-	-	5	5	5	-	-	1	1	1
Maladie valvulaire	1	-	-	5	5	1	-	-	1	1	1
Diabète	1	2	2	2	-	2	-	-	-	-	-
Cancer du côlon actif	1	-	2	2	-	2	-	-	-	-	-
Cancer du côlon sous surveillance	1	-	5	5	-	2	-	-	-	-	-
Cancer bronchopulmonaire actif	1	-	2	2	-	2	-	-	-	-	-
Cancer bronchopulmonaire sous surveillance	1	-	5	5	-	2	-	-	-	-	-
Troubles psychotiques	1	1	5	5	-	5	5	5	5	5	5
Troubles névrotiques et de l'humeur	1	1	5	5	-	5	5	5	5	5	5
Déficience mentale	1	-	2	2	-	-	2	2	2	2	2
Troubles addictifs	1	1 *(liés au tabac)*	2	2	-	-	2	2	2	2	2
Sclérose en plaques	1	1	5	5	-	-	-	-	5	5	-
Épilepsie	1	1	5	5	-	-	-	-	-	-	-

Source : Méthodologie médicale de la cartographie des pathologies et des dépenses, version G11

Des algorithmes centralisés et standardisés dans un référentiel médical

| = OU

Nom de la requête	*Libellé*	*Date de début de la période considérée*	*Date de fin de la période considérée*	*Code cim10 à inclure*	*Code cim10 à exclure*
REQ	**label**	**date_deb**	**date_fin**	**CIM_in**	**CIM_out**
DIA_ALD1	diabète	01/01/Y	31/12/Y	E10|E11|E12|E13|E14	

... ALD | MED | BIO | PRS | PMSI_MCO | PMSI_SSR | PMSI_PSY | PMSI_HAD | CCAM | Population | SCRIP

Nom de la requête	*Libellé*	*Date de début de la période considérée*	*Date de fin de la période considérée*	*Nombre de delivrance minimale sur la période (à différentes dates)*	*Mot-clef du referentiel de médicaments (refCIP_Gi) associé aux produits à rechercher*
REQ	**label**	**date_deb**	**date_fin**	**MED_nbdlv**	**kw_comb**
DIA_CAT_MED1_3D	antidiabétiques oraux ou injectables spécifiques	01/01/Y	31/12/Y	3	diab_autres|diab_insul|diab_glp_insul|diab_glp|diab_sglt
DIA_CAT_MEDm1_3D	antidiabétiques oraux ou injectables spécifiques l'année n-1	01/01/Y-1	31/12/Y-1	3	diab_autres|diab_insul|diab_glp_insul|diab_glp|diab_sglt
DIA_INS_MED1_3D	insulines	01/01/Y	31/12/Y	3	diab_insul|diab_glp_insul
DIA_AGL_MED1_3D	agonistes GLP-1 spécifiques du diabète	01/01/Y	31/12/Y	3	diab_glp|diab_glp_insul
DIA_SGL_MED1_3D	inibiteurs SGLT2 spécifiques du diabète	01/01/Y	31/12/Y	3	diab_sglt

... | triplets_labels | Synthese | Combinaisons | ALD | MED | BIO | PRS | PMSI_MCO | PMSI_SSR | PMSI_PSY | ...

Autres extraits du référentiel médical

MCO

REQ	label	date_deb	date_fin	typeET_OU	MCO_DP	MCO_DR	DP_DR	MCO_DA	MCO_DPrum	MCO_DRrum	DA_DPr_DRr
DIA_CAT_MCO2_DP_TR	diabète	01/01/Y-1	31/12/Y	OU	E10|E11|E12|E13|E14						
DIA_CAT_MCO2_DPDADPrDRr_TR	DP de complication du diabète associé à un DA (ou DP/DR de RUM) de diabète	01/01/Y-1	31/12/Y	ET_SEJ	L97|G590|G632|G730|G990|H280|H360|I792|M142|M146|N083						E10|E11|E12|E13|E14
DIA_CAT_MCO2_DR_TR	diabète	01/01/Y-1	31/12/Y	OU		E10|E11|E12|E13|E14					
DIA_CAT_MCO2_DRDADPrDRr_TR	DR de complication du diabète associé à un DA (ou DP/DR de RUM) de diabète	01/01/Y-1	31/12/Y	ET_SEJ		L97|G590|G632|G730|G990|H280|H360|I792|M142|M146|N083					E10|E11|E12|E13|E14

Combinaison des critères

I	J	K	L
Nom du top (output)	*Nom détaillé*	*Combinaison de requêtes/ tops / sups / combinaison complexes provenant de l'onglet Combinaison / Scripts*	*top à exclure*
DIA_CAT_CAT	Diabète	DIA_ALD1|DIA_CAT_MED1_3D|DIA_CAT_MEDm1_3D|(DIA_CAT_MCO2_DP_TR|DIA_CAT_MCO2_DR_TR)|(DIA_CAT_MCO2_DPDADPrDRr_TR|DIA_CAT_MCO2_DRDADPrDRr_TR)	
DIA_INS_IND	Diabète insulino-traité	DIA_INS_MED1_3D&DIA_CAT_CAT	
DIA_AGL_IND	Diabète traité par agoniste du GLP-1 spécifique	DIA_AGL_MED1_3D&DIA_CAT_CAT	

... | triplets_labels | **Synthese** | Combinaisons | ALD | MED | BIO | PRS | PMSI_MCO | PMSI_SSR | PMSI_PSY | ...

En dehors des résultats, que mettons-nous à disposition ?

- Données individuelles annuelles sur le SNDS (4 tables/année)
- Documentation méthodologique sur ameli
- Article scientifique : Rachas et al. Medical Care 2022
- Programmes de repérage des pathologies en open source

Partage des programmes R en open source

- Modifiés pour être utilisables hors Cartographie
- Toute étape liée aux dépenses est exclue
- Référentiel médical, liste des médicaments spécifiques, documentation
- Référentiel exemple pour tester
- Git du HDH et BOAS, sous licence GPL 3 ➔republier les modifications +++
- Nous sommes preneurs de retours constructifs !

ATTENTION !
NE PAS LANCER LES PROGRAMMES SUR TOUS LES ALGORITHMES EN 1 FOIS

image : https://www.freepik.com/premium-ai-image/server-computer-explosion-with-fire-smoke_37814761.htm

Limites de la Cartographie des pathologies

- Sous-estimation de la prévalence des pathologies ➔prises en charge
- Certaines pathologies non comprises
- Données médico-administratives
- Années calendaires
- Algorithmes très complexes ne peuvent être intégrés dans l'outil
- Référentiel médical complexe (cancers +++)
- Précautions :
  - Rappeler l'objectif de la Cartographie : répartir les dépenses
  - Travailler sur 1 version (évolution des méthodes)
  - Attention lors de l'extension à d'autres années

Quels usages pour la communauté scientifique ?

- Faciliter les études sur le SNDS
- Comprendre la méthodologie de la Cartographie dans le détail
- Réutiliser les requêtes
- Créer/modifier les algorithmes
- Perspectives (à vous de jouer !) :
  - Traduction des programmes dans d'autres langages
  - Adaptation pour une Cartographie « glissante »

Remerciements

- Groupe de travail organisé par l'AMDAC pour la rédaction du **Guide d'accompagnement à l'ouverture d'un code source**
- Théophile Daney de Marcillac, pour son aide juridique sur les licences
- Le Health Data Hub pour la mise à disposition sur le Git et la BOAS
- Sofiane Kab pour son aide sur le partage du programme du top diabète
- REDSIAM pour les échanges entre experts et les algorithmes qui ont largement nourri la Cartographie depuis sa création
- Toute l'équipe de la Cartographie, fortement mobilisée pour ce projet

SÉCURITÉ SOCIALE

l'Assurance Maladie

Agir ensemble, protéger chacun

annexes

Pourquoi passer à R ?

- Capable de faire l'ensemble des tâches du processus
- Logiciel libre et de plus en plus répandu. Réflexion générale dans le domaine public
- Facilite l'intégration de nouveaux collaborateurs, en particulier les plus jeunes formés
- Très efficient (cependant dépendant de l'infrastructure en place)
- Espace de stockage de données réduits
- Facilite le travail collaboratif (si module Github en place)
- Motivant pour les membres de l'équipe : nouvelles compétences à acquérir

Mise en œuvre

Mobilisation de l'ensemble de l'équipe cartographie pour la construction puis la mise en œuvre du projet

Médecins de santé publique
Antoine Rachas
Panayotis Constantinou
Aide à la construction et au remplissage du contenu du référentiel, suivi du projet

Expert R
Victor Bret
Supervision et aide à toutes les étapes du projet, force de proposition

Coordonateur de la carto
Pierre Denis
Motivation ++ pour cette évolution, impliqué dans toutes les étapes du projet

Statisticiens/data managers spécifiquement en charge des "tops"
Pierre Denis
Thomas Lesuffleur
Dimitri Lastier
Impliqués dans toutes les étapes du projet

Statisticiennes de l'équipe cartographie chargées des tops et de l'analyse des résultats
Corinne Mette
Muriel Nicolas
Laurence Pestel
Pauline Barthélémy
Impliquées dans les étapes de définition du projet et de rédaction des programmes R

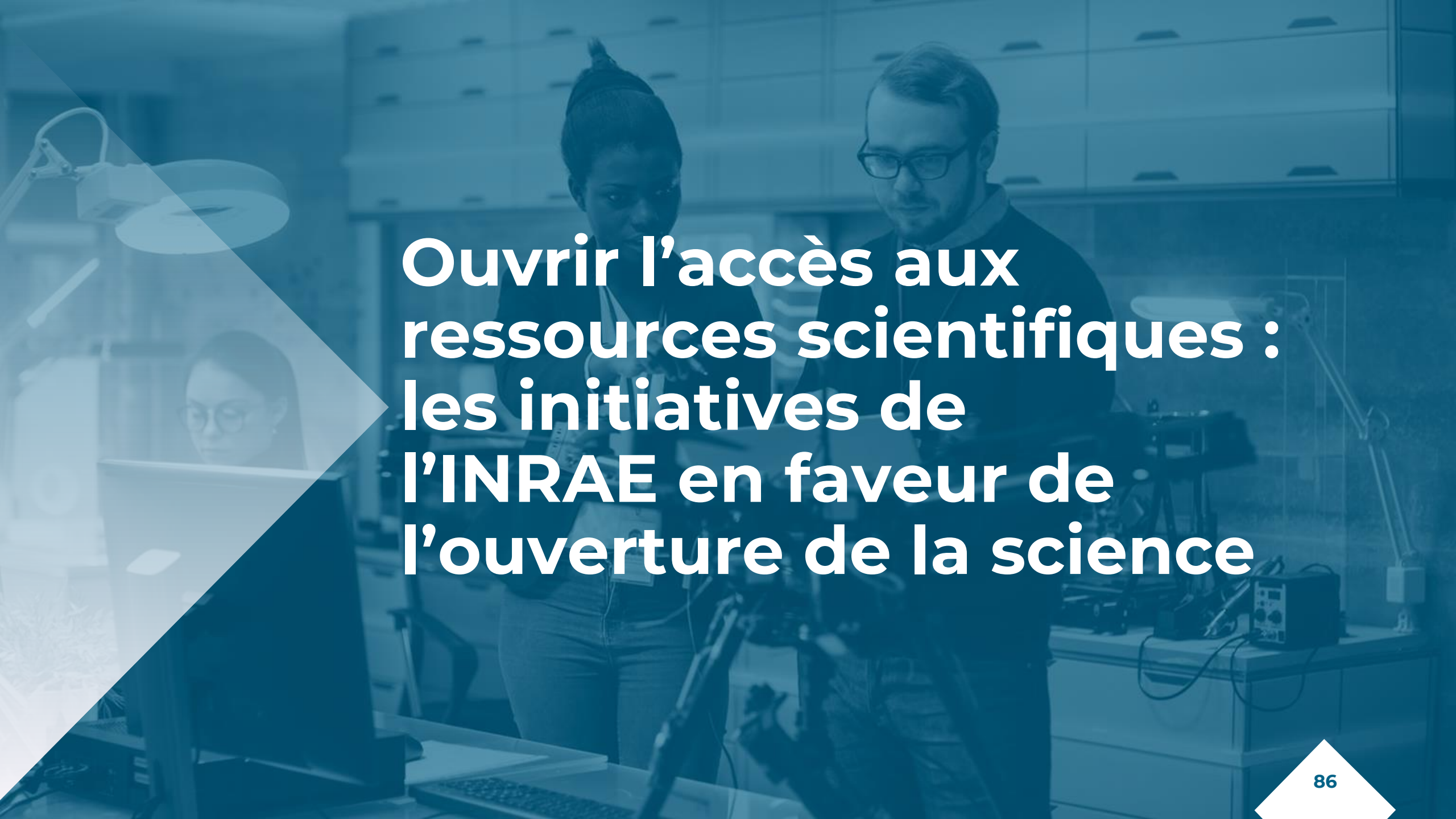
Retour d’expérience

- Difficultés rencontrées :
  - **Coût d’entrée important :**
    - Passage au logiciel R non maitrisé par l’équipe. idéalement commencer par un projet moins imposant ?
    - Remise à plat du processus, beaucoup de vérifications (création d’un programme scan)
    - Compléter le référentiel demande de connaitre ses règles
  - **Techniques :**
    - Problème de manque de mémoire sur R
    - Travailler sur deux logiciels (doublonnage de tables, transfert de tables d’un logiciel à l’autre)
  - **Difficile d’intégrer des algorithmes trop différents (dialyse)**

Retour d’expérience

❖ Points positifs:

- **Remise à plat des requêtes (correction de coquilles, homogénéisation des requêtes, vérification des filtres, etc)**
- **Remise à plat de la méthodologie médicale**
- **Processus plus rapide et plus simple d’intégrer/tester de nouveaux algorithmes (sauf exception)**
- **Avancées facilitées par l'outil : rattachement séjours HAD, création des sups cancers**
- **Séparation du code et de l’algorithme médical**
- **Apprentissage de R en mode projet**



Ouvrir l’accès aux ressources scientifiques : les initiatives de l’INRAE en faveur de l’ouverture de la science

HEALTH DATA HUB

Ouvrir l'accès aux ressources scientifiques : les initiatives de l'INRAE en faveur de l'ouverture de la science

11h15 - 11h45

Odile Hologne

Responsable de la Direction pour la science ouverte (DipSO) d'INRAE

Odile Hologne est Ingénieure Générale des Ponts des Eaux et des Forêts et directrice de la Direction pour la Science Ouverte depuis la création d'INRAE le 1 er janvier 2020. Cette direction coordonne la politique SO de l'institut et sa mise en œuvre pour ouvrir les processus de recherche et rendre librement accessible les résultats. Depuis plusieurs années Odile Hologne est impliquée dans la diffusion des connaissances scientifiques en ayant la responsabilité d'activités éditoriales, ou de vulgarisation scientifique et depuis 2011 elle est impliquée dans différentes initiatives pour le partage des données de la recherche que ce soit au niveau international (RDA, GOFAIR, ...), européen (membre du scientifc Advisory board de la plateforme de publication d'articles Open Research Europe de la commission européenne, coordination et participation à des projets et groupes de travail de l'European Open Science Cloud) ou national dans le cadre du comité national pour la science ouverte.

Les initiatives d'INRAE en faveur de l'ouverture de la science

Odile Hologne, Direction pour la Science Ouverte (DipSO)

INRAE : Institut national de recherche pour l'agriculture, l'alimentation et l'environnement

Agriculture

Environment

Food

Global challenges
Climate change and risks, One health, agri-food systems

Collaborer avec l'ensemble de la société

INRAE inscrit son activité dans le dispositif de l'enseignement supérieur, la recherche et l'innovation français et européens, développe des partenariats avec l'ensemble des acteurs publics et privés nationaux et internationaux de nos domaines et nourrit un dialogue avec les citoyens et parties prenantes.

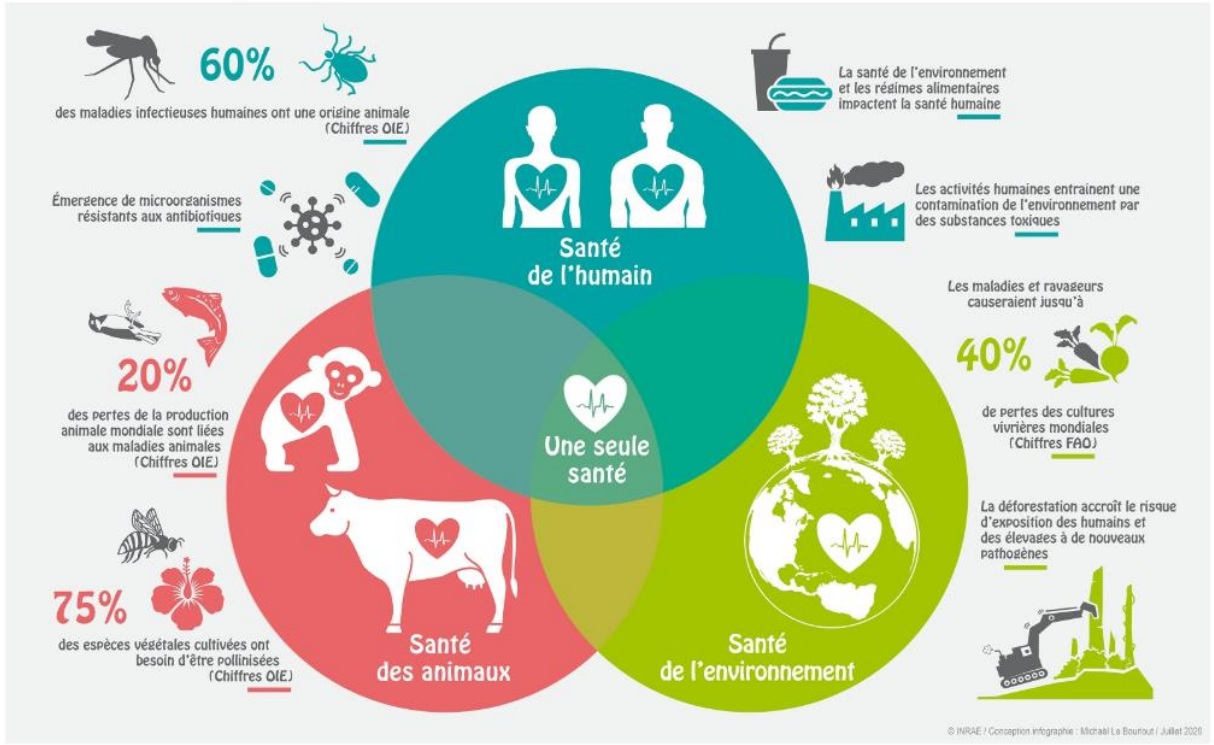
Notre engagement pour une science ouverte

Partager les résultats de recherche, non seulement entre scientifiques, mais aussi avec la société, est le meilleur moyen de faire progresser la connaissance et de développer des relations de confiance avec les différents acteurs.

Expertise et appui aux politiques publiques

Les expertises d'INRAE participent à éclairer la société et les décideurs et gestionnaires publics afin d'élaborer et accompagner des politiques adaptées et efficaces.

Des travaux en lien avec la santé humaine



PREZODE
Preventing zoonotic disease emergence

https://www.inrae.fr/alimentation-sante-globale

Sujets abordés

- Politique Science Ouverte d'INRAE et Gouvernance des données
- Exemples d'actions et analyse d'impacts

Politique Science Ouverte
Gouvernance des données

Les facettes de la Science ouverte

Avoir des outils ouverts pour collaborer

Efficience, économie

Rendre la science accessible aux citoyens

« Les communs » de la science

Évaluer la science autrement

Fecher B., Friesike S. (2014) Open Science: One Term, Five Schools of Thought. In: Bartling S., Friesike S. (eds) Opening Science. Springer, Cham. https://doi.org/10.1007/978-3-319-00026-8_2

Politique SO INRAE : Les axes

DipSO en animation d'une politique fédératrice impliquant différents acteurs

Une politique en phase avec nos enjeux scientifiques, et l'approche holistique de l'Unesco

Mise en œuvre d'une politique

Des opportunités et des risques

- INRAE met en œuvre une politique de science ouverte
  - Ouvrir ses résultats de recherche à différents acteurs socio-économiques
  - Ouvrir les processus de recherche
  - Etre lui-même utilisateur de données produites par d'autres pour répondre à ses propres enjeux de recherche

Une politique incontournable pour les projets

- Analyser l'impact de l'ouverture
  - Ouvrir comporte des risques de différentes natures (juridique, financier, économique, réputationnels, concurrence , …)
  - Création de valeur à partir des données ouvertes peut être réalisée par des tiers (nouvelles connaissances, nouveaux produits, ou nouveaux services).

Gouvernance des données : 4 Principes

- Science : Il faut réutiliser et partager les données en respectant les valeurs de la science
  - Intégrité, déontologie, éthique, collaboration, embargo.
- Gestion des données : Les données doivent être gérées en vue de faciliter leur découverte et leur réutilisation
  - Plan de gestion des données et principes FAIR
- Réglementation : Les données doivent être « aussi ouvertes que possible, aussi fermées que nécessaire »
  - Ouvert et gratuit par défaut (financement publique)
- Innovation : L'ouverture des données contribue à l'innovation et à la création de valeur pour la société
  - La création de valeur (économique ou autre) peut-être réalisée par des tiers.

https://science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/la-gouvernance-des-donnees-algorithmes-et-codes

Une organisation et des services

Sur toutes les facettes de la Science Ouverte https://science-ouverte.inrae.fr/fr

En particulier pour les données :

- Une direction en appui à la gestion, partage, ouverture
- Un administrateur des algorithmes, données et codes
- Une cellule « gouvernance des données »
- Des référents données :
  - « opérationnels » dans les unités de recherche
  - « Stratégiques » dans les départements de recherche

Exemples d’actions et analyse d’impacts

Ouverture des publications
Ouverture des données
Ouverture à la société
Diplomatie scientifique et données

Bonnes pratiques de publication – libre accès

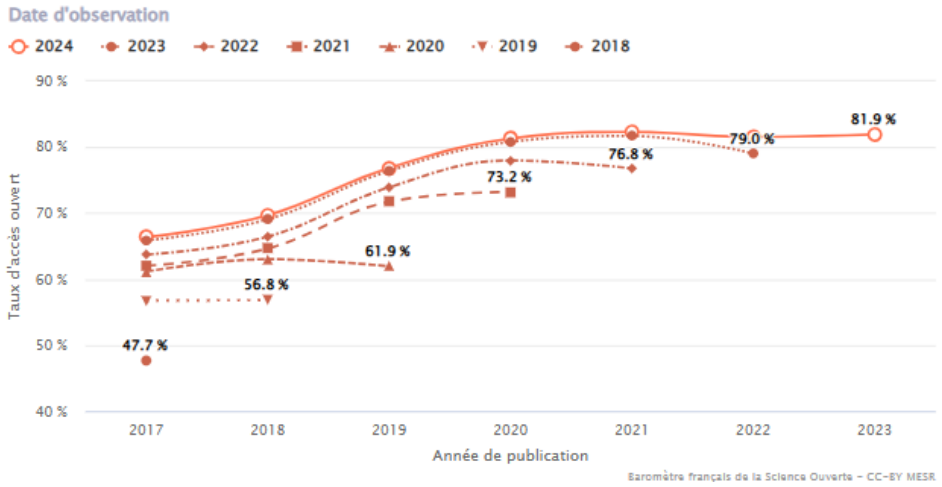
Bonnes pratiques

Politique : 100% libre accès à l'horizon 2030

En synthèse

Tout scientifique reste libre de choisir ses supports de publication quels que soient leurs modèles économiques. L'institut formule cependant des conseils aux auteurs pour que leurs résultats soient librement accessibles notamment en déposant le texte intégral de leurs publications dans HAL INRAE, en mettant également à disposition les données et les codes (résultats reproductibles). Il doit pour cela connaître les dispositifs mis à sa disposition et s'assurer en amont que ses productions répondent aux critères d'éligibilité à cette diffusion (sur le plan juridique, de l'intégrité, en matière de délai et de format compatibles avec la stratégie scientifique de son collectif). De plus, il est conseillé d'appliquer certains principes pour limiter les dépenses liées à la publication. Depuis 2023, l'évaluation-conseil et les promotions des chercheurs s'appuient sur la version librement accessible des articles publiés dans les revues scientifiques, via la liste des productions exportée de HAL INRAE et sur les autres productions librement accessibles.

Pour approfondir



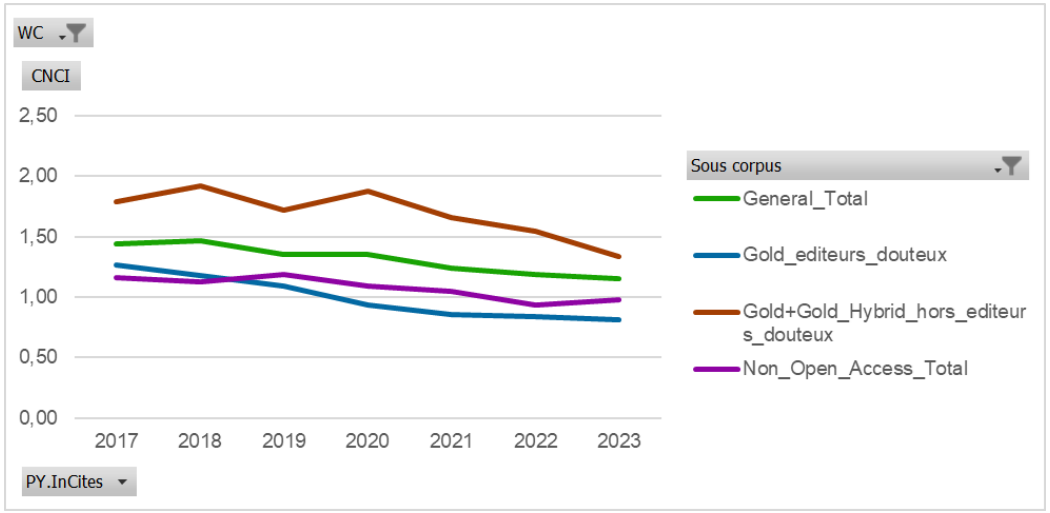
Effet de la politique sur les pratiques

https://science-ouverte.inrae.fr/fr/les-publications-scientifiques/bonnes-pratiques-de-publication

https://science-ouverte.inrae.fr/fr/la-science-ouverte/le-barometre-de-la-science-ouverte-inrae

Citations selon les modalités de publication

Les publications « gold » (① rouge) a l'exception de celles qui sont publiées chez MDPI et Frontiers (③ bleu) ont un CNCI supérieur aux publications sous abonnements (② violet). A partir de 2019, les publications chez MDPI et Frontiers sont moins citées que la moyenne



③
①
②

Le CNCI d'un document est calculé en divisant ses citations « a date » par le nombre moyen de citations reçues par les documents de même type, même année et même catégorie thématique.

Le CNCI d'un ensemble de documents, par exemple les œuvres complètes d'un individu, d'une institution ou d'un pays, est la moyenne des valeurs CNCI de tous les documents de l'ensemble.

	③	①	②	
Année	General_Total	Gold_editeurs_douteux	Gold+Gold_Hybrid_hors_editeurs_douteux	Non_Open_Access_Total	
2017	1,44	1,26	1,79	1,16	54%
2018	1,46	1,18	1,92	1,12	71%
2019	1,35	1,09	1,72	1,18	45%
2020	1,35	0,94	1,87	1,09	71%
2021	1,24	0,86	1,66	1,05	58%
2022	1,19	0,84	1,54	0,93	66%
2023	1,16	0,81	1,33	0,97	37%

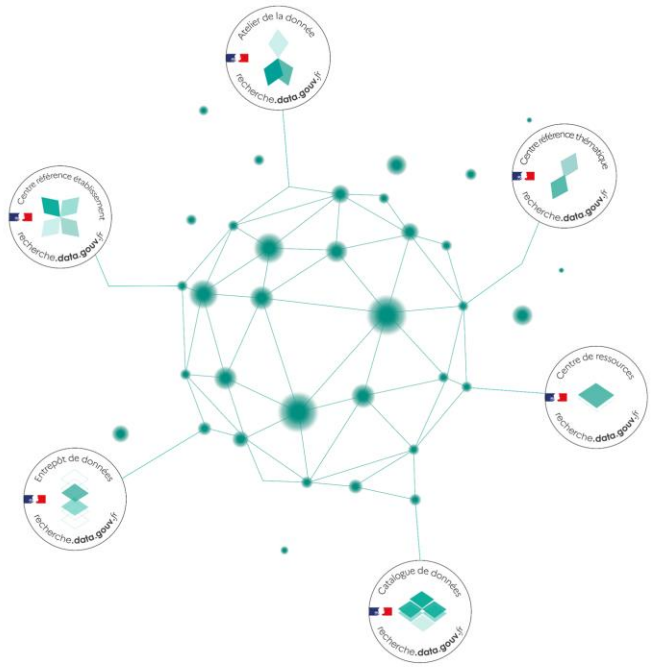
Sur la période, en moyenne les publications Gold (1) ont un CNCI supérieur de 58% au CNCI des articles sous abonnement (2)

Focus sur Recherche Data Gouv

un écosystème au service du partage et de l'ouverture des données de recherche
https://recherche.data.gouv.fr/

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

recherche.data.gouv.*fr*



- **Initiative du ministère** de la Recherche
- **Composé :**
  - **De centre de compétences** (Ateliers de la données animés par les universités, centres de ressources thématiques liés aux infrastructures de recherche)
  - **D'une plateforme nationale des données de la recherche** (entrepôt pour le dépôt de données et catalogue des données) qui permet à chaque institution d'avoir son entrepôt

  INRAE est responsable du projet de plateforme nationale
  Un travail collaboratif

Réutilisation de jeux de données

Recherche Data Gouv > Data INRAE >

Hyperspectral images of grape leaves including healthy leaves and leaves with biotic and abiotic symptoms

Version 1.0

Ryckewaert, Maxime, 2023, "Hyperspectral images of grape leaves including healthy leaves and leaves with biotic and abiotic symptoms", https://doi.org/10.57745/WW7TY7, Recherche Data Gouv, V1, UNF:6:kdwliLQtkYShebYAgppEUQ== [fileUNF]

Citer le jeu de données

Pour en apprendre davantage sur le sujet, consulter le document Data Citation Standards [en].

Modalités d'accès au jeu de données

Contact | Partager

Métriques Make Data Count (MDC)
depuis 2020-07-01

158 292 consultations

144 723 téléchargements

2 citations

Description A hyperspectral imaging database was collected on two hundred and five grape plant leaves. Leaves were measured with a hyperspectral camera in the visible/near infrared spectral range under controlled conditions. This dataset contains reflectance spectra of grape leaves of seven different varieties. For each variety, healthy leaves and leaves with foliar symptoms caused by different grapevine diseases showing clear symptoms of biotic or abiotic stress on other organs. English

Sujet Earth and Environmental Sciences; Agricultural Sciences

Mot-clé Hyperspectral

Licence/Conditions d'utilisation des données etalab 2.0

Téléchargements anonymes : qui réutilise, pour faire quoi ?

Plus de 300 projets de Recherche participatives

https://science-ouverte.inrae.fr/fr/sciences-et-recherches-participatives-srp/les-projets-de-science-ou-recherche-participative

FOCUS PROJET #1

NutriNet-Santé

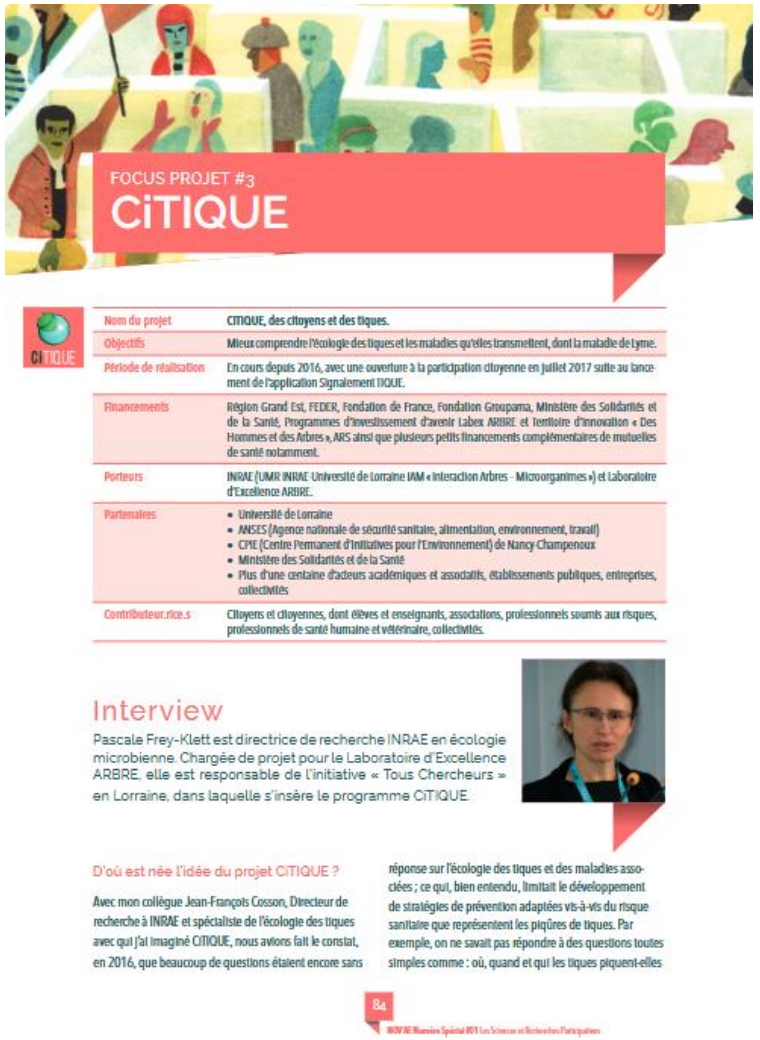
Nom du projet	NutriNet-Santé
Objectifs	Étudier les relations nutrition-santé
Période de réalisation	Depuis 2009
Financements	Financements publics pour l'étude NutriNet-Santé et les différents projets adossés
Partenaires	Université Sorbonne Paris Nord, Inserm, INRAE, Cnam
Contributeur.rice.s	Citoyens et citoyennes volontaires

Interview

Benjamin Allès est chargé de recherche en épidémiologie de la nutrition, en poste à INRAE depuis sept ans. Rattaché au Centre de Recherche en Épidémiologie et Statistiques (CRESS, Université de Paris et Université Sorbonne Paris Nord), il travaille au sein de l'équipe de recherche en épidémiologie (EREN – Université Sorbonne Paris Nord, Inserm, INRAE, Cnam), qui coordonne l'étude NutriNet-Santé, une étude de cohorte lancée en 2009 pour étudier les relations nutrition-santé et dont dérive un très grand nombre de projets aux objectifs plus spécifiques.

Peux-tu nous présenter l'étude NutriNet-Santé en quelques mots ?

L'étude NutriNet-Santé a été lancée en 2009 avec pour objectif d'étudier les liens entre nutrition et santé et de comprendre les déterminants des comportements alimentaires et d'activité physique, que ce soient des déterminants socio-démographiques, économiques, géographiques, psychologiques ou autres, et les facteurs de modes de vie associés, comme le statut tabagique. NutriNet-Santé, c'est en fait une cohorte, c'est-à-dire un groupe de personnes suivies et observées dans le temps, sans pour autant chercher à modifier leurs habitudes. Cette cohorte compte, aujourd'hui, plus de 171 000 volontaires, que l'on nomme les « nutrinautes », et a déjà permis de générer un peu plus de 200 études spécifiques, depuis son lancement. Il s'agit de la première « web-cohorte » de cette taille dans le monde dans le domaine nutrition-santé. Elle est caractérisée par une évaluation très fine des comportements alimentaires et des expositions nutritionnelles. L'étude est coordonnée par l'EREN (Équipe de recherche en Épidémiologie Nutritionnelle du CRESS), sous tutelle de Sorbonne Paris Nord, de l'Inserm, d'INRAE, et du Cnam. Nous travaillons bien sûr avec des collègues extérieurs à l'équipe, par exemple, côté INRAE, avec des chercheur.e.s des départements ALIMH et EcoSocio. L'étude NutriNet-Santé nous permet également d'étudier les liens entre profils alimentaires végétalisés (incluant les végétarismes)



FOCUS PROJET #3

CiTIQUE

Nom du projet	CITIQUE, des citoyens et des tiques.
Objectifs	Mieux comprendre l'écologie des tiques et les maladies qu'elles transmettent, dont la maladie de Lyme.
Période de réalisation	En cours depuis 2016, avec une ouverture à la participation citoyenne en juillet 2017 suite au lancement de l'application Signalement TIQUE.
Financements	Région Grand Est, FEDER, Fondation de France, Fondation Groupama, Ministère des Solidarités et de la Santé, Programmes d'investissement d'avenir Labex ARBRE et Territoire d'innovation « Des Hommes et des Arbres », ARS ainsi que plusieurs petits financements complémentaires de mutuelles de santé notamment.
Porteurs	INRAE (UMR INRAE Université de Lorraine IAM « Interaction Arbres - Microorganismes ») et Laboratoire d'Excellence ARBRE.
Partenaires	• Université de Lorraine • ANSES (Agence nationale de sécurité sanitaire, alimentation, environnement, travail) • CPIE (Centre Permanent d'Initiatives pour l'Environnement) de Nancy-Champenoux • Ministère des Solidarités et de la Santé • Plus d'une centaine d'acteurs académiques et associatifs, établissements publiques, entreprises, collectivités
Contributeur.rice.s	Citoyens et citoyennes, dont élèves et enseignants, associations, professionnels soumis aux risques, professionnels de santé humaine et vétérinaire, collectivités.

Interview

Pascale Frey-Klett est directrice de recherche INRAE en écologie microbienne. Chargée de projet pour le Laboratoire d'Excellence ARBRE, elle est responsable de l'initiative « Tous Chercheurs » en Lorraine, dans laquelle s'insère le programme CiTIQUE.

D'où est née l'idée du projet CITIQUE ?

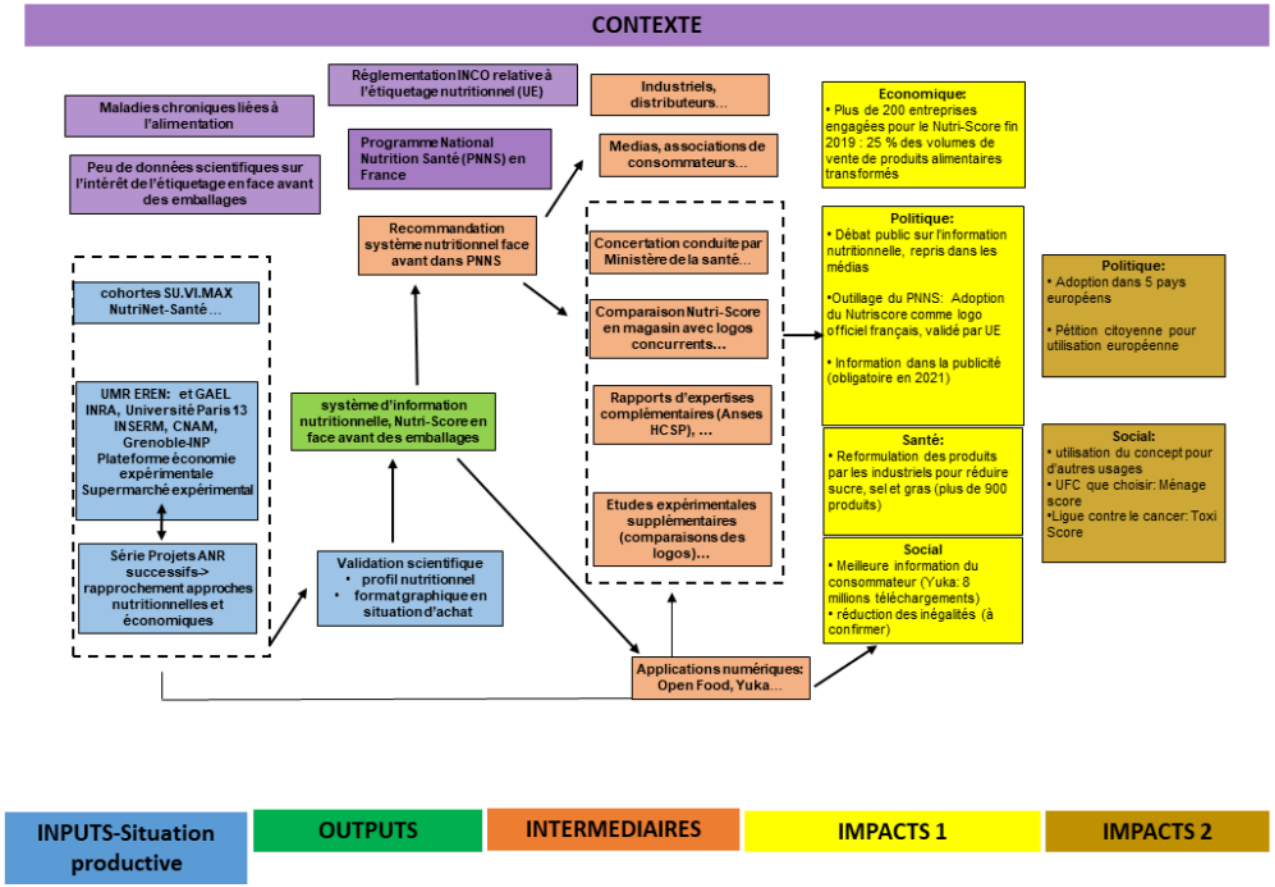
Avec mon collègue Jean-François Cosson, Directeur de recherche à INRAE et spécialiste de l'écologie des tiques avec qui j'ai imaginé CITIQUE, nous avions fait le constat, en 2016, que beaucoup de questions étaient encore sans réponse sur l'écologie des tiques et des maladies associées ; ce qui, bien entendu, limitait le développement de stratégies de prévention adaptées vis-à-vis du risque sanitaire que représentent les piqûres de tiques. Par exemple, on ne savait pas répondre à des questions toutes simples comme : où, quand et qui les tiques piquent-elles

Analyser les impacts des Recherches participatives

Méthode ASIRPA – chaine d'impacts

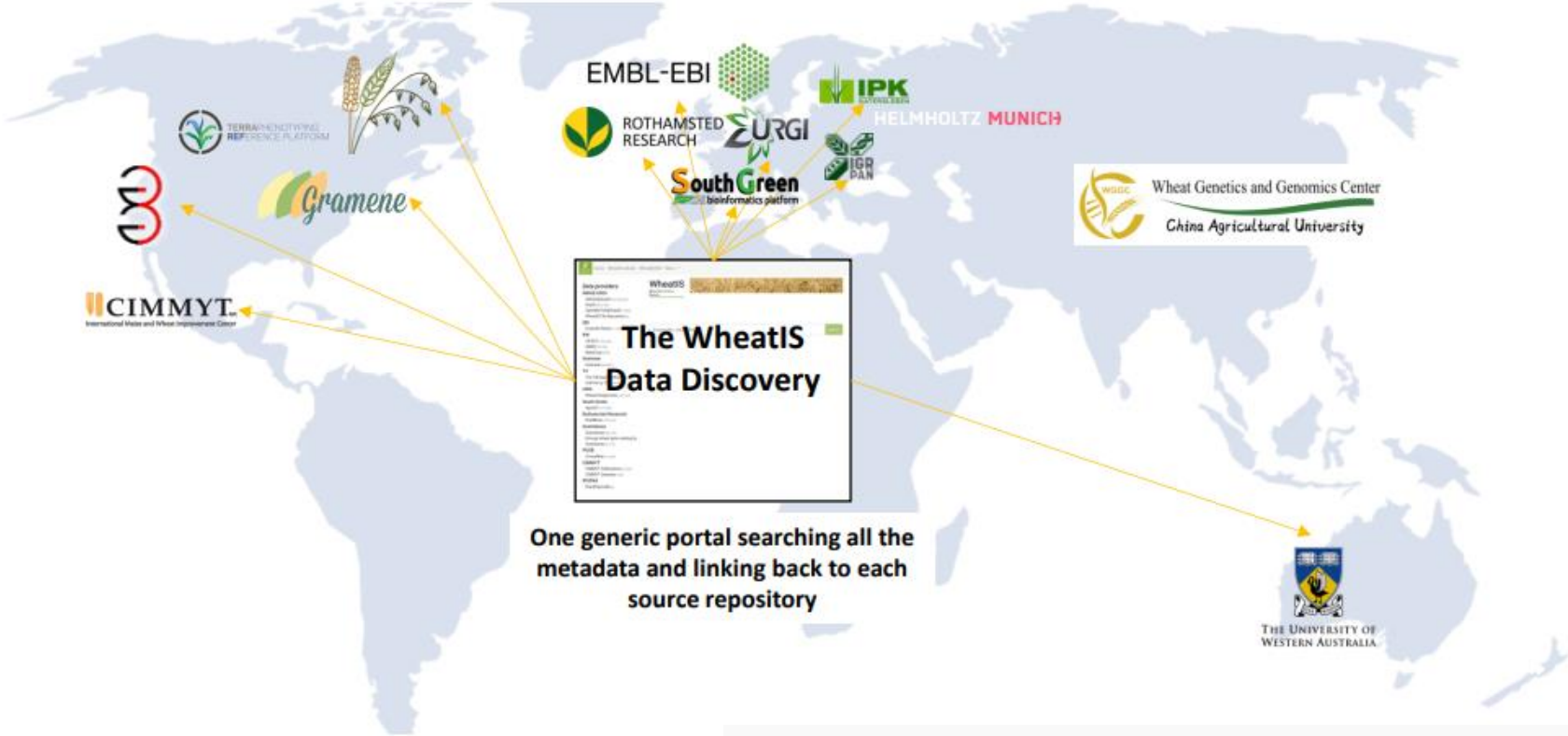
https://asirpa.hub.inrae.fr/60-cas-etudes/etudes-de-cas/alimentation/nutri-score

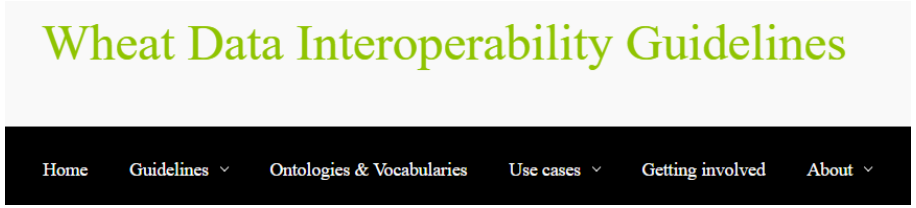
la méthode ASIRPA, développée par INRAE, permet d'évaluer les effets des projets dans cinq dimensions : **économique, sociale, environnementale, politique et en matière de circulation des connaissances.**



Wheat initiative (2011) et Système d'information sur le blé

http://www.wheatis.org/





https://ist.blogs.inrae.fr/wdi/

Les infrastructures de données au cœur des stratégies de recherche et de partenariat scientifique

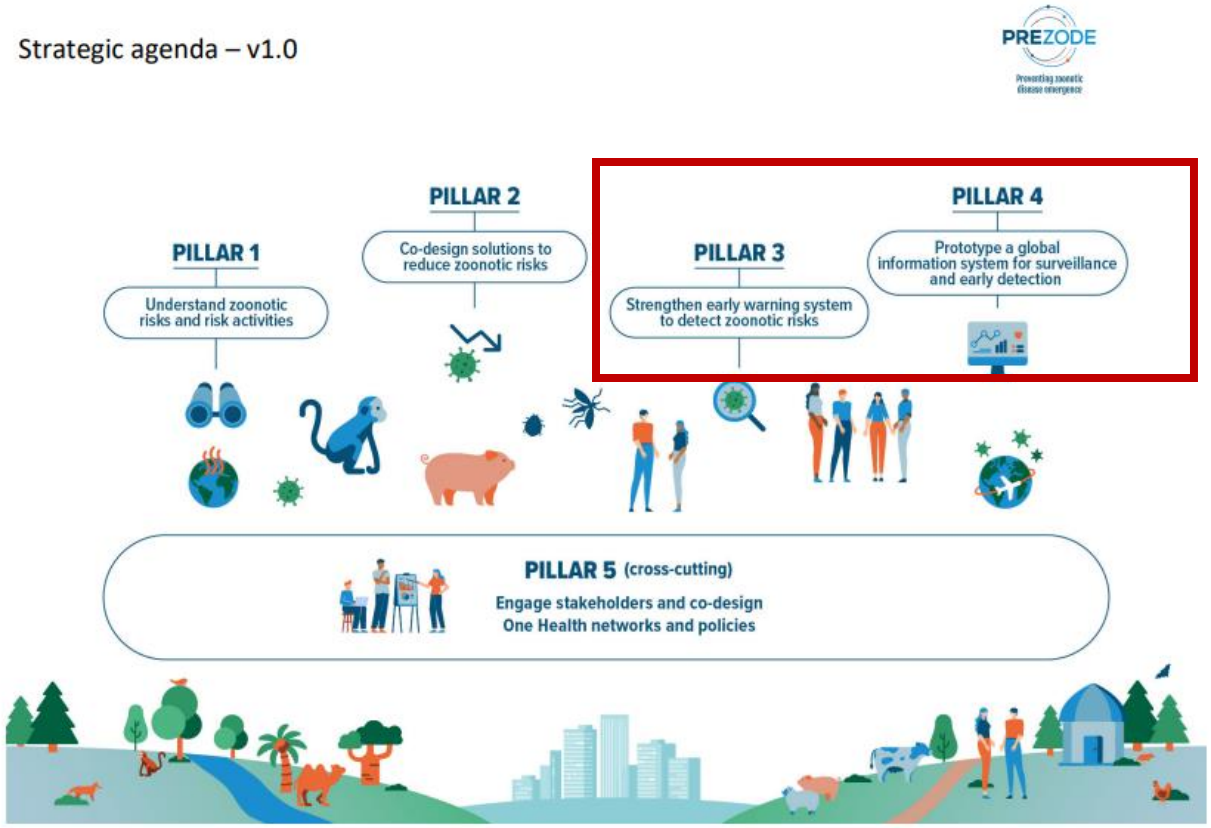


Figure 4. *The five pillars of the PREZODE Initiative.*

https://prezode-initiative.org/wp-content/uploads/2025/03/PREZODE_Strategic_Agenda.pdf

INRAE DipSO

conclusion

Les promesses de la Science Ouverte

En regard d'objectifs politiques

Soutenir à l'innovation ?

- L'ouverture des résultats favorise t-elle la création de valeur (économique ou non) ? Oui : ex des IA génératives

Faciliter l'interdisciplinarité ?

- L'ouverture décloisonne t-elle les communautés ?
- L'interdisciplinarité a-t-elle augmenté avec l'ouverture ?

Augmenter la confiance dans la science ?

- Ce n'est pas parce que les résultats sont ouverts et accessibles que les citoyens ont « confiance »

Faciliter l'impact et poursuivre les analyses

- Impact académique : mesures d'usage et de citation
  - Publications, données, codes
- Impacts sociétaux et économiques
  - Amélioration des connaissances des participants des projets de SRP – partage de connaissances
  - Réutilisation des données - > innovation
- Impacts « organisationnels »
  - sur les participants des projets de SRP, posture des chercheurs
  - Les stratégies de partenariat scientifique

Merci pour votre attention

http://www.inrae.fr/dipso

https://intranet.inrae.fr/evaluation/Chercheurs

Annonce des lauréats de la 8e vague de l’AMI BOAS

HEALTH DATA HUB

Annonce des lauréats de la 8e vague de l'AMI BOAS

11h45 - 12h30

Maxime Caillet

Responsable du pôle Projets Partenaires à la Direction des Partenariats du Health Data Hub

Au sein de la direction des Partenariats du Health Data Hub, Maxime pilote les appels à projets lancés par le HDH, participe au conventionnement des projets lauréats et assure le suivi de leur avancement. Il a participé à l'organisation des quatre dernières éditions de l'AMI BOAS et est régulièrement en interaction avec les équipes qui bénéficient de l'accompagnement du HDH.

I - Présentation de l’AMI BOAS

Introduction

Le Health Data Hub, une structure impulsée et soutenue dans le cadre de politiques publiques d'ampleur

Les usages des données de santé se multiplient et accéder aux sources de données dans des délais les plus courts possibles est essentiel. Le Health Data Hub est une structure publique créée fin 2019 visant à faciliter l'accès aux données de santé aux projets d'intérêt public, dans la continuité de l'ouverture du SNDS en 2016.

Un **guichet unique** sur les **données de santé en France**

Une **plateforme sécurisée** et **à l'état de l'art**

Un **catalogue de données**, incluant **une des plus larges bases de données médico-administratives au monde**

Un projet impulsé et soutenu par des politiques publiques d'ampleur

Créé à la suite du **rapport Villani** promouvant la **mise en place de grandes plateformes de partage de données pour favoriser l'IA**

Projet phare et un contributeur actif de la **Stratégie Nationale pour l'intelligence artificielle**

L'un des grands projets du **FTAP** et l'un des rares auditionnés en mars 2022 par la Ministre Amélie de Montchalin

L'une des actions de la **feuille de route du Numérique en Santé** 2019-2022 du Ministère de la Santé et de la Prévention

Appel à manifestation d'intérêt - BOAS
Objectifs du programme BOAS

*Soutenir le **développement et la validation d'algorithmes et de programmes** informatiques **facilitant la réutilisation** de toutes les bases de données*

***Mutualiser les efforts** autour des outils de manipulation des données de santé*

***Consolider et fiabiliser les outils** d'analyse des données*

***Promouvoir l'open source** dans la recherche sur les données de santé*

Afin de mettre à disposition de nouveaux algorithmes et outils dans la BOAS,
L'AMI BOAS du Health Data Hub est ouvert en continu et plusieurs relèves des dossiers de candidatures sont organisées.

Appel à manifestation d'intérêt - BOAS

25 projets sont dores et déjà accompagnés dans le cadre du programme BOAS

Maltraitance : épidémiologie de la maltraitance envers les enfants

Elan : épidémiologie de résilience chez les enfants

PerinatAlgoVal : Périnalité

Predict EDI : Position socio économique

EAI Algo : encéphalopathie chez les nouveaux nées

ALIA : sous-types d'AVC hospitalisés

DIONISOS : sclérose en place

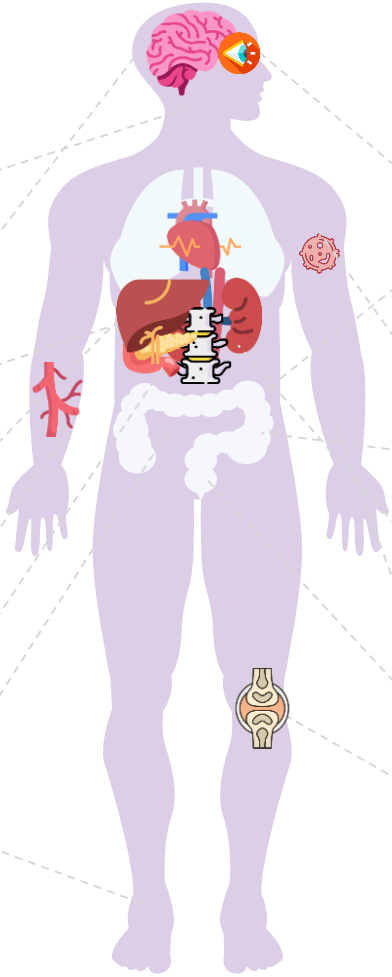
Hepavir-Algo: hépatite B et C

EMIR-Algo : maladies inflammatoires systémiques rares

Papricap 2 : Cancers du pancréas

Thanatos : chirurgie digestive

Poseidon : orthopédie



API-DMLA : DMLA

Poumalgo : cancer du poumon

LMC RCT : leucémie myéloïde chronique

Renalgo : maladies rénales chroniques (MRC)

SNDS GraftLoss : Perte de greffon

AICC-MC : cancers cutanés

ALGO-IBD : Maladies inflammatoires chroniques intestinales

Projet Rhumatismes inflammatoires

RISH : indicateurs de risques de handicap

ACR : hospitalisations suite aux accidents de route

Ican Algo : cancers incidents

Adhérence : maladies chroniques

SCOP : score de mode de vie

1ère vague (2021) | 2e vague (2022) | 3e vague (2022) | 4e vague (2023) | 5e vague (2024) | 6e vague (2024) | 7e vague (2024)

Evolution de l'AMI BOAS

Suite à la relève de la vague 7, l'AMI évolue

Historiquement

Thématique 1

Documentation médicale et technique d'**algorithmes de ciblage** existants pour favoriser leur réutilisation large

Thématique 2

Évaluation et sophistication d'algorithmes développés sur la base principale du SNDS par des institutions nationales

Thématique 3

Développement, évaluation et validation de nouveaux algorithmes de ciblage sur la base principale du SNDS

Elargissement du cahier des charges

*Élargissement à **tout algorithmes ou codes sources développé sur des données de santé***

*Fusion des deux thématiques et élargissement à **tout algorithmes, codes sources ou autres programmes visant à faciliter la réutilisation de la base principale du SNDS (thématique 2) ou une autre base (thématique 3)***

Les trois thématiques de travail ont évolué pour de nouvelles opportunités

Thématique 1

Documentation médicale et technique d'algorithmes ou codes sources développés sur des **données de santé** et permettant leur réutilisation large

Thématique 2

Développement, validation et sophistication d'algorithmes de ciblage ou autres programmes visant à faciliter la réutilisation de la **base principale du SNDS**

NOUVELLE THÉMATIQUE !

Thématique 3

Développement, validation et sophistication d'algorithmes, de codes sources ou de programmes visant à faciliter la réutilisation de **données de santé autres que celles issues de la base principale du SNDS**

Le HDH accompagne financièrement les projets lauréats

Les projets lauréats
seront **financés** par
le HDH sur **une période**
de 12 à 24 mois

Thématique 1

Un montant pouvant aller **jusqu'à 10 000€[1] sur une période de 12 mois**

[1][Conditions]

Thématique 2 et 3

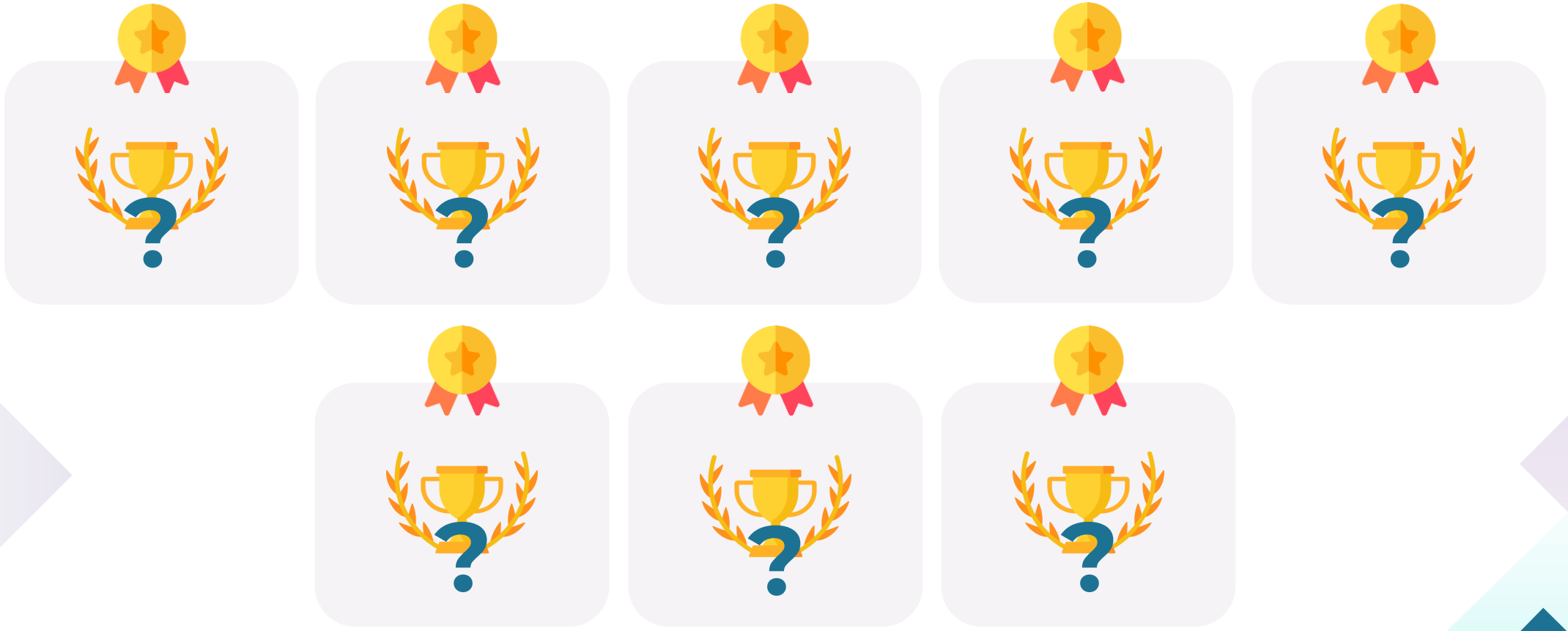
Un montant pouvant aller **jusqu'à 100 000€[2] sur une période de 24 mois**

[2][Conditions]

II - Présentation des projets lauréats BOAS 8

Présentation des lauréats AMI BOAS 8

#8 nouveaux projets lauréats à l'AMI BOAS 8



Projet MORS

HUGO
Hopitaux Universitaires Grand Ouest

Porteur.s : Groupe de coopération sanitaire HUGO

Thématique 1 : Documentation médicale et technique d'algorithmes ou codes sources développés sur des données de santé et permettant leur réutilisation large

Intervenant : Valentine Guitton

Objectifs :

- **Structurer une solution intégrée**, sous forme de **package**, à partir d'un **algorithme de rapprochement entre les entrepôts de données de santé et la base nationale des décès**, en s'appuyant sur des ressources existantes.

- L'objectif est de proposer une **solution open source**, **robuste**, **conforme au RGPD**, facilement déployable localement, et favorisant une **meilleure valorisation des données pour la recherche clinique et épidémiologique**

Présentation des lauréats AMI BOAS 8

#8 nouveaux projets lauréats à l'AMI BOAS 8

MORS

HUGO
Hopitaux Universitaires Grand Ouest







Projet PULM-ALGO

Porteur.s : CHU de Grenoble Alpes / PulmoTension

Thématique 2 : Développement, validation et sophistication d'algorithmes de ciblage ou autres programmes visant à faciliter la réutilisation de la base principale du SNDS

Intervenant : Dr Alex Hlavaty

Objectifs :

- **Evaluer et valider un algorithme facilitant l'identification de l'hypertension artérielle pulmonaire**, une maladie rare et incurable dont la prévalence est très variable.
- **Évaluer et valider un algorithme évaluant la gravité de l'HTAP**

Données mobilisées :

- Couplage des données du **SNDS** avec celle du **registre français de l'hypertension pulmonaire**

Présentation des lauréats AMI BOAS 8

#8 nouveaux projets lauréats à l'AMI BOAS 8

MORS

HUGO Hopitaux Universitaires Grand Ouest

PULM-ALGO

ulmoTension Centre de Référence de l'Hypertension Pulmonaire

FHU COURNAND Transforming Pulmonary Vascular Medicine

CHU GRENOBLE ALPES

Projet ALGO-DS

Porteur.s : Unité de recherche clinique HEGP, AP-HP

Thématique 2 : Développement, validation et sophistication d'algorithmes de ciblage ou autres programmes visant à faciliter la réutilisation de la base principale du SNDS

Intervenant : Claire Rives-Lange

Objectifs :

- **Développer et valider un algorithme de détection automatique de la dénutrition sévère lors des séjours hospitaliers dans le SNDS.**

Données mobilisées :

- **Données de l'entrepôt des données (EDS) de l'AP-HP** : données cliniques et biologiques, données médico-administratives (PMSI MCO/ HAD/ SSR et données INSEE)
- **Combinaison des données disponibles dans le SNDS** (diagnostics, actes PMSI MCO, HAD, SSR, données INSEE)

Présentation des lauréats AMI BOAS 8

#8 nouveaux projets lauréats à l'AMI BOAS 8

MORS

HUGO Hopitaux Universitaires Grand Ouest

PULM-ALGO

ulmoTension Centre de Référence de l'Hypertension Pulmonaire

FHU COURNAND Transforming Pulmonary Vascular Medicine

CHU GRENOBLE ALPES

ALGO-DS

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Projet HEPALGO

Centre Eugène Marquis RENNES

Porteur.s : Centre Eugène Marquis (CEM) - CLCC de Rennes

Thématique 2 : Développement, validation et sophistication d'algorithmes de ciblage ou autres programmes visant à faciliter la réutilisation de la base principale du SNDS

Intervenant : Estelle NEVEU, Biostatisticienne au CEM

Objectifs :

- **Identifier le stade BCLC** (0-A ; B ; C ; D), **des patients primo-diagnostiqués d'un carcinome hépatocellulaire** à partir du SNDS.

Données mobilisées :

- La **cohorte nationale CHIEF en tant que gold standard** de cette étude.
- Les données de la **base principale du SNDS** : DCIR, PMSI, données sur les Causes de Décès

Présentation des lauréats AMI BOAS 8

#8 nouveaux projets lauréats à l'AMI BOAS 8

MORS

HUGO Hopitaux Universitaires Grand Ouest

PULM-ALGO

ulmoTension Centre de Référence de l'Hypertension Pulmonaire

FHU COURNAND Transforming Pulmonary Vascular Medicine

CHU GRENOBLE ALPES

ALGO-DS

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

HEPALGO

Centre Eugène Marquis RENNES

?

?

?

?

Projet Propolos

HeKA

Porteur.s : HeKA, INRIA

Thématique 2 : Développement, validation et sophistication d'algorithmes de ciblage ou autres programmes visant à faciliter la réutilisation de la base principale du SNDS

Intervenant : Anne Sophie Jannot

Objectifs :

- **Développement d'un jeu de données flouté du SNDS, anonyme et conservant les corrélations au sein de la base de données**, permettant de développer des méthodes pour analyser le parcours de soin des patients.

Données mobilisées :

- Données du **projet Drómos** permettant de résumer les données utiles du SNDS sur le parcours de soin en une cinquantaine d'indicateurs.

Présentation des lauréats AMI BOAS 8

#8 nouveaux projets lauréats à l'AMI BOAS 8

MORS

HUGO
Hopitaux Universitaires Grand Ouest

PULM-ALGO

ulmoTension
Centre de Référence de l'Hypertension Pulmonaire
FHU COURNAND
Transforming Pulmonary Vascular Medicine
CHU GRENOBLE ALPES

ALGO-DS

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

HEPALGO

Centre Eugène Marquis
RENNES

PROPOLOS

HeKA

Projet LOKAN

Porteur.s :
- **Institut National du Cancer (INCa)**
- **Registre des cancers de Poitou-Charentes**
- **Registre des cancers de la Gironde**

Thématique 2 : Développement, validation et sophistication d'algorithmes de ciblage ou autres programmes visant à faciliter la réutilisation de la base principale du SNDS

Intervenant : Sophie Houzard

Objectifs :

- **Développer et valider** des algorithmes de **détermination des localisations de cancer incident du sein, de prostate, du côlon-rectum et du poumon, ainsi que de leurs dates diagnostiques, de leur comportement invasif ou in situ.**

Données mobilisées :

- **Données PMSI et socles des registres Poitou-Charentes et de la Gironde.**
- Données de la **Cohorte Cancer**, extrait du SNDS sur les patients atteints ou à risque de cancer en France, hébergées sur la Plateforme de Données en Cancérologie de l'INCa

Présentation des lauréats AMI BOAS 8

#8 nouveaux projets lauréats à l'AMI BOAS 8

MORS

HUGO
Hopitaux Universitaires Grand Ouest

PULM-ALGO

ulmoTension
Centre de Référence de l'Hypertension Pulmonaire
FHU COURNAND
Transforming Pulmonary Vascular Medicine
CHU GRENOBLE ALPES

ALGO-DS

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

HEPALGO

Centre Eugène Marquis
RENNES

PROPOLOS

HeKA

LOKAN

INSTITUT NATIONAL DU CANCER
REGISTRE DES CANCERS POITOU-CHARENTES
Registre Général des CANCERS de la Gironde

Projet PENELOPALGO

Porteur.s : AP-HP

Thématique 3 : Développement, validation et sophistication d'algorithmes de ciblage ou autres programmes visant à faciliter la réutilisation d'une autre base de santé

Intervenant : Dr Christel Daniel

Objectifs :

- Développer de **nouveaux algorithmes de ciblage de patients potentiellement éligibles à la participation à des recherches dans les domaines des cancers colo-rectaux et ORL**.

Données mobilisées :

- **Données de l'EDS de l'AP-HP**

Présentation des lauréats AMI BOAS 8

#8 nouveaux projets lauréats à l'AMI BOAS 8

MORS

HUGO Hopitaux Universitaires Grand Ouest

PULM-ALGO

ulmoTension Centre de Référence de l'Hypertension Pulmonaire

FHU COURNAND Transforming Pulmonary Vascular Medicine

CHU GRENOBLE ALPES

ALGO-DS

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

HEPALGO

Centre Eugène Marquis RENNES

PROPOLOS

HeKA

LOKAN

INSTITUT NATIONAL DU CANCER

REGISTRE DES CANCERS POITOU-CHARENTES

Registre Général des CANCERS de la Gironde

PENELOPALGO

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

?

Projet FALCON

HUGO
Hopitaux Universitaires Grand Ouest

Porteur.s : Groupe de coopération sanitaire HUGO

Thématique 3 : Développement, validation et sophistication d'algorithmes de ciblage ou autres programmes visant à faciliter la réutilisation d'une autre base de santé

Intervenant : Morgane Pierre-Jean, PhD

Objectifs :

- **Développer un outil capable de restituer et visualiser**, de manière compréhensible, **des indicateurs portant sur les dérives temporelles** des données de biologie et le monitoring des tests réalisés.

Données mobilisées :

- **EDS du CHU de Rennes**
- **EDS d'un autre établissement HUGO** parmis les CHU d'Angers, de Brest, de Nantes ou de Tours
- **Base HUGOSHARE**

Présentation des lauréats AMI BOAS 8

#8 nouveaux projets lauréats à l'AMI BOAS 8

MORS

HUGO Hopitaux Universitaires Grand Ouest

PULM-ALGO

ulmoTension Centre de Référence de l'Hypertension Pulmonaire

FHU COURNAND Transforming Pulmonary Vascular Medicine

CHU GRENOBLE ALPES

ALGO-DS

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

HEPALGO

Centre Eugène Marquis RENNES

PROPOLOS

HeKA

LOKAN

INSTITUT NATIONAL DU CANCER

REGISTRE DES CANCERS POITOU-CHARENTES

Registre Général des CANCERS de la Gironde

PENELOPALGO

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

FALCON

HUGO Hopitaux Universitaires Grand Ouest

Félicitations aux équipes lauréates

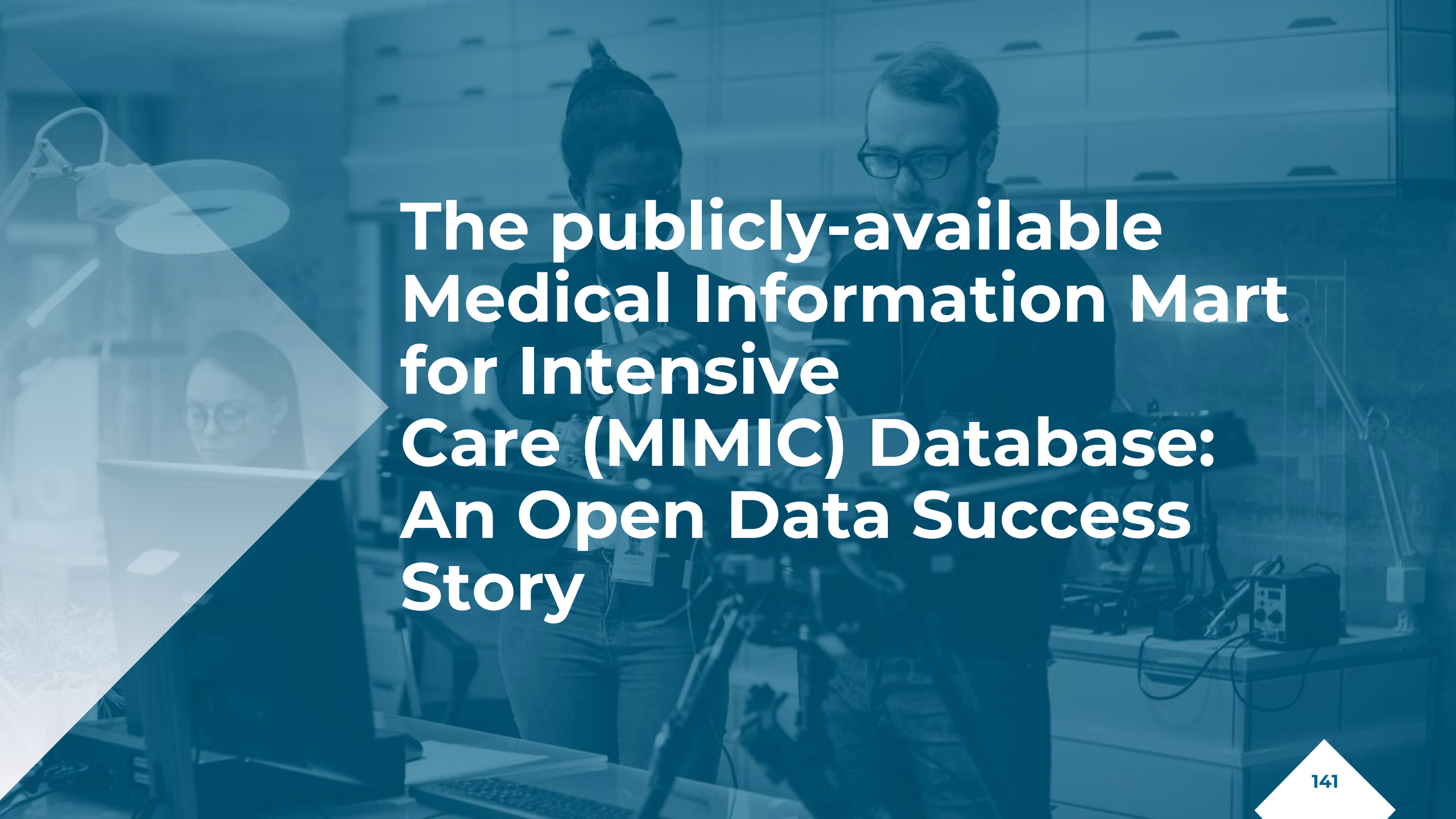
HEALTH DATA HUB

PAUSE

12h30 - 14h00

Journée de l’open science en santé - Programme

14h00 - 14h25	**The publicly-available Medical Information Mart for Intensive Care (MIMIC) Database: An Open Data Success Story**, Leo Anthony Celi (MIT)
14h25 - 14h50	**Les initiatives du Health Data Hub en faveur de l'ouverture de la science**, Laurie Alla (HDH)
14h50 - 15h05	**Le programme Data Challenges en santé, catalyseur d’innovations ouvertes**, Lauriane Armand (HDH)
15h05 - 15h20	**Le Data Challenge CytologIA - Améliorer le diagnostic en hématologie biologique grâce à l’IA**, Dr Samy Dahmani (Algoscope)
15h20 - 15h40	**Données ouvertes et IA en hématologie biologique : quelles perspectives après le Data Challenge CytologIA ?**, Dr Thomas Boyer (GFHC)
15h40 - 16h00	**Pause**
16h00 - 17h00	**Présentation des résultats et remise des prix du Data Challenge CytologIA**
17h00 - 17h20	**L'open data, un catalyseur de l'engagement citoyen au service de la santé**, Augustin Courtier (Latitudes)
17h20 - 17h30	**Conclusion et remerciements**
17h30 - 19h00	**Cocktail**



The publicly-available Medical Information Mart for Intensive Care (MIMIC) Database: An Open Data Success Story

HEALTH DATA HUB

The publicly-available Medical Information Mart for Intensive Care (MIMIC) Database: An Open Data Success Story

14h00 - 14h25

Leo Anthony Celi

MD MS MPH - Research director and principal research scientist at the MIT Laboratory for Computational Physiology (LCP)

Dr. Celi is the principal investigator behind the Medical Information Mart for Intensive Care (MIMIC) and its offsprings, MIMIC-CXR, MIMIC-ED, MIMIC-ECHO, and MIMIC-ECG. With close to 100k users worldwide, an open codebase, and close to 10k publications in Google Scholar, the datasets have undoubtedly shaped the course of machine learning in healthcare in the United States and beyond. His group has written 3 open-access textbooks: “Secondary Analysis of Electronic Health Records” in 2016, “Global Health Informatics: Principles of eHealth and mHealth to Improve Quality of Care” in 2017, and “Leveraging Data Science for Global Health” in 2020. The first has been downloaded over 1.7 million times and translated into Mandarin, Spanish, Korean and Portuguese. The group has created two open online courses, “Global Health Informatics” and “Collaborative Data Science for Healthcare”. Finally, in partnership with hospitals, universities and professional societies across the globe, Dr. Celi and his team have organized over 50 datathons in 22 countries, bringing together students, clinicians, researchers, and engineers to leverage data routinely collected in the process of care.

Science by Us, for All of Us: From Open Science to Community Science and the role of Artificial Intelligence

Leo Anthony Celi

On behalf of MIT Critical Data (we/us)

- No significant disclosure pertinent to this presentation.
- Research support received from the National Institute of Health (US), National Science Foundation (US) and Ministry of Health & Welfare (South Korea)

What was wrong with science pre-AI?

Can we fix it with AI? How?

More importantly, who can fix it?

A Case Study of Epistemic Injustice

The slowly evolving truth about heart disease and women

By Laura Williamson, American Heart Association News

(melitas/iStock via Getty Images)

THE LANCET
Digital Health

VIEWPOINT · Volume 2, Issue 9, E489-E492, September 2020 · *Open Access* Download Full Issue

The myth of generalisability in clinical research and machine learning in health care

Joseph Futoma, PhD [a] · Morgan Simons, MD [b] · Trishan Panch, MD [c,e] · Finale Doshi-Velez, PhD [a,†] · Leo Anthony Celi, MD [d,f,g,†]

Affiliations & Notes Article Info

a School of Engineering & Applied Sciences, Harvard University, Cambridge, MA, USA
b Department of Medicine, NYU Langone Health, New York, NY, USA
c Department of Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA
d Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
e Wellframe, Boston, MA, USA
f Division of Pulmonary, Critical Care, and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
g Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Health Research Policy and Systems

Research | Open access | Published: 15 May 2020

Exploring why global health needs are unmet by research efforts: the potential influences of geography, industry and publication incentives

Alfredo Yegros-Yegros, Wouter van de Klippe, Maria Francisca Abad-Garcia & Ismael Rafols

Health Research Policy and Systems **18**, Article number: 47 (2020) | Cite this article

How the creation and validation of knowledge will be affected by ~~AI~~

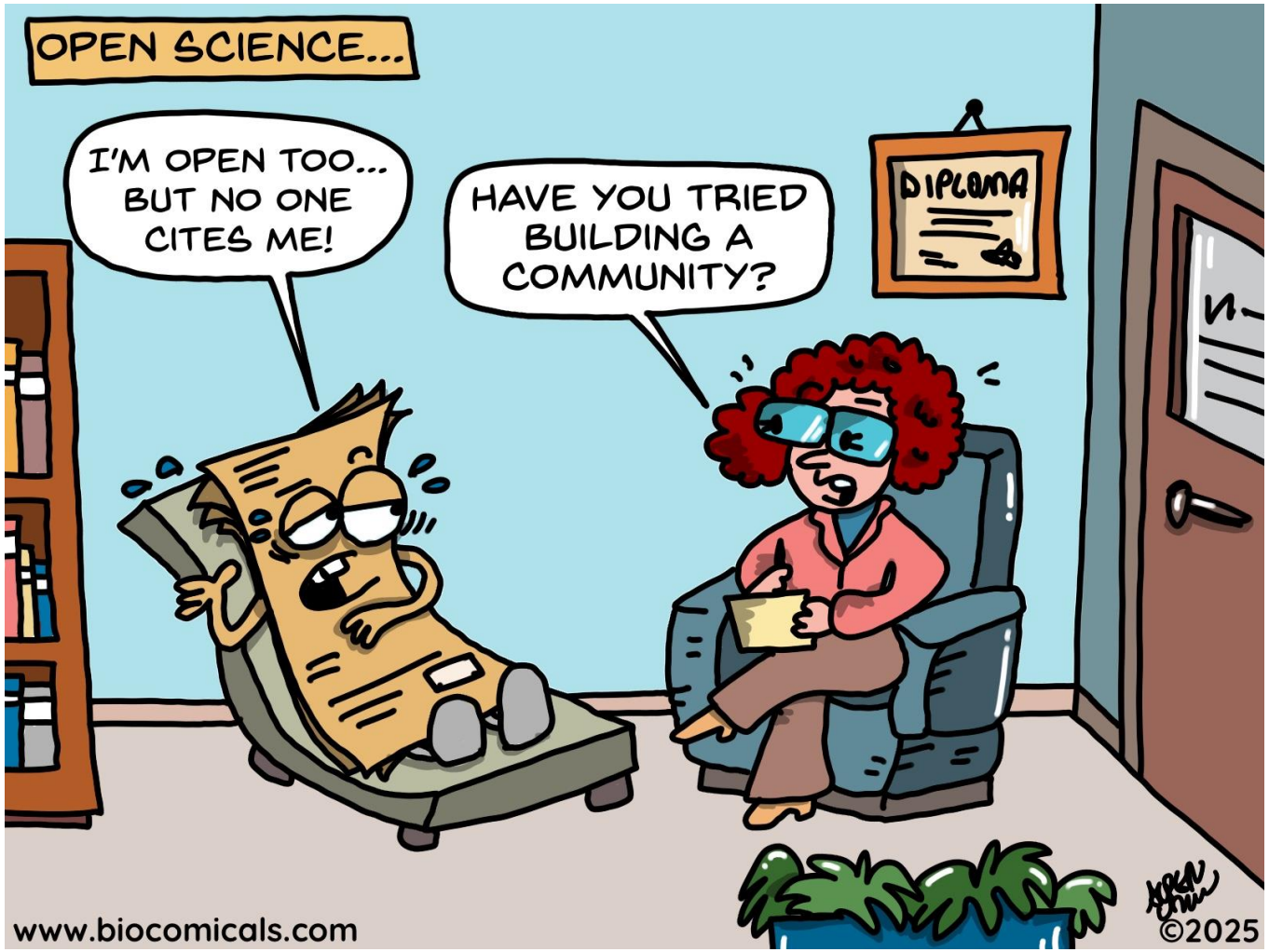
how we change the way we think, the way we learn, the way we work together

iScience. 2021 Jun 25; 24(6): 102656.
Published online 2021 Jun 10. doi: 10.1016/j.isci.2021.102656
PMCID: PMC8209268
PMID: 34169236

Village mentoring and hive learning: The MIT Critical Data experience

Christopher V. Cosgriff,[1] Marie Charpignon,[2] Dana Moukheiber,[3] Mary E. Lough,[4] Judy Gichoya,[5] David J. Stone,[6] and Leo Anthony Celi[7,8]

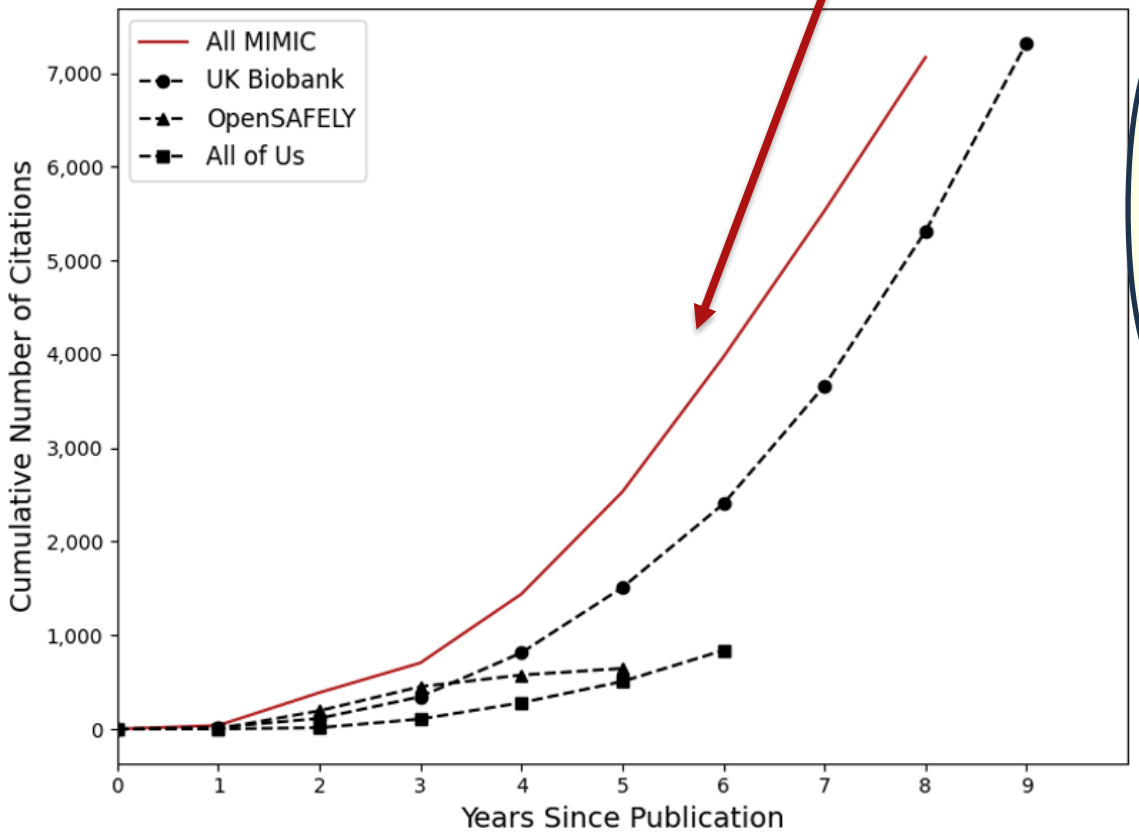
From Open Science to Community Science

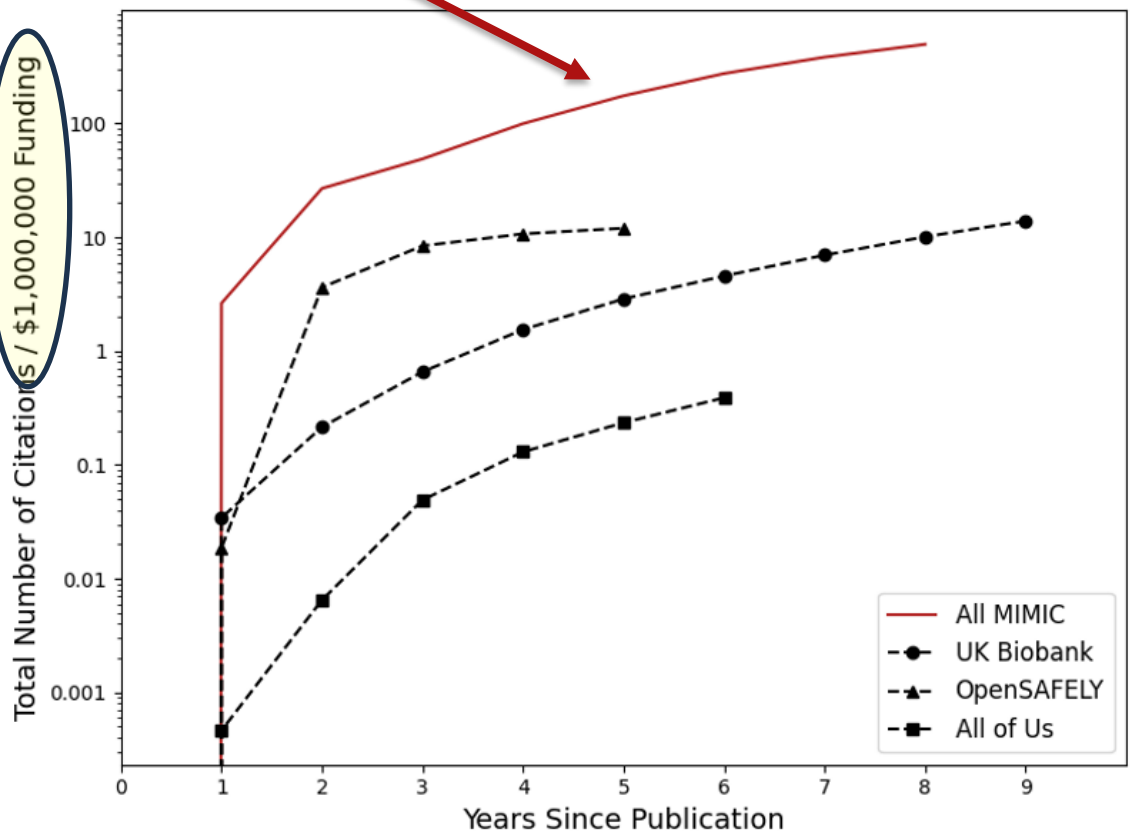


From Open Science to Community Science

MIMIC-III, a freely accessible critical care database AEW Johnson, TJ Pollard, L Shen, LH Lehman, M Feng, M Ghassemi, ... Scientific data 3 (1), 1-9, 2016	8464
The eICU Collaborative Research Database, a freely available multi-center database for critical care research TJ Pollard, AEW Johnson, JD Raffa, LA Celi, RG Mark, O Badawi Scientific data 5 (1), 1-13, 2018	1527
MIMIC-IV, a freely accessible electronic health record dataset AEW Johnson, L Bulgarelli, L Shen, A Gayles, A Shammout, S Horng, ... Scientific data 10 (1), 1, 2023	1431
The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care M Komorowski, LA Celi, O Badawi, AC Gordon, AA Faisal Nature medicine 24 (11), 1716-1720, 2018	1229

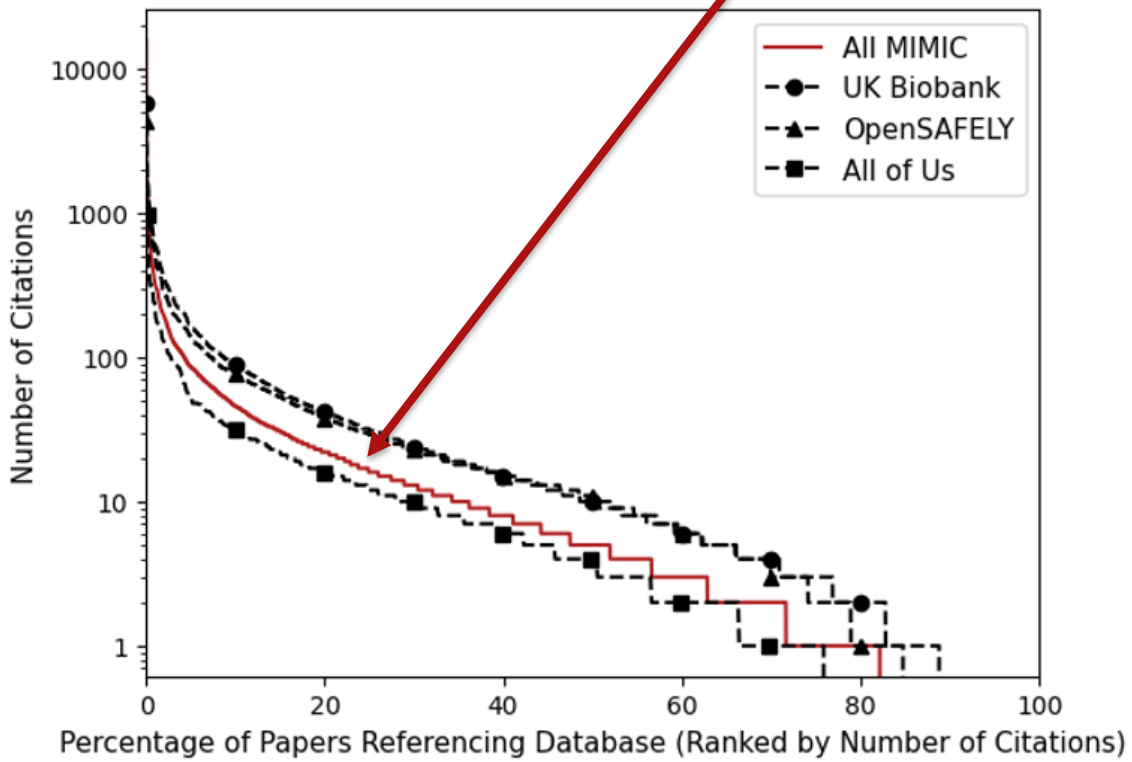
MIMIC-I / -II / -III / -IV

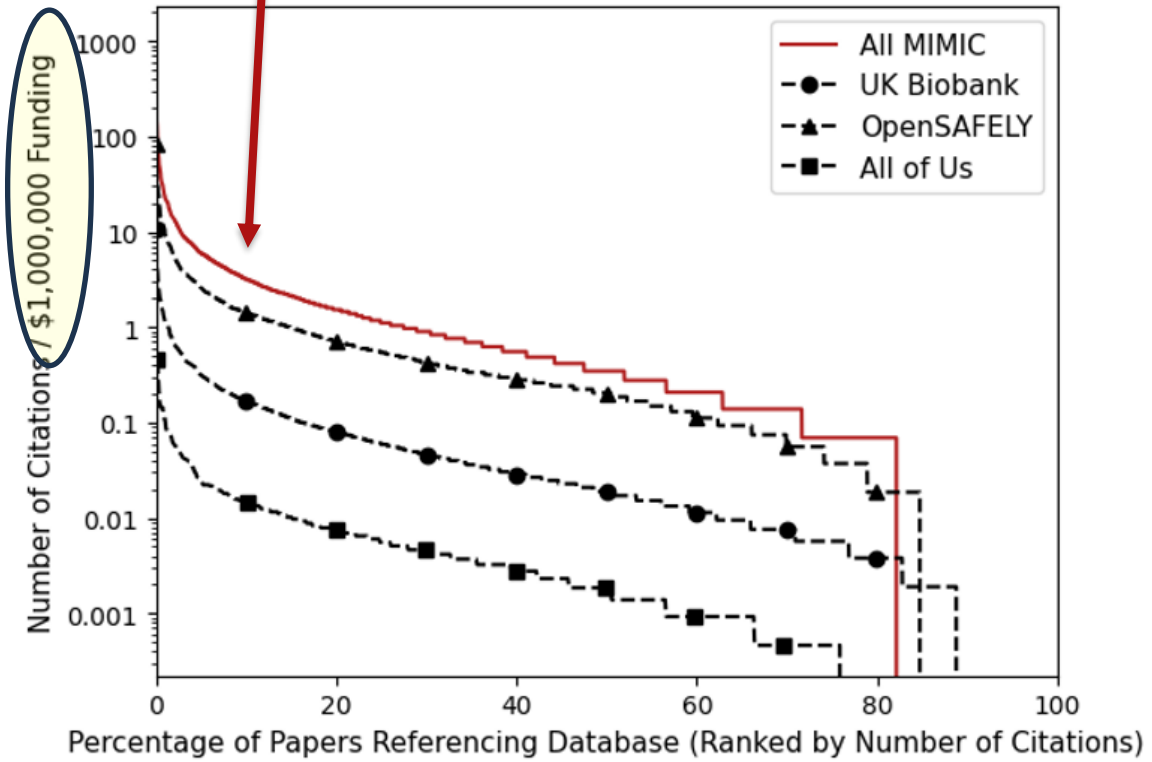




- Total citations since public release
- MIMIC = highest public research generation / $

MIMIC-I / -II / -III / -IV





- "Database h-index" for papers citing each database
- Normalized by total amount of citing papers
- MIMIC = highest public research impact / $ / paper

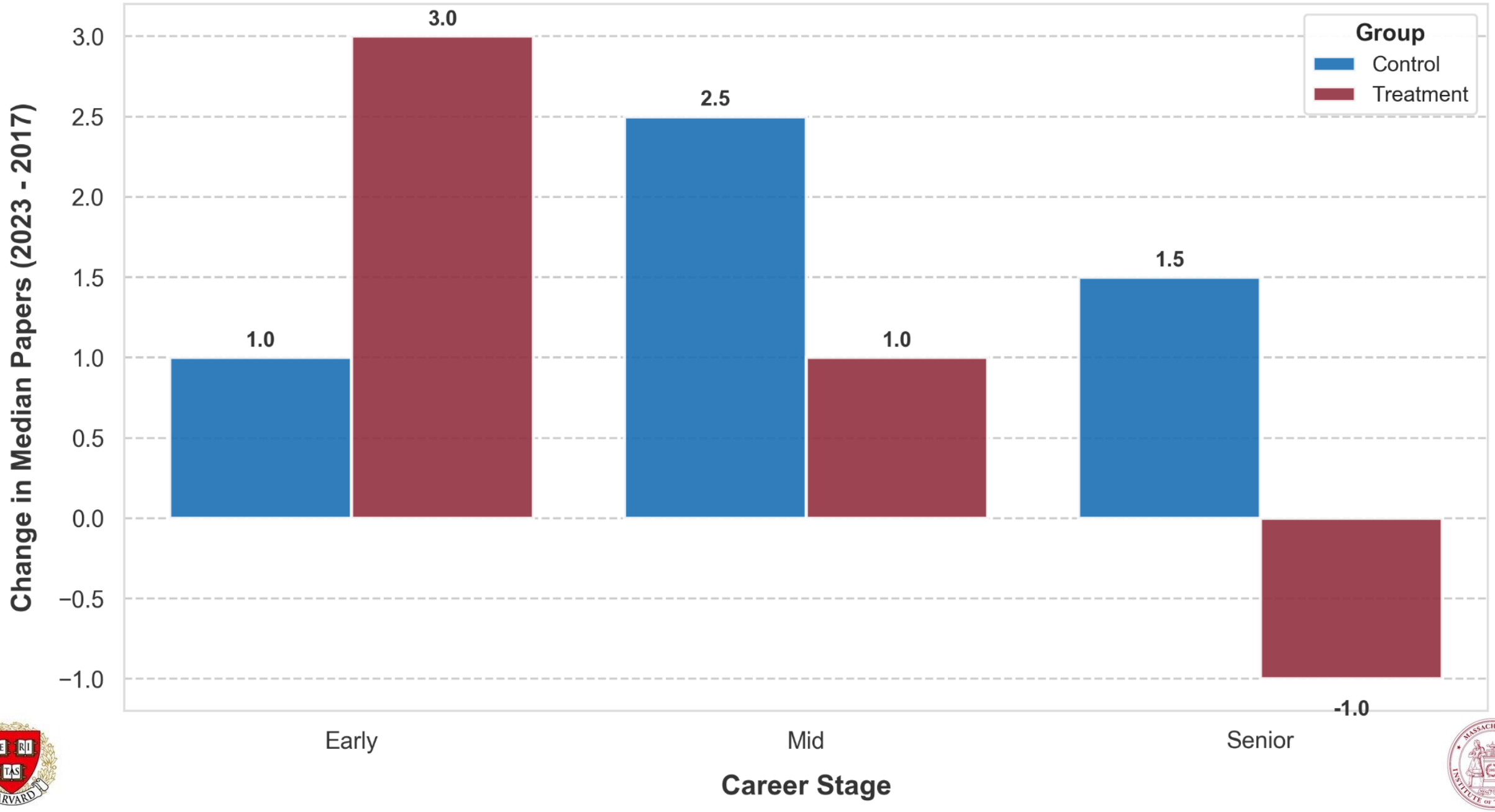
MIT Critical Data

LTARC Real-World Evaluation of LLMs

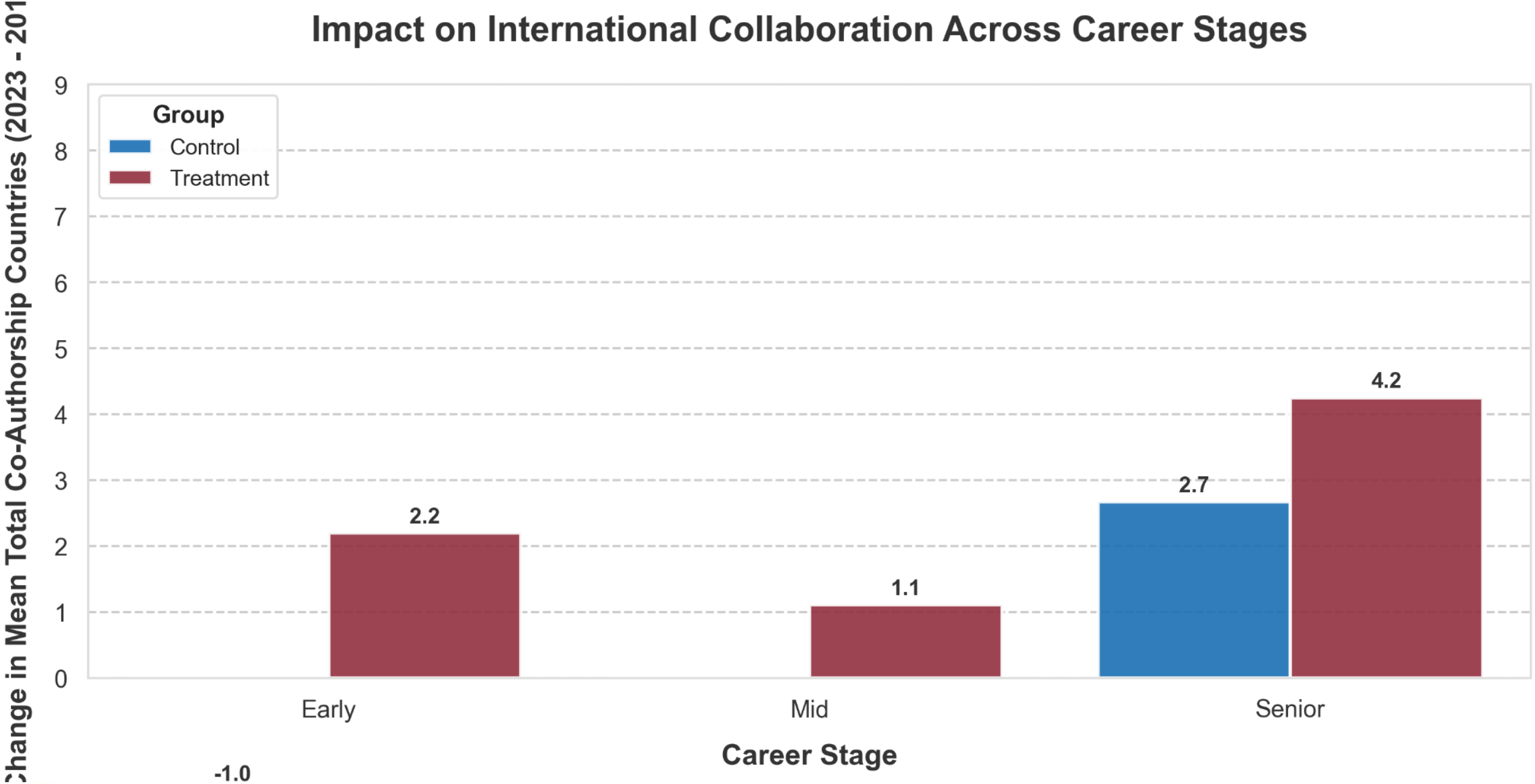
- Local
- Task-Specific
- Agile
- Reflective & Reflexive
- Community-operated & Continuous

Health AI Systems Thinking for Community (HASTC) Policy Camps

Impact on Publication Output Across Career Stages

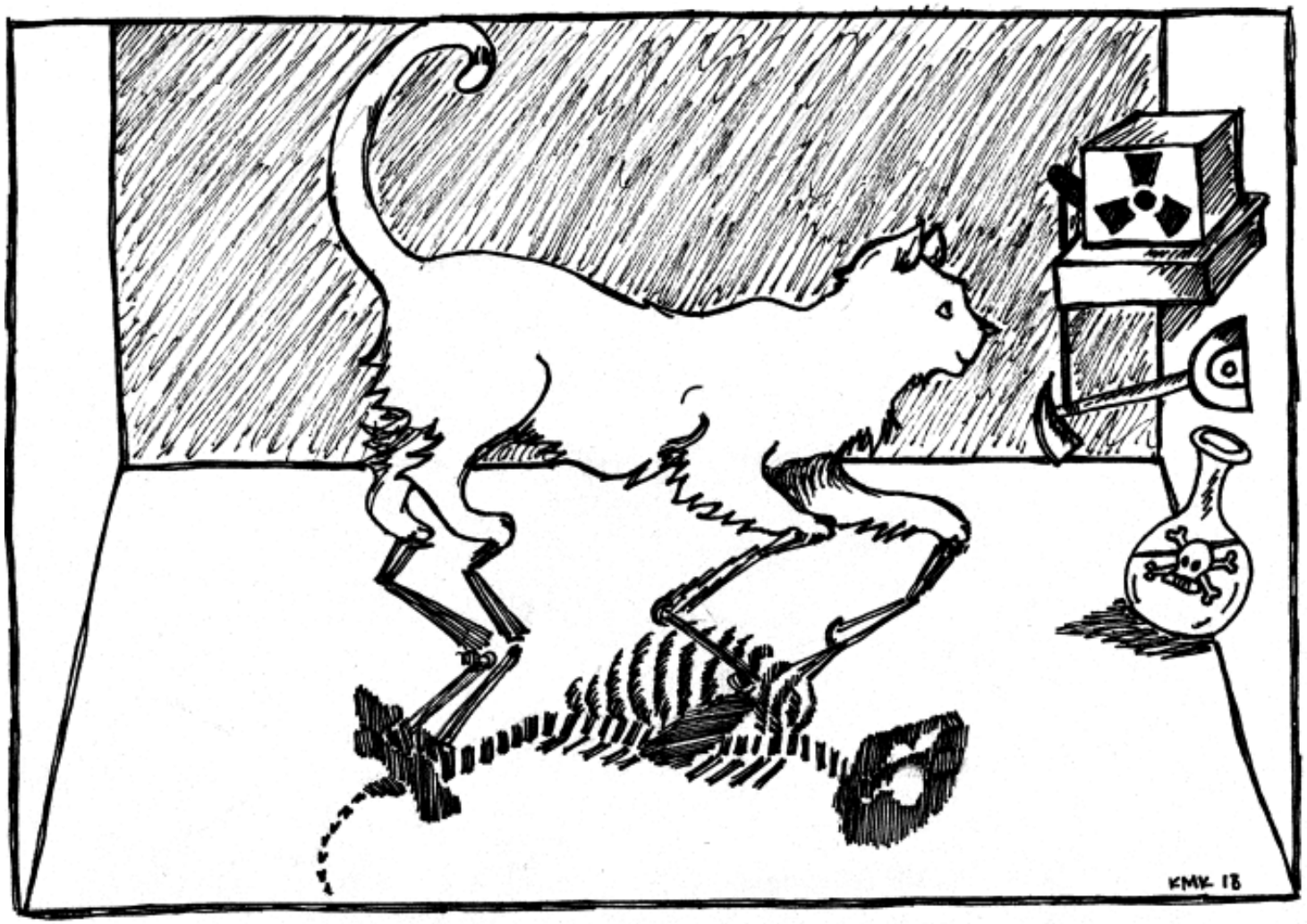


Impact on International Collaboration Across Career Stages

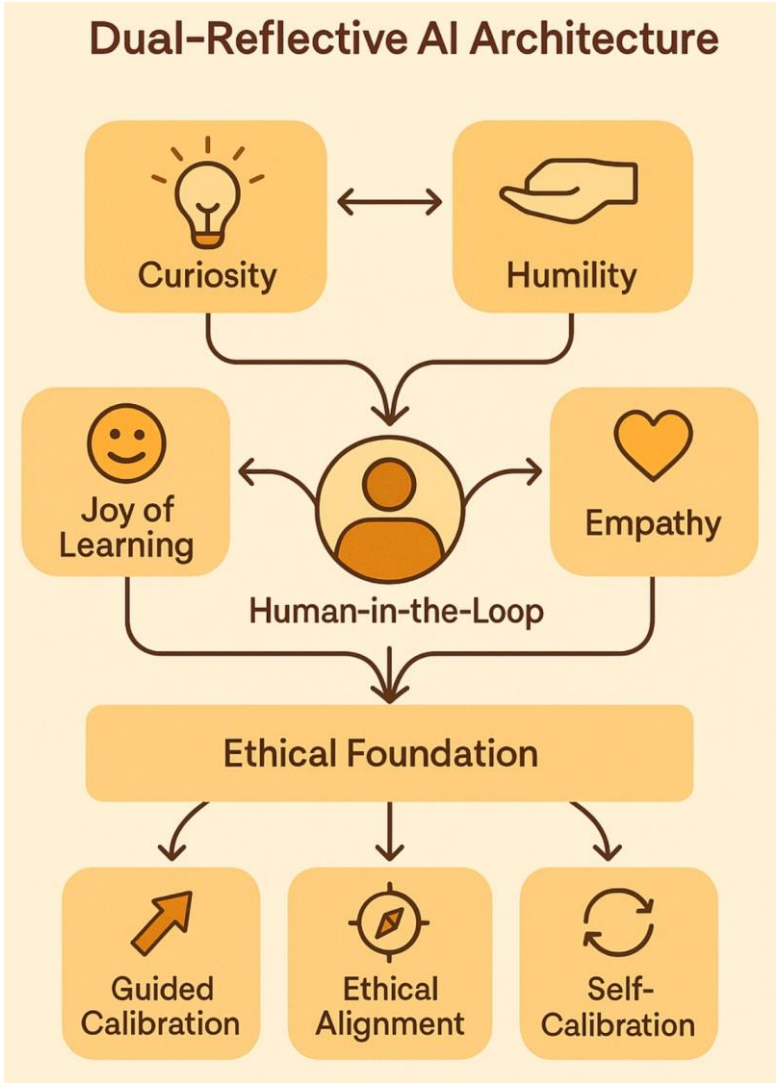


What do we teach about ~~AI~~ knowledge creation & validation?

- Health system science / systems thinking / critical thinking
- Epistemic humility and pursuit of plurality
- “When everyone is thinking alike, no one is truly thinking.”
- Regular reflection on what motivates scientific thinking: publications & citations, validation & ego, professional identity
- Reflection only works when done amid those who think differently (and when there is psychological safety to challenge each other)


KMK 18

Designing Human-AI Systems



Les initiatives du Health Data Hub en faveur de l'ouverture de la science

HEALTH DATA HUB

Les initiatives du Health Data Hub en faveur de l'ouverture de la science

14h25 - 14h50

Laurie Alla

Cheffe de projet Open Science
au Health Data Hub

Au sein de la direction scientifique du Health Data Hub, Laurie pilote les actions liées à la stratégie d'Open Science, avec pour objectif de promouvoir une science ouverte, collaborative et accessible dans le domaine de la santé. Elle contribue activement à la valorisation et à la diffusion des résultats issus des projets partenariaux du HDH, notamment issus de l'initiative Data Challenge en santé, qui vise à mobiliser la communauté autour de cas d'usage concrets et innovants.

Introduction

*Afin de répondre au **besoin d'ouverture dans le domaine des données de santé**, le Health Data Hub a créé, en janvier 2024, le pôle Open Science*

Action Open Source du HDH :

- **Promotion** de l'Open Source,
- **Coordination** de la mise en **Open Source** des ressources internes et des partenaires du HDH.

Action Open Data du HDH :

- **Promotion** de l'Open Data
- **Développement de solutions** pour faciliter l'**Open Data**

Action Data Challenge du HDH :

- Organisation de projets **Data Challenges**,
- Accompagnement de la mise en **Open Data** des **bases de données anonymisées**,
- **Référencement** des **algorithmes** publiés en **Open Source**.

depuis janvier 2024

Le pôle Open Science du HDH :

- Participation aux **travaux nationaux relatifs à la science ouverte** dans le domaine des données de **santé** ;
- **Organisation d'événements fédérateurs** pour rassembler les acteurs de l'innovation et contribuant à l'**ouverture des connaissances** dans le domaines des données de santé ;
- Développement de **solutions Open Source et Open Data** pour la communauté des donnés de santé

Plan

I. L’offre de service open science du Health Data hub

- L’anonymisation des données de santé, un pré-requis obligatoire à l’open data
- Les étapes successives de l’offre de service

I. Etat des lieux sur les contributions open science du Health Data Hub et de ses partenaires

- Contributions open data sur la plateforme publique data.gouv.fr
- Contributions open source à la BOAS et perspectives via le comité de suivi

I. Les données synthétiques : une solution face aux obstacles liés aux données de santé ?

#1

L’offre de service open science du Health Data Hub

L'offre de service open science du Health Data Hub

L'anonymisation des données de santé, un pré-requis obligatoire à l'open data

*Une solution d'anonymisation doit être construite **au cas par cas** par le partenaire et adaptée aux usages prévus. Pour aider à évaluer la bonne anonymisation d'un jeu de données, le **G29** a dégagé **trois critères** :*

Individualisation

Il ne doit pas être possible d'**isoler** une partie ou la totalité des enregistrements liés à un individu **dans le jeu de données**.

Corrélation

Il ne doit pas être possible de **relier deux enregistrements distincts** concernant un même individu (que les **bases** de données soient **distinctes ou non**).

L'inférence

Il ne doit pas être possible de **déduire**, avec un degré de probabilité élevé, de **nouvelles informations sur un individu.**

- Dans le cas où il n'est pas possible de conclure parfaitement à l'absence de ces trois critères, le responsable de traitement qui souhaite anonymiser un jeu de données doit démontrer, via une **analyse des risques d'identification**, que le risque de ré-identification avec des moyens raisonnables est **négligeable**.
- **Le caractère anonyme des données est sous la responsabilité juridique du partenaire, qui doit être en mesure de le démontrer.**

Documentation :

- Qu'est-ce qu'une donnée anonyme en santé ?
- Guide d'évaluation du caractère anonyme d'un jeu de données dans le cadre d'un projet de recherche

L’offre de service open science du Health Data Hub

Étapes successives à l’ouverture

Garantir l'interopérabilité des données et des algorithmes

Utiliser des **formats de données standardisés** et des **langages de programmation** largement utilisés pour **faciliter l'accès et l'utilisation** par le plus grand nombre :

Pour les **données** :

- **Tabulaires** : csv
- **Imageries** : tiff

Pour les **codes** :

- Python, SAS

L’offre de service open science du Health Data Hub

Étapes successives à l’ouverture

Garantir l'interopéra-bilité

Fournir une documentation détaillée

Cette **documentation** doit contenir les informations **nécessaires à la compréhension et la réutilisation :**

Pour les **données** :

- Les **métadonnées**,
- La **signification** des données,
- L’**origine**, le **contexte** et la **méthodologie** de collecte des données,
- La **structure** et le **format** des données,
- La **qualité** et **fiabilité** des données (mais aussi les erreurs potentielles et les **limites** du jeu de données),
- Les **mises à jour** et **évolutions**

Les métadonnées

Les métadonnées sont des informations permettant de **décrire une base de données** de manière descriptive et statistique.

Exemple:

- *Dictionnaire de variables,*
- *Description statistique des variables,*
- *Valeurs possibles ou intervalle,*
- *Unité de mesure,*
- *Structure de la base,*

L'offre de service open science du Health Data Hub

Étapes successives à l'ouverture

Garantir l'interopéra-bilité

Fournir une documentation détaillée

Cette **documentation** doit contenir les informations **nécessaires à la compréhension et la réutilisation :**

Des **templates** de **documentation standardisés** sont mis à disposition des porteurs souhaitant ouvrir un résultat afin de les guider dans cette étape

Pour les **données** :

- Les **métadonnées**,
- La **signification** des données,
- L'**origine**, le **contexte** et la **méthodologie** de collecte des données,
- La **structure** et le **format** des données,
- La **qualité** et **fiabilité** des données (mais aussi les erreurs potentielles et les **limites** du jeu de données),
- Les **mises à jour** et **évolutions**

Pour les **codes** :

- **Titre** du projet et **auteurs**,
- **Objectifs** de l'algorithme,
- **Méthodologie**,
- **Données** utilisées,
- **Validation**,
- Date de **mise à jour,**
- **Maintenance**,
- Comment installer et utiliser l'algorithme

L’offre de service open science du Health Data Hub

Étapes successives à l’ouverture

Garantir l'interopéra-bilité

Fournir une documenta-tion détaillée

Attribuer un DOI pour garantir la parentalité

DOI : Digital Object Identifier

- Fournit un **lien stable** vers un objet scientifique et sa description
- Il s’agit d’un **identifiant numérique unique et pérenne** pour un objet scientifique que l'on souhaite rendre **citable**

Un DOI peut être attribué à l’oeuvre à ouvrir via la plateforme fabrica de DataCite sur laquelle le HDH a un compte

L'offre de service open science du Health Data Hub

Étapes successives à l'ouverture

Garantir l'interopérabilité → **Fournir une documentation détaillée** → **Attribuer un DOI pour garantir la parentalité** → **Choisir une licence**

La licence définit **les modalités de réutilisation** d'une oeuvre et de **distribution** des versions modifiées.

Soumettre un contenu à plusieurs licences permet de:

✓ **élargir les compatibilités**

✓ **profiter des communautés** utilisatrices des deux licences

✓ **protéger la distribution** du contenu dans l'éventualité où une juridiction nationale annule une des deux licences

Pour les **données** : nous recommandons l'**utilisation conjointe** de **deux licences** :

Etalab 2.0 & **CC-BY**

Pour les **codes** : nous recommandons les licences :

- **Apache 2.0 ou**
- **Licence MIT**

Ces licences sont **permissives,** autorisées aux **administrations** et permettent une **réutilisation libre, gratuite et sans restriction** dans le **respect de la citation des auteurs**

Documentation : ✓Outil de la Commission Européenne

L'offre de service open science du Health Data Hub

Étapes successives à l'ouverture

Garantir l'interopérabilité → **Fournir une documentation détaillée** → **Attribuer un DOI pour garantir la parentalité** → **Choisir une licence** → **Choisir un outil de diffusion**

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
data.gouv.fr

Le catalogue des données françaises

63,9k	351,1k	5,5k
JEUX DE DONNÉES ET API	FICHIERS	ORGANISATIONS

Une communauté dynamique et engagée

4,4k	123,2k	15,4k
RÉUTILISATIONS	UTILISATEURS	DISCUSSIONS

Pour les **données** : nous recommandons de référencer les bases de données via la **plateforme nationale data.gouv.fr**

- Si la volumétrie de la base est **≤ 1 Go**, celle-ci peut directement être **hébergée sur data.gouv.fr**
- Si la volumétrie de la base est **> 1 Go**, la base doit être **hébergée sur une infrastructure propre**

Le Health Data Hub met à disposition des porteurs une infrastructure de stockage

L'offre de service open science du Health Data Hub

Étapes successives à l'ouverture

Garantir l'interopérabilité → **Fournir une documentation détaillée** → **Attribuer un DOI pour garantir la parentalité** → **Choisir une licence** → **Choisir un outil de diffusion**

Pour les **algorithmes** : nous proposons un **référencement** dans la **Bibliothèque Ouverte d'Algorithmes en Santé (BOAS), un projet collaboratif ouvert pour faciliter le traitement des données de santé**

Codes et algorithmes facilitant l'étude des données de santé **issus de tous type de projet** (incluant les codes et algorithmes développés dans le cadre de l'**AMI BOAS**)

L'Appel à Manifestation d'Intérêt BOAS soutient et encourage le **développement, la mise à jour et la validation d'outils informatiques sur les données de santé**

La BOAS est une bibliothèque centralisée d'algorithmes en santé partagés en Open Source

Une **bibliothèque centralisée** d'algorithmes en santé

Une **initiative collaborative** entre les différents **acteurs** de l'écosystème

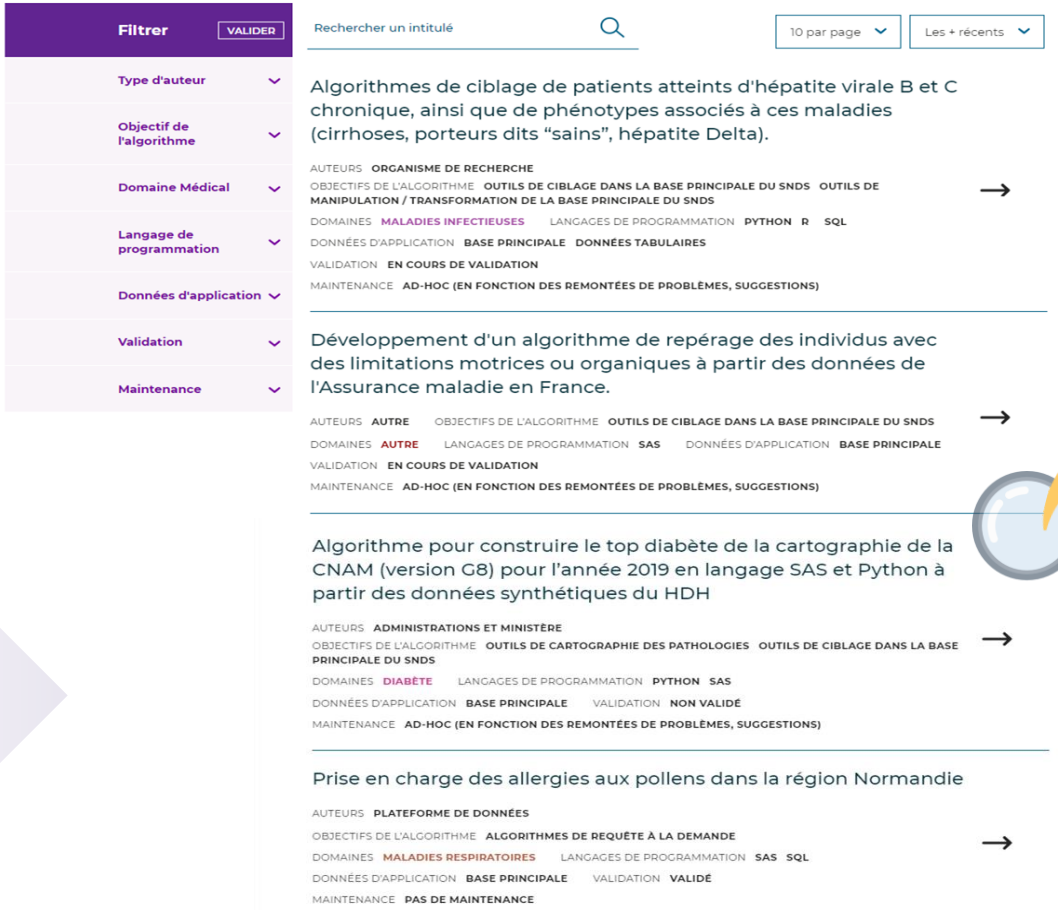
Partage des **bonnes pratiques**

Un **accompagnement** dans l'ouverture de codes sources

Focus sur la BOAS

A quoi ressemble cette bibliothèque ?

Bibliothèque Ouverte d'Algorithmes en Santé (BOAS)



Algorithme pour construire le top diabète de la cartographie de la CNAM (version G8) pour l’année 2019 en langage SAS et Python à partir des données synthétiques du HDH

Lien vers le repo : Gitlab

Objectifs de l’algorithme

OUTILS DE CARTOGRAPHIE DES PATHOLOGIES OUTILS DE CIBLAGE DANS LA BASE PRINCIPALE DU SNDS

Objectifs et périmètre

L'algorithme ici présenté a pour objectif de cibler les personnes prises en charge pour un diabète dans la base principale du SNDS afin de créer le « Top Diabète » de la cartographie des pathologies créée et maintenue par la CNAM (version G8). Il s’appui sur le programme source partagé par la CNAM et a été conçu pour le périmètre suivant :

- Caractéristique(s) ciblée(s) : Diabète, quel que soit son type
- Perspective : Construction du top diabète de la cartographie de la CNAM
- Périmètre géographique : France entière
- Périmètre historique programme source: Années 2015 à 2019 (incluses)
- Périmètre historique programmes adaptés : Années 2018-2019
- Régimes retenus : Ensemble des régimes d'assurance maladie

Version de la cartographie des pathologies : G8

Le programme source en SAS de la CNAM tourne sur les données des années 2015 à 2019.

Les versions Python et SAS adaptés de ce programme portent sur des données synthétiques pour les années 2018-2019 mais peuvent être étendues à d’autres années.

Traduction du top diabète CNAM - version Python

Focus sur la BOAS

A quoi ressemble cette bibliothèque ?

GitLab

Health Data Hub / ... / CNAM / Top Diabete

Top Diabete

Ajouter aux favoris

main | top-diabete / | Historique | Rechercher un fichier | Modifier | Code

Update README.md
Laurie Alla rédigé il y a 15 heures — 0f4d7b03

Nom	Dernière validation	Dernière mise à jour
Version python	Téléverser un nouveau fichier	il y a une semaine
Version sas	Ajout versions sas	il y a une semaine
Programme_source_top_FDi...	Update Programme_source_-_top_FDi...	il y a une semaine
README.md	Update README.md	il y a 15 heures
Tables_et_variables_du_SN...	Téléverser un nouveau fichier	il y a une semaine

README.md

Top Diabete

Titre du projet :

Algorithme pour construire le top diabète de la cartographie de la CNAM (version G8) pour l'année 2019 en langage SAS et Python à partir des données synthétiques du HDH

Algorithme pour construire le top diabète de la cartographie de la CNAM (version G8) pour l'année 2019 en langage SAS et Python à partir des données synthétiques du HDH

Lien vers le repo : Gitlab

Objectifs de l'algorithme

OUTILS DE CARTOGRAPHIE DES PATHOLOGIES OUTILS DE CIBLAGE DANS LA BASE PRINCIPALE DU SNDS

Objectifs et périmètre

L'algorithme ici présenté a pour objectif de cibler les personnes prises en charge pour un diabète dans la base principale du SNDS afin de créer le « Top Diabète » de la cartographie des pathologies créée et maintenue par la CNAM (version G8). Il s'appui sur le programme source partagé par la CNAM et a été conçu pour le périmètre suivant :

- Caractéristique(s) ciblée(s) : Diabète, quel que soit son type
- Perspective : Construction du top diabète de la cartographie de la CNAM
- Périmètre géographique : France entière
- Périmètre historique programme source: Années 2015 à 2019 (incluses)
- Périmètre historique programmes adaptés : Années 2018-2019
- Régimes retenus : Ensemble des régimes d'assurance maladie

Version de la cartographie des pathologies : G8

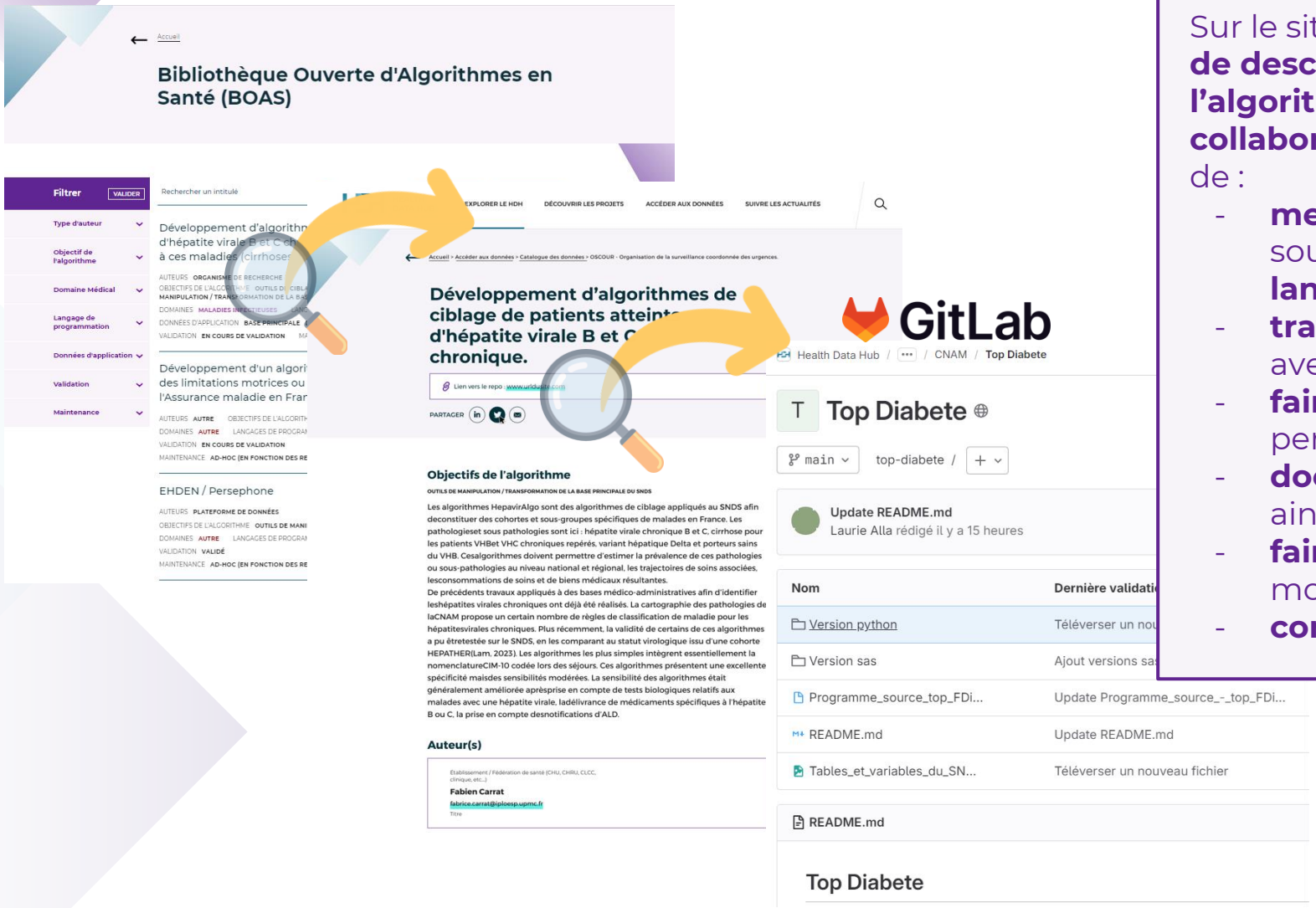
Le programme source en SAS de la CNAM tourne sur les données des années 2015 à 2019.

Les versions Python et SAS adaptés de ce programme portent sur des données synthétiques pour les années 2018-2019 mais peuvent être étendues à d'autres années.

Traduction du top diabète CNAM - version Python

Focus sur la BOAS

A quoi ressemble cette bibliothèque ?



Sur le site internet du Health Data Hub, chaque **fiche de description sur la BOAS** renvoie au **dépôt de l'algorithme** sur une **plateforme ouverte** et **collaborative** (telle que GitLab ou GitHub) permettant de :

- **mettre à disposition** des algorithmes et codes sources à la communauté **quelque soit le langage de programmation**,
- **travailler en collaboration** avec la communauté avec un système de tickets,
- **faire évoluer** les codes et algorithmes afin de les perfectionner,
- **documenter les changements** de version et ainsi **historiser les versions** des codes sources,
- **faire valider** (ou approuver) les propositions de modification,
- **comparer** les différentes versions.

#2

Etat des lieux sur nos contributions open science et celles de nos partenaires

Contributions à l'open data

Etat des lieux

Près de 6000 visites en 1 an sur notre page data.gouv.fr

Bases de données ouvertes

Données synthétiques de la base principale du Système National des Données de Santé (Health Data Hub) : jeu de données généré en utilisant le **schéma des tables de la base principale du SNDS** (50 patients fictifs).

Données synthétiques du Top Diabète (Health Data Hub) : jeu de données créée dans le cadre de la traduction et l'implémentation de l'algorithme utilisé par la CNAM pour construire le top diabète

Allergen Chip challenge (Société Française d'Allergologie) : base de données **tabulaire** correspondant au **profil immunologique et clinique de près de 3 000 patients.**

TissueNet (Société Française de Pathologie) : base de données d'imagerie composée de **1 000 lames de biopsie de col utérin numérisées**

Cytolog-IA (Groupement Francophone d'Hématologie Cellulaire, AlgoScope) : base de données d'**imagerie** composée de **70 000 images de frottis sanguins annotées issues de 23 centres hospitaliers francophones.**

DigiLut (Hôpital Foch) : base de données d'imagerie composée de **500 images de biopsie pulmonaire annotées.**

Bases de données en cours d'ouverture

VisioMel (Société Française de Pathologie) : base de données d'imagerie composée de **2 300 images numérisées d'exérèse de mélanomes localisés associée et données cliniques**.

Contributions à l'open data

Etat des lieux

Bases de données ouvertes

Données synthétiques de la base principale du SNDS (Health Data Hub)

- **2,6 k visites** (271 en mai 2025)
- **1,2 k téléchargements** (146 en mai 2025)
- Depuis l'ouverture (05/2024), en moyenne **227 visites/mois et 100 téléchargements**

Données synthétiques du Top Diabète (Health Data Hub)

- **2,2 k visites** (213 en mai 2025)
- **227 téléchargements** (10 en mai 2025)
- Depuis ouverture (05/2024), en moyenne **193 visites/mois et 16 téléchargements**

Allergen Chip challenge (Société Française d'Allergologie)

- **2,2 k visites** (158 en mai 2025)
- **296 téléchargements** (29 en mai 2025)
- Depuis l'ouverture (05/2024), en moyenne **202 visites/mois et 29 téléchargements**

TissueNet (Société Française de Pathologie)

- **1,2 k visites** (264 en mai 2025)
- **1,8 k téléchargements** (395 en mai 2025)
- Depuis ouverture (01/2025), **en moyenne 241 visites et 357 téléchargements**

DigiLut (Hôpital Foch)

Depuis ouverture le 23 mai 2025 :

- **30 visites**
- **20 téléchargements**

Contributions à l'open source

Etat des lieux de la bibliothèque

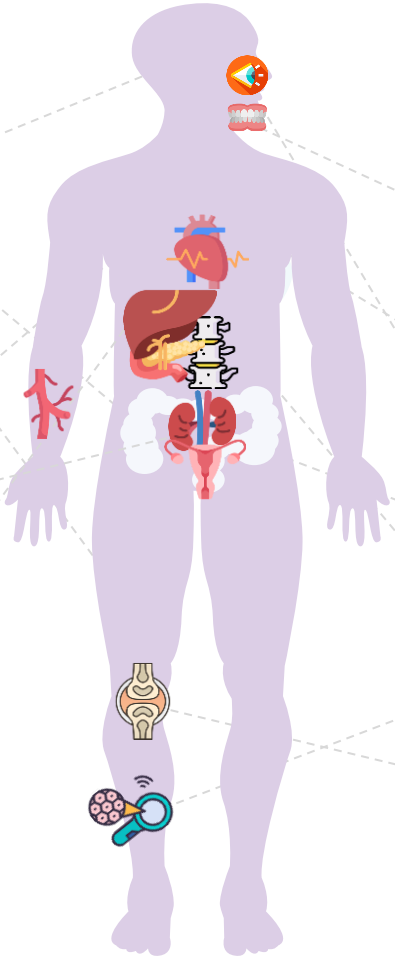
- **33 ressources** sont disponibles dans la BOAS depuis sa mise en ligne en avril 2024

Algorithmes de ciblage dans la base principale du SNDS

- **Top Diabète**
- **Requête type :** affection longue durée
- **Requête type :** prestation dans le DCIR
- **Requête type :** séjours hospitaliers
- **MICI :** rectocolite hémorragique
- **BIS : Asthme** de l'enfant
- **RISH :** limitations cognitives
- **RISH :** limitations motrices/handicap
- **DIONISOS :** sclérose en plaques
- **HepaVirAlgo :** hépatite B et C
- **Insuffisance rénale chronique**
- Distinction **diabète T1 vs T2**

Requêtes à la demande

- **Requête à la demande :** femmes ayant droits
- **Requête à la demande :** dépistage des cancers féminins
- **Requête à la demande :** socio-démographie
- **Requête à la demande :** recours à l'IVG
- **Requête à la demande :** allergies en Normandie



Algorithmes de traitement des données de santé

- **ALIA :** sous-types d'AVC hospitalisés
- **Allergologie :** prédiction des allergènes
- **Maladies rares** dentaires, orales et faciales
- **Cancer du col de l'utérus :** détection de lésions
- **Greffe pulmonaire :** prédiction du rejet
- **Mélanome :** prédiction de la récidive
- **Cytologie :** classification de leucocytes
- **Cartographie des pathologies**

Outils de manipulation, transformation des données

- **OMOP :** Standardisation de la base principale
- **Magic Loop :** génération des tables de consommation de soin
- **Deep.piste** : pseudonymisation de fichiers DICOM
- **Extract metadata** : extraction de métadonnées (données tabulaires et d'imagerie)
- **Macro SAS :** d'extraction de données SNDS
- **Librairies** de **manipulation de données sur les l'EDS**

En moyenne **380 visites/mois**

En cours de référencement

https://www.health-data-hub.fr/bibliotheque-ouverte-algorithmes-sante

183

Contributions à l'open science

Participation aux travaux nationaux avec l'AMDAC MSS

2024

Groupe de travail open source

- Un **guide d'accompagnement à l'ouverture des algorithmes et codes sources** a été produit dans l'optique de fournir des **fiches thématiques pour réussir une ouverture d'algorithmes/codes source.**

2025

Groupe de travail open data

- Des documents **pratiques** et **synthétiques** seront produits dans le but de servir de **guides de référence** permettant de **définir les principaux concepts de la mise en œuvre de la pseudonymisation et de l'anonymisation** des données et **proposer un état des lieux des principales méthodes** et des **lignes directrices** des cas dans lesquels ces méthodes sont pertinentes.

Convaincu que les **découvertes les plus impactantes seront issues de l'intelligence collective**, ces **travaux collaboratifs** impliquant des **acteurs majeurs** ont pour objectif **de soutenir et d'accompagner l'écosystème de la santé numérique dans la démarche open science.**

Contributions à l'open science

Organisation d'événements fédérateurs

Ces événements ont pour objectif de ***rassembler les acteurs de l'innovation et contribuant à l'ouverture des connaissances dans le domaine des données de santé***

Deuxième éditions de la journée open science en santé :

- La première édition avait rassemblée une centaine de personnes en présentiel et en ligne.
- Cette première édition avait permis de **mettre en lumière les bénéfices concrets du partage de ressources à travers des exemples inspirants**.

L'organisation régulière de meetups :

- Les Meetups sont des **rencontres ouvertes à tous** réunissant des **acteurs variés** dans le but d'échanger autour de thématiques telles que :
  - **Les avancées sur le SNDS,**
  - **L'ouverture en open source des travaux sur des données de santé**
  - **L'ouverture en open data de bases de données de santé,**
  - etc.
- Le groupe **Meetup**, lancé en 2019, rassemble aujourd'hui plus de 3 900 membres.

OMOPisation des données de santé : retour d'expérience et contrôle qualité

Inscription :

présentiel // distanciel

Tous les replays sont disponibles sur la page Youtube du Health data Hub

Contributions à l'open science

Les actions en faveur de la science ouverte au delà du pôle open science

Partager avec l'écosystème

- **Forum entraide :** permet de poser des questions et de poster des sujets de discussion ou annonces autour de la base principale du SNDS.

~ 611 utilisateurs/ utilisatrices

Visites/mois (moyenne) : **1 800**

- **BOAS :** bibliothèque centralisée d'algorithmes en santé partagés en open source

33 fiches créées en 1 an

Visites/mois (moyenne) : **380**

Se former

- **Catalogue de formations en accès libre :** procédures réglementaires, informations sur les données du SNDS et ses modalités d'accès, MOOC SNDS (4 modules)

16 formations disponibles

Visites/mois (moyenne) : **1 500**

Se documenter

- **Documentation SNDS collaborative** riche en ressources et informations autour des données de santé

70 fiches thématiques

Visites/mois (moyenne) : **13 000**

- **Dictionnaire interactif** développé par le Lab Santé de la DREES, cet outil permet la navigation par variable, table et nomenclature, schéma interactif illustrant les liens entre les tables et leurs clés de jointure, actuellement sur les métadonnées 2022

Visites/mois (moyenne) : **2 100**

#3

Les données synthétiques : une solution face aux obstacles liés aux données de santé ?

Les données synthétiques : une solution face aux obstacles liés aux données de santé ?

Les données synthétiques sont des données ***artificielles générées pour reproduire des caractéristiques statistiques de données réelles sans contenir d'informations personnelles.***

Dans un contexte où les **contraintes réglementaires** et **éthiques** freinent parfois l'accès et l'ouverture de données de santé, les données synthétiques représentent une **solution prometteuse** concernant les aspects de :

Confidentialité & conformité	Les données synthétiques permettent le **partage** de données **sans compromettre la vie privée** (toutes les données ne sont pas partageables)
Accessibilité et complémentarité	Elles permettent d'agir comme **levier pour ouvrir la science** là où les données réelles sont restreintes et peuvent être **complémentaires des données réelles** (par exemple en équilibrant des classes rares, créant des scénarios cliniques sous-représentés, etc.)
Innovation accélérée	Elles permettent également d'offrir des **environnements** de développement, test et validation **sans contrainte réglementaire ni risque** pour les patients

Les données synthétiques représentent un enjeu national de plus en plus important et font partie intégrante de notre feuille de route 2025.

Les données synthétiques : une solution face aux obstacles liés aux données de santé ?

Les données synthétiques sont des données ***artificielles générées pour reproduire des caractéristiques statistiques de données réelles sans contenir d'informations personnelles.***

Afin de mieux comprendre les **besoins**, **usages** et **attentes** autour de ces données, et d'**identifier des projets existants à valoriser**, le Health Data Hub et l'agence de programme santé numérique d'Inria lancent un **sondage** à destination des acteurs de l'écosystème.

Votre avis compte : contribuez à faire avancer l'usage des données synthétiques en santé en participant au sondage et/ou en le diffusant directement à celles et ceux qui sont concerné(s) par ces données !

Scan me!

Recrutement post doctorant

Générateur d'Imagerie Synthétique

Objectif du projet : ***Développer** un **générateur de données synthétiques complexes** pour l'**imagerie médicale**, à des fins **pédagogiques** et d'**entraînement d'IA**.*

Ce recrutement est intégré au **projet PFDS – Programme de Formation aux Données de Santé** - porté par l'Université Grenoble Alpes, dans le cadre de l'AMI **Compétences et Métiers d'Avenir**.

Profil recherché

- Profil spécialisé en **intelligence artificielle, apprentissage profond ou traitement de l'image**
- Expertise en **traitement d'images**, **génération de données synthétiques** et **entraînement de modèles basés sur l'apprentissage profond**
- Compétences en programmation (Python, TensorFlow/PyTorch)
- Sensibilité aux enjeux **éthiques et réglementaires** liés à l'utilisation des données de santé
- Aisance en travail collaboratif avec chercheurs et professionnels de santé

Missions principales

- **Concevoir une architecture de génération d'images médicales réalistes et de qualité**
- Contrôler la **complexité** et les **caractéristiques** des images produites
- Mettre en place des systèmes de **validation des données**

Cadre du recrutement

- **Durée :** 12 mois (renouvelable)
- **Lieu et encadrement :** Institut de recherche partenaire en collaboration étroite avec le Health Data Hub
- **Budget :** 102 600 € (HDH)
- **Convention entre HDH et l'institut d'accueil**

Pour toute question ou pré-manifestation d'intérêt, veuillez contacter l'équipe en charge du projet PFDS aux l'adresse suivante :
agathe.delaune@health-data-hub.fr et laurie.alla@health-data-hub.fr

Avez-vous des questions ?

Le programme Data Challenges en santé, catalyseur d'innovations ouvertes

HEALTH DATA HUB

Le programme Data Challenges en santé, catalyseur d'innovations ouvertes

14h50 - 15h05

Lauriane Armand

Cheffe de projet au Health Data Hub, PharmD

Au sein de la direction scientifique du HDH, Lauriane travaille sur la stratégie et la gestion de l'initiative Data Challenges en santé et accompagne les porteurs de projets dans toutes les étapes de l'organisation de la compétition, depuis le cadrage scientifique de la problématique jusqu'à la valorisation et l'ouverture des résultats.

Le programme Data Challenges en santé, catalyseur d'innovations ouvertes

Data Challenge

Approche innovante et participative

Enrichissement du catalogue de données de santé

Développement d'algorithmes d'assistance aux professionnels de santé

Démarche d'intelligence collective

Ouverture de la science

Le HDH est une institution promouvant **l'innovation par l'utilisation secondaire des données de santé**

Dans le cadre de sa mission d'animation de l'écosystème des données de santé, le **HDH en partenariat avec la Délégation au Numérique en Santé, le SGPI et Bpifrance, soutient et organise des Data Challenges en santé**

Ces initiatives font partie du programme **France 2030 et s'inscrivent dans la Stratégie d'Accélération Santé Numérique**

GOUVERNEMENT
Liberté
Égalité
Fraternité

FRANCE 2030

bpifrance

HEALTH DATA HUB

Qu'est-ce qu'un Data Challenge ?

Un Data Challenge est une **compétition en science des données** ouverte dont l'objectif est de **résoudre une problématique** spécifique de data science grâce à des solutions **d'apprentissage automatique**. Cette compétition se déroule **en ligne** et repose sur **un large jeu de données** mis à disposition par les organisateurs via une **plateforme dédiée** -la plus connue étant Kaggle-.

Pour servir de source d'information à l'analyse et permettre le développement d'un modèle

BASE DE DONNÉES A ANALYSER

Pour participer au challenge et tenter de remporter un prix

COMPÉTITEURS

PROBLÉMATIQUE

A laquelle on peut répondre par des techniques de data science/machine learning

PLATEFORME DE DATA CHALLENGE

Pour héberger la compétition et recruter des participants du monde entier

CLASSEMENT ET RÉCOMPENSE

Pour identifier les compétiteurs ayant conçu le meilleur modèle

Comment se présente un Data Challenge ?

Un Data Challenge est une compétition de data science se déroulant en ligne sur une **plateforme dédiée**.

La plateforme de Data Challenge permet aux participants de…

COMPRENDRE LA PROBLÉMATIQUE ET LES RÈGLES DU CHALLENGE

… grâce à une introduction pédagogique et à de la **documentation** sur les **enjeux** et **la problématique**, une présentation de la **métriques d'évaluation,** l'accès à un **règlement intérieur**

ACCÉDER AUX JEUX DE DONNÉES MIS À DISPOSITION ET À UN ESPACE DE CALCUL

… plusieurs **jeux de données** sont mis à disposition des compétiteurs pour qu'ils puissent les **analyser** et **développer leur modèle**. Les participants auront aussi la possibilité d'utiliser la **puissance de calcul** de la plateforme pour **entraîner** et **valider** leur algorithme

SE MESURER AUX AUTRES PARTICIPANTS EN AYANT ACCÈS À LEUR CLASSEMENT

… en **soumettant leurs algorithmes**, qu'ils soient intermédiaires ou finaux, afin de les **tester** et de visualiser leur classement dans un « **Leaderboard** » en temps réel

ÉCHANGER AVEC LES AUTRES PARTICIPANTS

… grâce à des **espaces collaboratifs** (ex. FAQ, Wiki, GitHub, accès aux notebooks des participants)

Le programme Data Challenges en santé

Data Challenge

2020

TissueNet
SFP

2021

AMI 2021
Sélection de
6 projets

AAP 2022
Sélection de
3 projets

2022

2023

VisioMel
SFP - SFD

Allergen Chip
SFA

D-IA-GNO-DENT
HUS

2024

AAP 2023 - 2025
Sélection de
5 projets

DigiLut
Hôpital Foch

CytologIA
GFHC - Algoscope

2025

AAP 2023 - 2025
Relève le
30 juin

2026

Les Data Challenges en santé finalisés

Data Challenge

TISSUENET

Terminé

Détection de lésions cancéreuses du col de l'utérus
Données de plus de **4 000** patientes
2020 - 550 participants

VISIOMEL

Terminé

Prédiction de l'évolution métastatique de mélanomes
Données de plus de **2 000** patients
2023 - 550 participants

ALLERGEN CHIP

Terminé

Prédiction des allergènes en cause d'une allergie
Données de plus de **4 000** patients
2023 - 300 participants

D-IA-GNO-DENT

Terminé

Aide au diagnostic de maladies bucco-dentaire rares
Données de **316** patients
2023 - 90 participants

DIGILUT

Terminé

Détection de rejet de greffe pulmonaire
Plus de **2 500** lames de biopsie de greffon
2024 - 260 participants

CYTOLOGIA

Terminé

Automatisation de la lecture de frottis sanguin
Plus de **70 000** images de leucocytes
2025 - 250 participants

Les Data Challenges en santé à venir

Data Challenge

DAT-HUB

SFMN

Détection de la maladie de Parkinson
Plus de **2 000** examens de patients
Lancement en **2025**

CARDI-HACK

IHU ican

Pronostic des cardiomyopathies hypertrophiques
Données de près de **500** patients
Lancement en **2025**

ANNITIA

IHU ican

Prédiction de la progression de la stéatose hépatique non alcoolique
Données de près de **2 000** patients
Lancement en **2026**

ML-KIDCAR

IHU ican

Prédiction de l'insuffisance rénale aiguë post-chirurgie cardiaque
Données de près de **4 000** patients
Lancement en **2026**

MYOCARDIA

MIRACL.ai

Diagnostic de myocardite aiguë
Données de plus de **800** patients
Lancement en **2026**

AID-ORAL

Hôpital Pitié-Salpêtrière AP-HP

Détections de lésions de la muqueuse buccale
Plus de **2 000** photographies
Lancement en **2026**

RESPRED-UC

HOPITAL FOCH

Prédiction de la réponse des tumeurs urothéliales aux immunothérapies
Données de plus de **500** patients
Lancement en **2026**

OUTSAIDER

GHU PARIS PSYCHIATRIE & NEUROSCIENCES

Corréler l'art brut à un diagnostic psychiatrique
Lancement en **2026**

L'Appel à Projets Data Challenge

Data Challenge

APPEL A PROJETS DATA CHALLENGES EN SANTÉ

Le HDH en partenariat avec la Délégation Ministérielle au Numérique en Santé, le Secrétariat Général Pour l'Investissement et Bpifrance a lancé en juillet 2023 un nouvel appel à projets !

Prochaine relève des candidatures :

30 juin 2025

Des équipes sont **appelées à candidater** pour organiser un challenge sur une **thématique médicale de leur choix**

Le dépôt de candidature est ouvert sur la **plateforme en ligne de Bpifrance**

GOUVERNEMENT
Liberté
Égalité
Fraternité

FRANCE 2030

bpifrance

HEALTH DATA HUB

Critères de sélection des projets

Data Challenge

Les projets sélectionnés lors de cet appel à projets devront notamment répondre aux critères suivants…

Intérêt de la question clinique

La question médicale posée doit présenter un **intérêt clinique** et être **originale**.

Pertinence de l'approche prédictive pour y répondre

La problématique doit pouvoir être traitée par une **solution d'apprentissage automatique**, le score de performance des algorithmes doit être mesurable et les données d'une **qualité suffisante** pour que les résultats soient fiables.

Faisabilité de la collecte des données et de leur anonymisation

Les données doivent pouvoir être **collectées en nombre suffisant** dans un temps imparti à un ou des formats numériques homogènes. Les données doivent pouvoir être **totalement anonymisées dans le respect du RGPD**.

Partage des données et des résultats

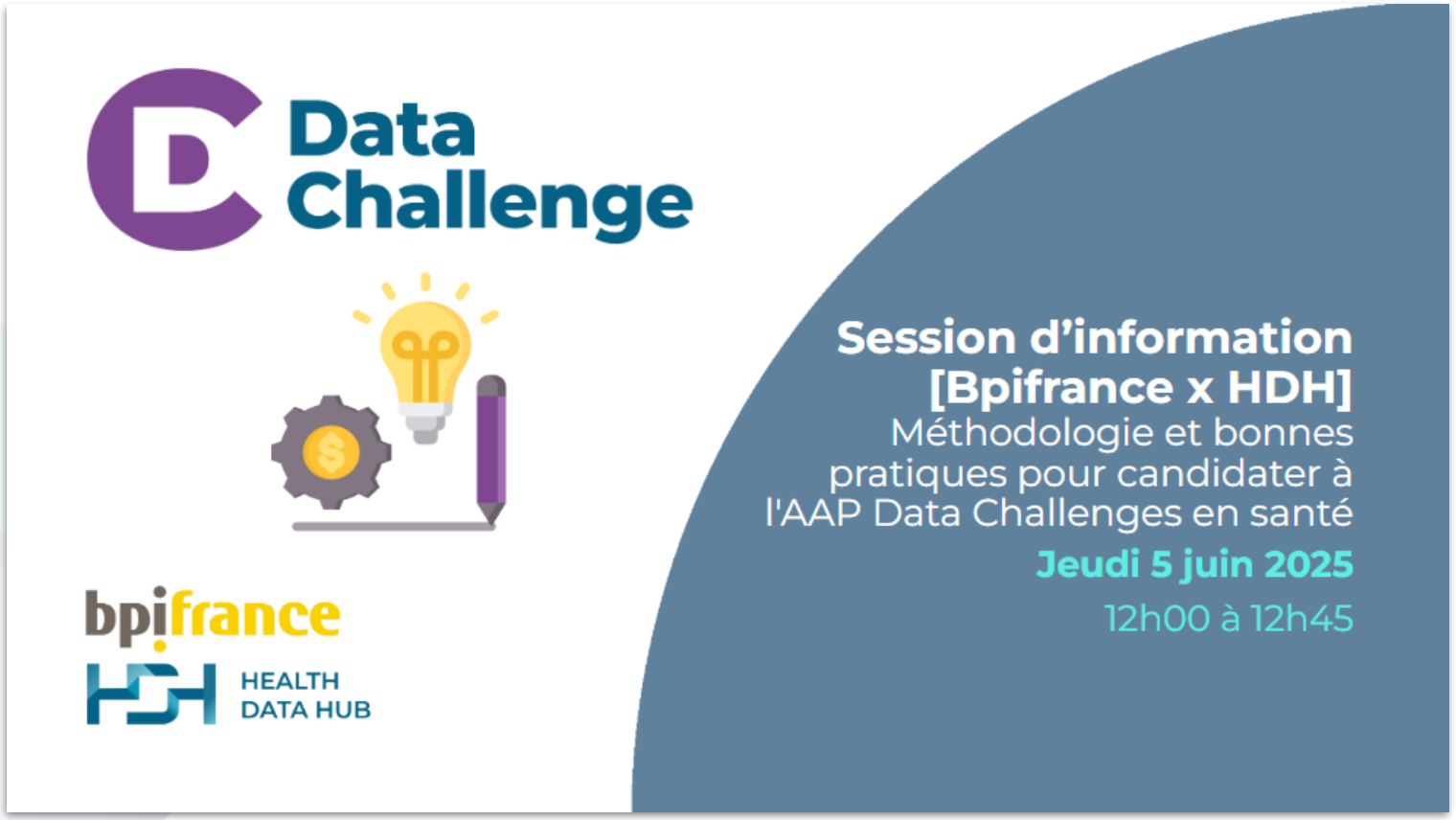
Les données collectées et anonymisées dans le cadre du projet seront **partagées en Open Data** à l'issue du Data Challenge. Des modalités de citation des bases sont toutefois prévues. Les gagnants devront **partager leur algorithme en Open Source** pour obtenir le prix de récompense.

Forte implication et faisabilité du projet

Une **forte implication de votre part** est requise sur toute la durée de l'organisation du data challenge. Les grands jalons proposés pour le projet doivent le rendre **réalisable en 18 mois**. La demande de financement ne doit pas dépasser **300k €**.

Save the Date ! Session d'information le 5 juin

Data Challenge



Bpifrance et le Health Data Hub organisent le **5 juin 2025 à 12h00** une **session d'information en ligne** dédiée aux **porteurs de projets souhaitant déposer une candidature en vue de la relève du 30 juin 2025**.

- Présentation des **modalités de candidature**
- Passage en revue des **pièces du dossier de candidature**
- Conseils pour le remplissage des **annexes financières et administratives**
- Partage de **bonnes pratiques**
- Session collective de **Q&A**

Inscrivez-vous ici !

Accompagnement du HDH et organisation type

Data Challenge

Sélection de projets via l'AAP Data Challenges en santé

Bpifrance dans le cadre du plan France 2030 apporte un soutien financier à hauteur de 300K€

Le HDH accompagne les projets sélectionnés depuis la conception scientifique jusqu'à l'ouverture des résultats en offrant un soutien logistique, scientifique, technique et humain.

- Définition et cadrage du projet
- Aide à la collecte, à l'anonymisation, à l'annotation des données
- Mise à disposition d'infrastructures de stockage et de traitement des données
- Communication et promotion des résultats

18 mois

1. Cadrage scientifique, réglementaire, opérationnel & financier
   - Constitution des comités
2. Collecte des données auprès des centres inclueurs
   - Convention avec les centres
   - Information des patients
3. Préparation des données : Annotation et labellisation
   - Base de données finale
4. Data Challenge (1 à 4 mois)
   - Hébergement sur la plateforme
   - Remise des prix
5. Communication, valorisation des résultats
   - Ouverture des données et algorithmes

RESTEZ INFORMÉS !
DATA CHALLENGES EN SANTÉ

Pour rester informé sur les ouvertures des Data Challenges, les actualités et les événements, abonnez-vous à la Newsletter Data Challenges en santé

Pour toute autre question ou pour organiser un data challenge en santé, vous pouvez nous écrire :

data.challenge@health-data-hub.fr

Le Data Challenge CytologIA - Améliorer le diagnostic en hématologie biologique grâce à l'IA

HEALTH DATA HUB

Le Data Challenge CytologIA - Améliorer le diagnostic en hématologie biologique grâce à l'IA

15h05 - 15h20

Dr Samy Dahmami

Biologiste médical, co-fondateur d'Algoscope

Le Dr Samy Dahmani, biologiste médical, entrepreneur et innovateur, se consacre à la modernisation de la médecine de laboratoire par l'intelligence artificielle et l'automatisation. C'est à l'issue de ses études de médecine qu'il a cofondé Algoscope, avec pour mission d'automatiser, de standardiser et de fiabiliser les processus de la médecine de laboratoire afin d'en optimiser l'efficacité et de renforcer la sécurité des patients.

Sa thèse de médecine sur l'application de l'IA à la lecture du frottis sanguin, a ensuite inspiré l'organisation du Data Challenge Cytologia. Ce projet vise à améliorer le diagnostic en hématologie biologique (notamment sa précocité) et à encourager la valorisation des données de santé, via des solutions d'IA open source.

Le Datachallenge CytologIA : Amélioration du diagnostic en hématologie biologique

Mercredi 4 juin 2025

Samy Dahmani

Journée Open Science en Santé

HEALTH DATA HUB

GROUPE FRANCOPHONE D'HÉMATOLOGIE CELLULAIRE

ALGOSCOPE

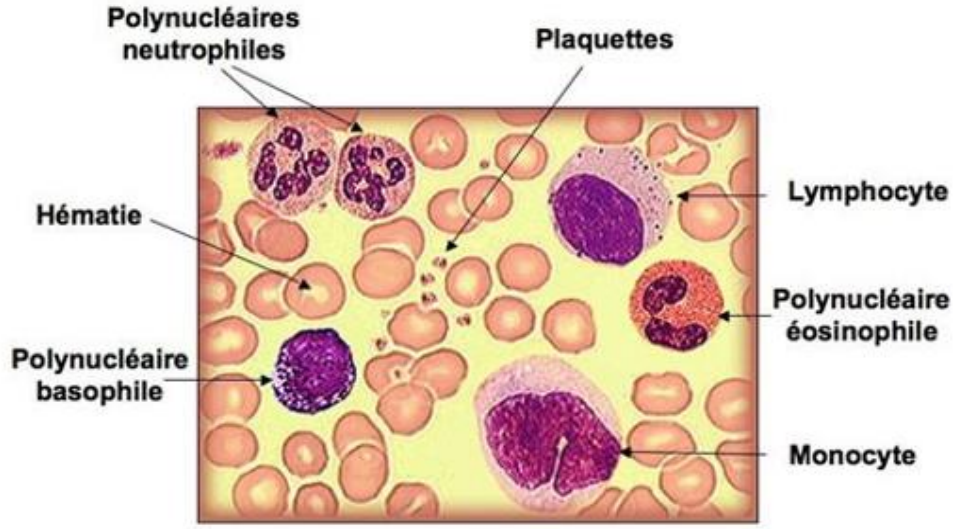
Trustii.io

Introduction

- L'**Hématologie** est l'étude des maladies du sang
- La **cytologie** est l'étude des éléments figurés du sang
- La **numération formule sanguine** ou **hémogramme** est la première analyse prescrite en France

Hémogramme

- Etude **quantitative** et **qualitative** des éléments figurés du sang
- Il y a 3 types de cellules :
  - Les leucocytes
  - Les hématies
  - Les plaquettes
- Très nombreuses indications
- De nombreuses classes et sous classes de leucocytes, discrimination délicate.
- Expertise du cytologiste



Catégories de cellules sanguines		Nombre moyen chez l'homme	Nombre moyen chez la femme
Hématies (millions/mm³de sang)		4,2 à 5,7	4,0 à 5,3
Plaquettes (milliers/mm³de sang)		150 à 400	150 à 400
Leucocytes	Total par mm³de sang	4 000 à 10 000	4 000 à 10 000
Détail des leucocytes	Granulocytes neutrophiles	45 à 70 % soit 1 700 à 7 500/microlitre	
	Granulocytes éosinophiles	1 à 3 % soit 40 à 300/microlitre	
	Granulocytes basophiles	0,5 % soit moins de 50/microlitre	
	Lymphocytes	20 à 40 % soit 1 000 à 4 000/microlitre	
	Monocytes	2 à 8 % soit 200 à 1 000/microlitre	

Réalisation de l'hémogramme

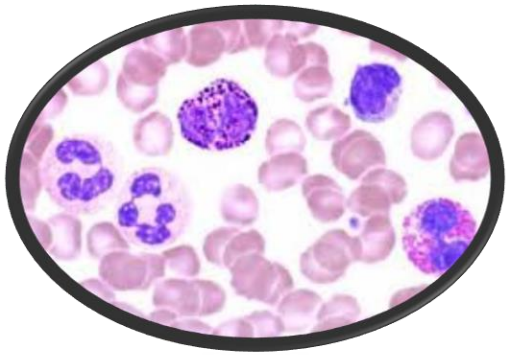
- Etape Automatisée :

Numération automatisée des éléments figurés du sang

Selon critères

- Etape Manuelle :

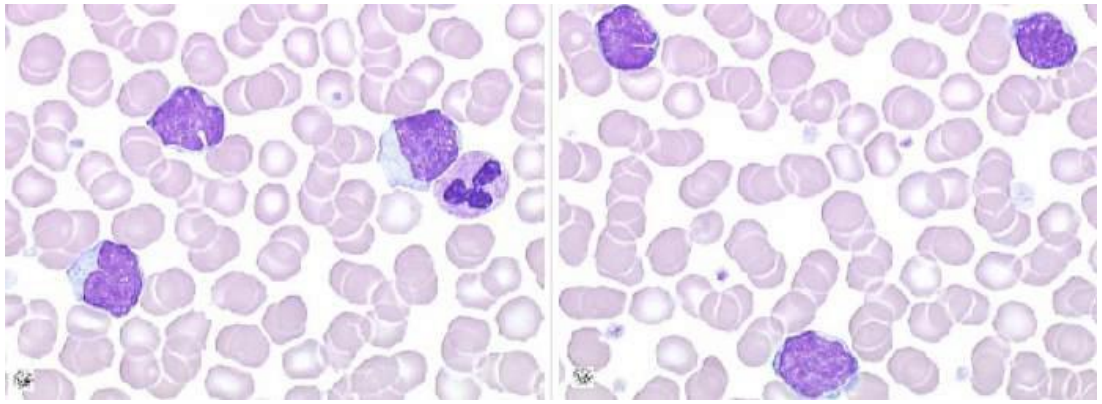
Lecture du **frottis sanguin**



Lecture frottis sanguin

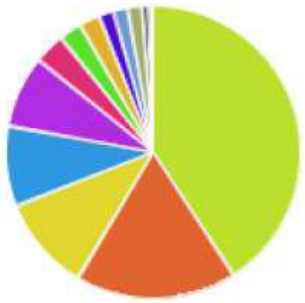
- Analyse morphologique frottis sanguin : **élément clé du diagnostic et du suivi des hémopathies malignes et non malignes**
- Gold standard : **Microscope optique (MO)**
- **Hétérogénéité des compétences sur le territoire**

La cytologie est une discipline difficile



Association de Biologie Praticienne, Campagne 2017 (CYT-17-3-2), Pr M.Zandecki

Nombre de réponses analysées: 852



- Leucémie lymphoïde chronique avec atypies morphologiques (40.7%)
- Dissémination sanguine d'un lymphome à cellules du manteau (17.6%)
- Dissémination sanguine d'un lymphome folliculaire (10.4%)
- Dissémination sanguine d'un lymphome (à petites cellules matures) (8.8%)
- Leucémie lymphoïde chronique (LLC) aspect typique (7.9%)
- Réponse en texte libre (3.4%)
- Hémopathie lymphoïde chronique non classable (2.3%)
- Dissémination sanguine d'un lymphome de la zone marginale à lymphocytes non villeux (2.1%)
- Suspicion de lymphocytose B monoclonale (1.5%)
- LLC forme prolymphocytoïde (1.5%)
- Je ne sais pas (préciser à quel niveau se situe votre problème) (1.5%)
- Lymphocytose réactionnelle (0.5%)
- Dissémination sanguine d'un lymphome à grandes cellules (0.5%)
- Leucémie prolymphocytaire (0.4%)
- Dissémination sanguine d'un lymphome splénique à lymphocytes villeux (0.2%)
- Aspect évoquant une leucémie aiguë monoblastique (LAM5 - FAB) (0.1%)
- Aspect évoquant une leucémie aiguë lymphoblastique (0.1%)
- Suspicion de macroglobulinémie de Waldenström (0.1%)
- Syndrome de Sézary (0.1%)
- Leucémie à tricholeucocytes (0.1%)

Automatisation de la lecture du frottis sanguin

- "Scanner" de cellules (environ 150 par frottis sanguin)

- Algorithme de classification excellent pour les leucocytes normaux **mais rapidement mis en défaut sur les lames pathologiques**

- **Développement d'une IA pour la classification des leucocytes**

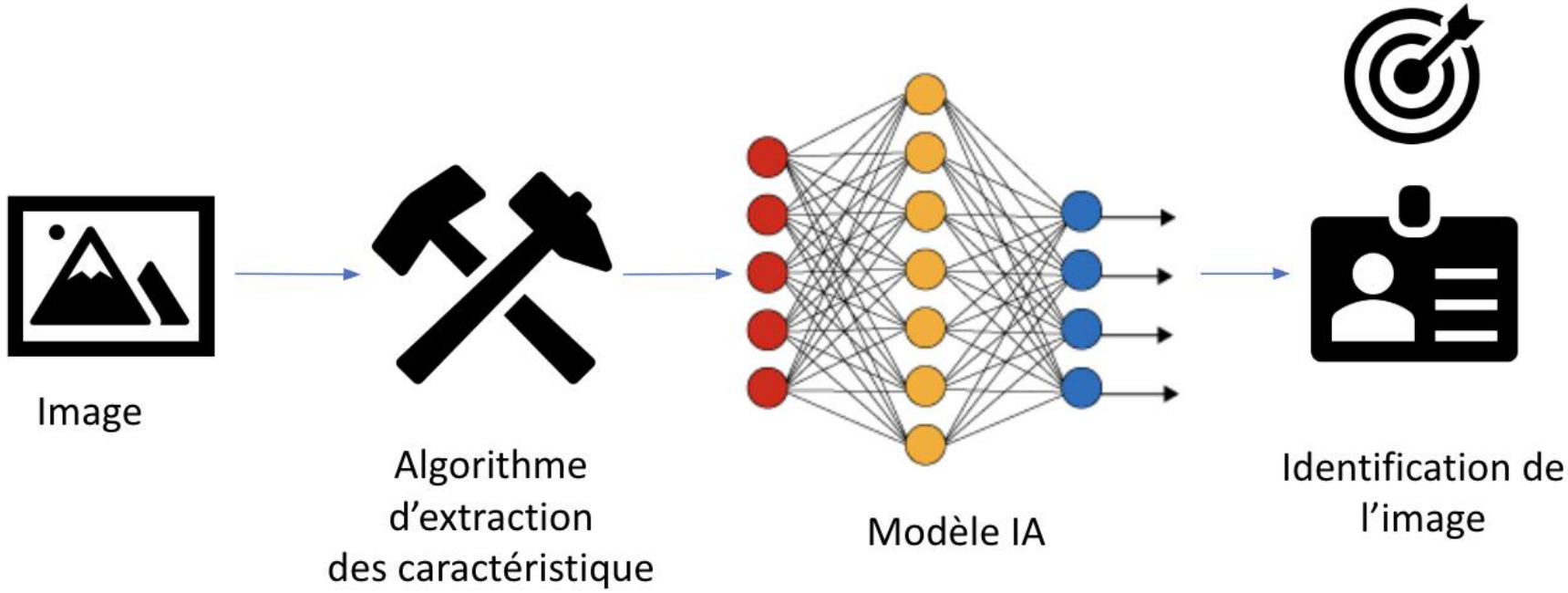
CellaVision

Sysmex

IA en cytologie

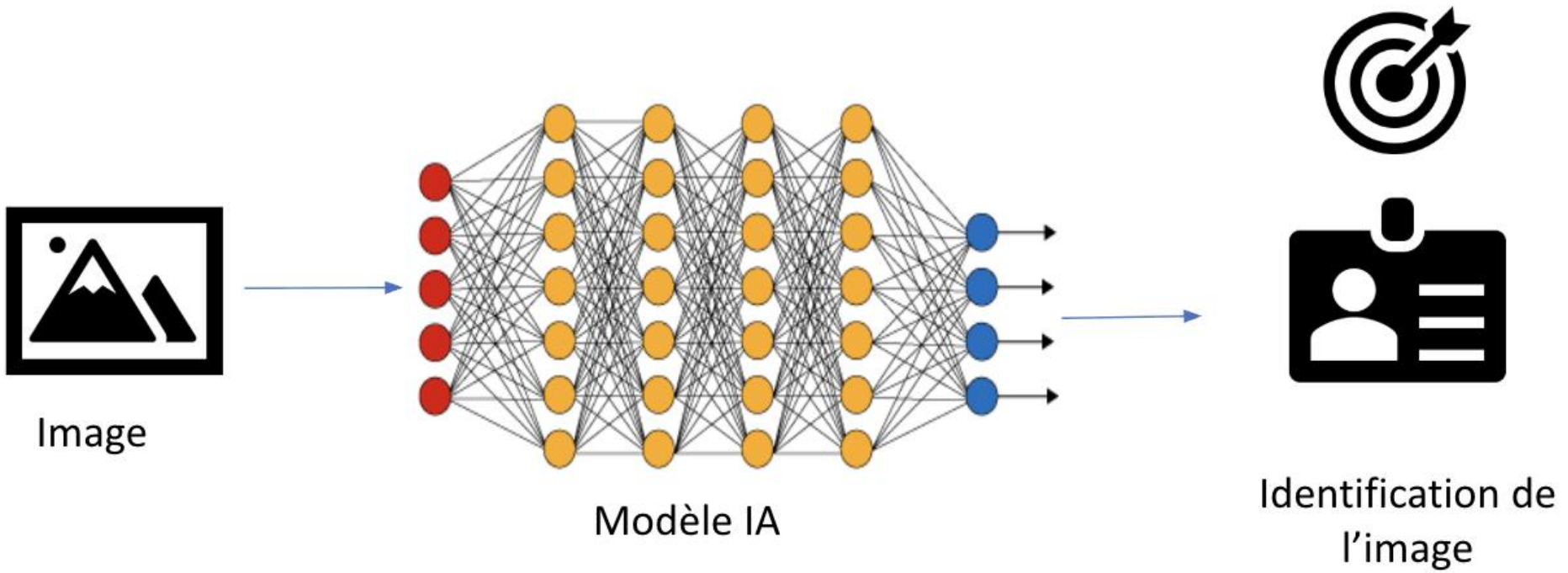
- **Vraie disparité au niveau du territoire en termes de compétence cytologique**
- Donner un outil de screening de 1ere ligne au centres périphériques/ labos de ville
- **Égalité des chances sur le territoire** (compétences spécialisées au CHU)
- Participer à la formation continue

Intelligence Artificielle



Machine Learning

Intelligence Artificielle



Deep Learning

IA en cytologie

Les jeux de données :

Nombreux datasets présents mais

- Peu de classes
- Monocentrique
- Faible nombre d'images

Dataset	No of Images
ALL-IDB [57]	130
Private Dataset [60]	70
CellaVision [47]	100
AA-IDB2 [8]	108
Hayatabad Medical Dataset [63]	1030
Isfahan Al-Zahra and Omid hospital [64]	312
Private Dataset [66]	431
ALL-IDB2/ Leishman-stained peripheral blood smears [59]	160/160
CellaVision [31]	450
Public Dataset [73]	92,800
BCCD [41]	12,444
Kaggle [42]	12,444
BCCD [33]	12,500
BCCD [76]	375
Kaggle/LISC [77]	12,500/400
LISC and BCCD [78]	6250
Jiangxi Tecom Science Corporation/ CellaVision/ Bsisc/ LISC [79]	300/ 100/ 268/ 257
KMC hospital, Manipal, India [80]	280
ALL-IDB [85]	108
Hybrid-Leukocyte database/ e Hybrid-Slide database [86]	891/ 377
Acquired from Sixth People's Hospital of Shenzhen [88]	21
Kaggle [89]	12,494
SMC-IDB/ IUMS-IDB/ ALL-IDB [92]	367/ 195/ 108
BCCD [94]	12447
SBILab [97]	76
BCCD [101]	2487
Kaggle [103]	12,444

https://doi.org/10.1371/journal.pone.0292026.t003

IA en cytologie

Les publications :

De nombreuses publications mais

- Pas de modèle disponible
- Pas d'utilisation en pratique

Author	Year	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)
Macawile et al. [39]	2018	AlexNet	96.63	98.85	99.61
Liang et al. [73]	2018	CNN + RNN	91	-	-
Sharma et al. [41]	2019	CNN	97	94	98
Togacar et al. [42]	2019	CNN	97.78	-	-
Mohamed et al. [74]	2020	Pre-trained Deep Learning Models	97.03	71	91
Ergen et al. [33]	2020	CNN, Feature Selection	97.95	98	97.75
Zhao et al. [75]	2021	TWO-DCNN	96	-	-
Cinar et al. [76]	2021	Alexnet- GoogleNet-SVM	99.73, 98.23	98.75	-
Wang et al. [77]	2019	CNN Architecture SSD and YOLOv3	90.09	-	-
Kutlu et al. [14]	2020	R-CNN	97.52	88.9	-
Fan et al. [78]	2019	ResNet50	98	-	-
Hegde et al. [79]	2019	Pre-trained AlexNet model	98.9	98.6	98.7
Acevedo et al. [80]	2019	Pre trained CNN	96.2	-	-
Qin et al. [81]	2018	Deep Residual Learning	76.84	-	-
Tiwari et al. [82]	2018	Double CNN model	97	83	-
Hung et al. [83]	2017	AlexNet and Fast R CNN Model	72	-	-
Naz et al. [84]	2017	CNN, faster R CNN	94.71	95.42	99.27
Vogado et al. [85]	2018	CNN with SVM	99.20	99.2	-
Habibzadeh et al. [86]	2018	ResNet and Inception	99.46	99.89	-
Song et al. [87]	2014	CNN	94.5	87.26	-
Fatih et al. [88]	2019	MRMR feature selection -ELM and CNN	97.37	-	-
Rehman et al. [89]	2018	Deep CNN	97.78	-	-
Bani-Hani et al. [90]	2018	CNN with the optimized genetic method	91	91	97
Di Ruberto et al. [91]	2020	Pre trained AlexNet	97.93	99.6	-
Loey et al. [92]	2020	Pre trained CNN AlexNet	100	100	98.2
Ma et al. [93]	2020	Generative Adversarial Network and residual neural network	91.7	92	-
Baydilli et al. [94]	2020	Capsule Networks	96.86	92.5	98.6
Tobias et al. [95]	2020	Faster Residual Neural Network	83.25	-	-
Kassani et al. [96]	2019	Hybrid DL based model	96.17	95.17	98.58
Baghel et al. [97]	2022	CNN	98.51	98.4	-
Shahzad et al. [98]	2022	CNN	98.44	99.96	99.98
C. Cheuque et al. [99]	2022	Multilevel CNN	98.4	98.3	-
Hosseini et al. [100]	2022	Convolutional Neural Network	97	94	98
Ramya et al. [101]	2022	CNN-PSO	99.2	94.56	98.78
Khalil et al. [102]	2022	CNN	98	-	-
Sharma et al. [103]	2022	DenseNet121	98.84	98.85	99.61

https://doi.org/10.1371/journal.pone.0292026.t008

IA en cytologie

- L'IA en cytologie hématologique est un sujet de recherche en constante évolution.
- Les études publiées en la matière ne partagent généralement pas de modèle d'IA et le nombre de classes reconnues est restreint ce qui limite leur impact et leur applicabilité en pratique.
- Il y a de nombreuses équipes qui y travaillent mais pour avoir un impact, il leur faut un dataset, varié, riche et représentatif de la routine
- Le projet CytologIA

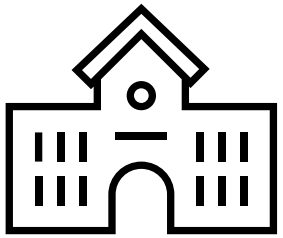
Projet CytologIA

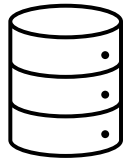
- **Réaliser une intelligence artificielle permettant de classer les leucocytes normaux et pathologiques utilisable en routine**
- 1ère étape : Constituer une base de données d'image de leucocytes normaux et pathologiques de bonne qualité, riche et variée
- 2ème étape : Datachallenge afin de sélectionner la meilleure IA

Intérêts du réseau du GFHC

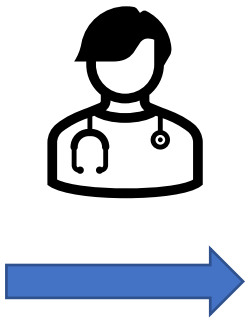
- GFHC = Groupe Francophone d'Hématologie Cellulaire
- Environ 300 biologistes spécialisés en hématologie
- **Grande quantité d'images fournies**, dans les différentes classes de leucocytes (normaux/pathologiques)
- **Pas d'effet centre** (overfitting)
- **Recueil rétrospectif** : anonymisation des images à 100%

Constitution de la base de données

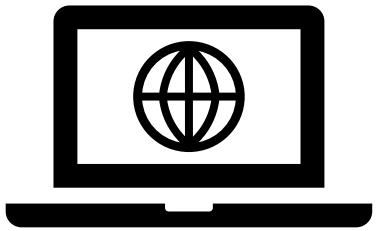




Base de données rétrospective anonymisée des centres inclueurs (GFHC)



Choix des images à uploader par le cytologiste



Plateforme

ALGOSCOPE

Revue et Labélisation des images par Des cytologistes

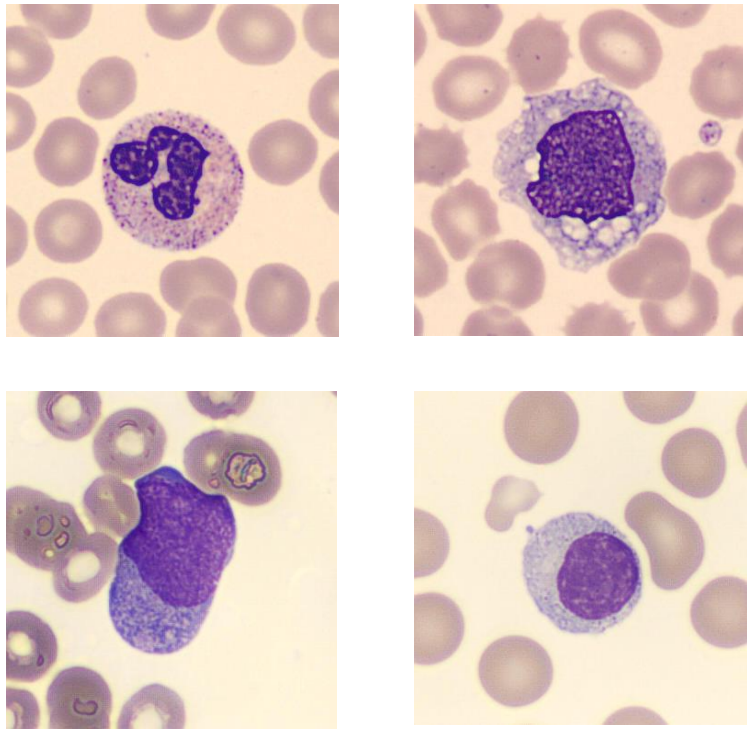
Base de données d'entraînement

Base de données test

Base de données de validation

Dataset CytologIA

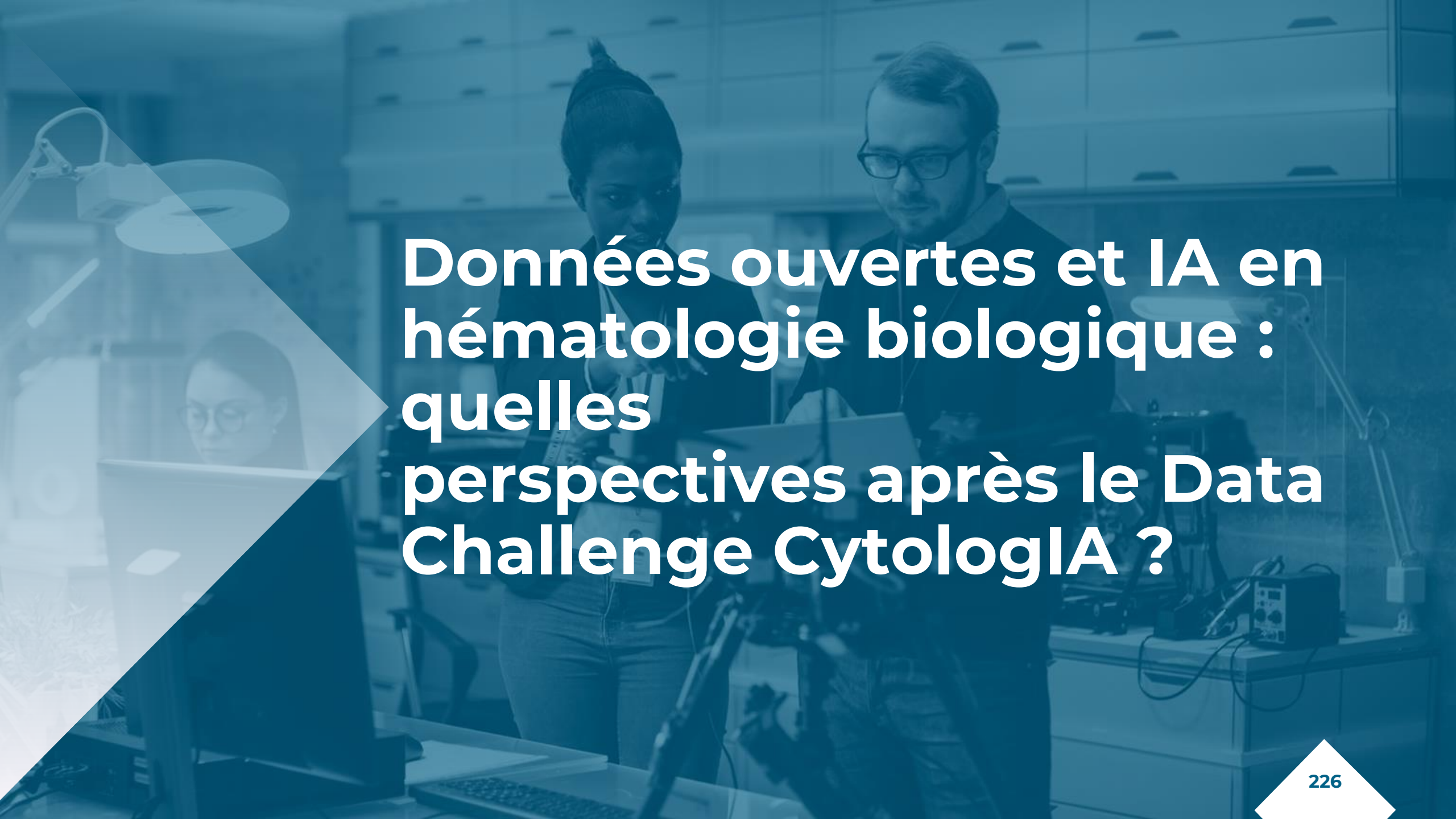
- **23 classes** de leucocytes, couvrant la routine
- **Multicentrique** : Données provenant de **23 centres experts**
- **69168 images labélisées**
- **Résolutions d'image variées**



Le Datachallenge

- 6 semaines
- **245 équipes** du monde entier
- Plus de **2000** modèles soumis
- Des scores finaux dépassant **93% de précision**
- **3 lauréats**

Merci



Données ouvertes et IA en hématologie biologique : quelles perspectives après le Data Challenge CytologIA ?

HEALTH DATA HUB

Données ouvertes et IA en hématologie biologique : quelles perspectives après le Data Challenge CytologIA ?

15h20 - 15h40

Dr Thomas Boyer

Hématologue et biologiste au CHU d'Amiens, Secrétaire adjoint du GFHC

Le Dr Thomas Boyer est MCU-PH à l'Université Picardie Jules Verne et dans le service d'Hématologie Biologique du CHU d'Amiens. En tant que secrétaire adjoint du Groupe Francophone d'Hématologie Cellulaire (GFHC), il a co-organisé le Data Challenge CytologIA en 2024 consacré au diagnostic des pathologies hématologiques sur frottis sanguin.

Données ouvertes et IA en hématologie biologique : quelles perspectives après le Data Challenge CytologIA?

Thomas Boyer

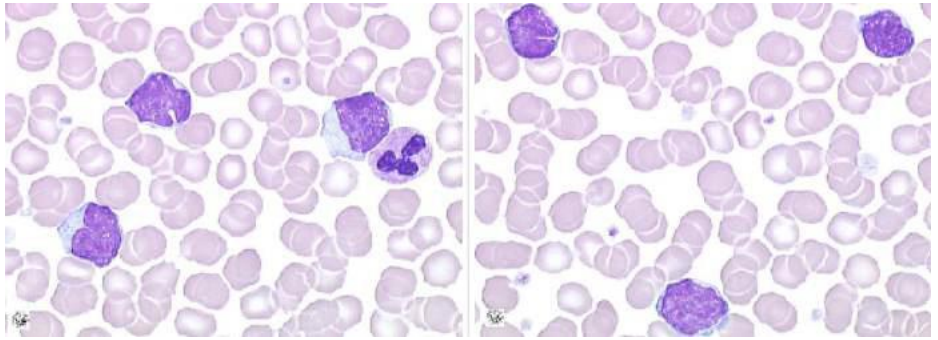
Journée Open Science en Santé

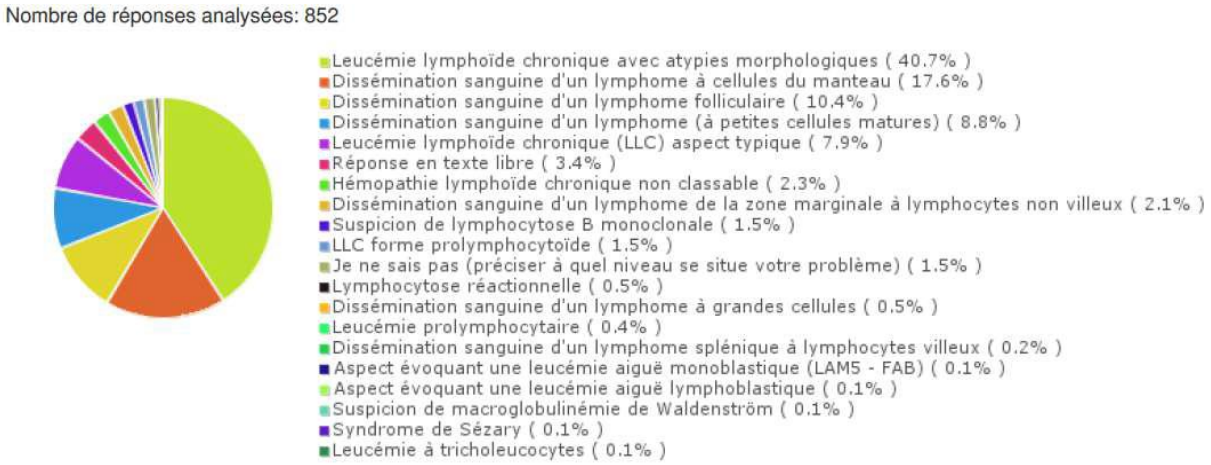
Mercredi 4 juin 2025

Pourquoi CytologIA?

- **La cytologie c'est DIFFICILE**

Association de Biologie Praticienne, Campagne 2017 (CYT-17-3-2), Pr M.Zandecki



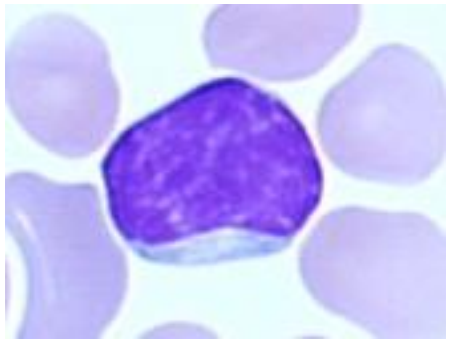


Expérience du biologiste +++

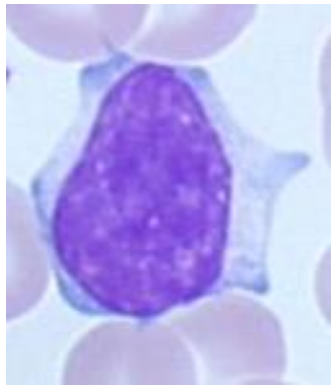
Disparité des compétences sur le territoire

Une place pour l'IA?

La base de données CytologIA comme outil pédagogique



Leucémie Lymphoïde Chronique
(= cancer)



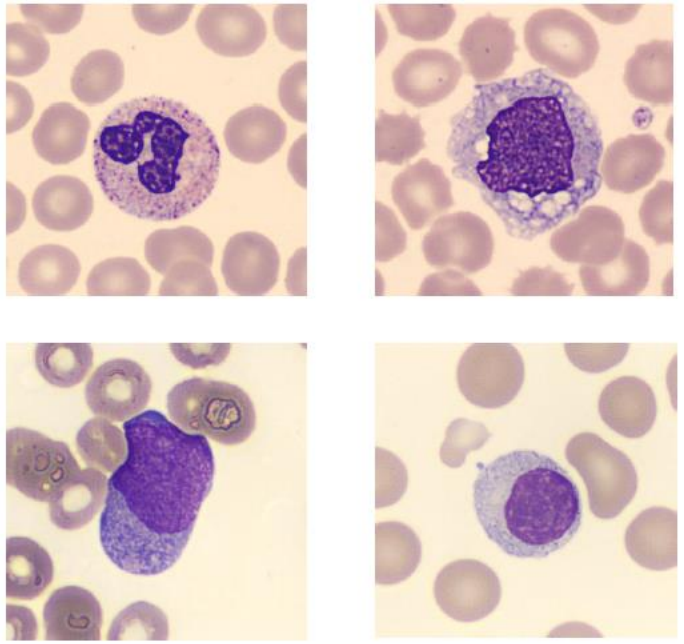
Lymphocyte normal

Data Challenge CytologIA

Réaliser une intelligence artificielle permettant de classer les leucocytes normaux et pathologiques

1ère étape : constituer une base de données d'image de leucocytes normaux et pathologiques de bonne qualité et variée

2ème étape : Data Challenge avec sélection de l'algorithme le plus performant



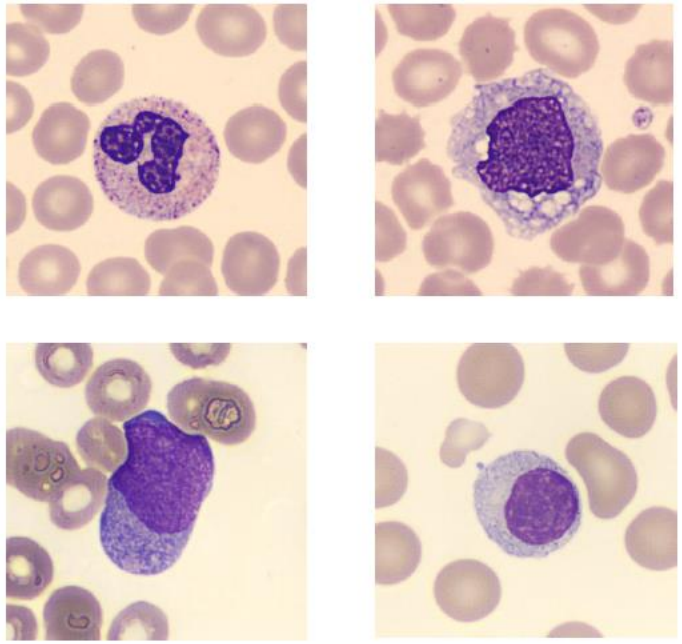
Datachallenge CytologIA

Point fort +++

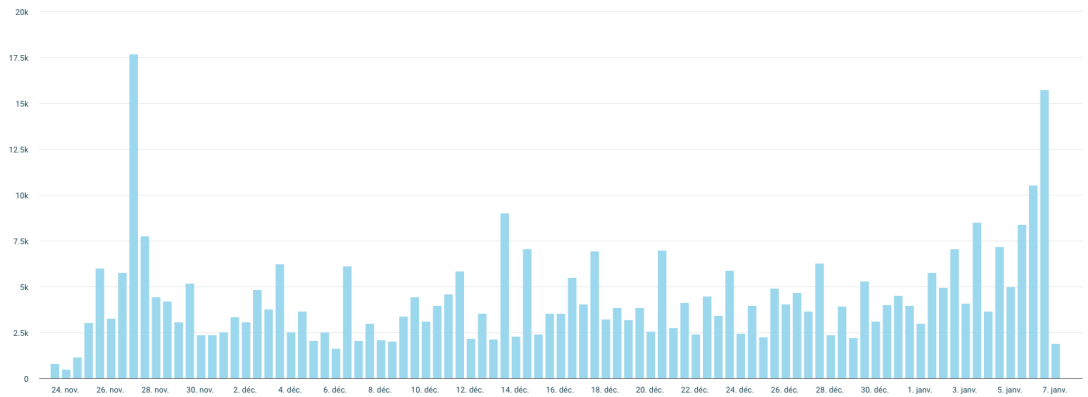
Réaliser une intelligence artificielle permettant de classer les leucocytes normaux et pathologiques

1ère étape : constituer une base de données d'image de leucocytes normaux et pathologiques de bonne qualité et variée

2ème étape : Data Challenge avec sélection de l'algorithme le plus performant



Datachallenge CytologIA



- 6 semaines : du 29/11/2024 au 06/01/2025
- **245** participants
- Base de données en Open Data
- Algorithmes d'IA en Open Source

Quelles suites? L'histoire ne s'arrête pas là!

- Suite immédiate : publication scientifique

Classification of normal and pathological blood leukocytes is possible thanks to artificial intelligence and deep learning : results of the CytologIA DataChallenge

Elise Sourdeau[1], Franck Geneviève[2], Lucile Baseggio[3], Bouchra Badaoui[4], Charles Chevalier[5], Mélanie Pannetier[6], Alexandre Janel[7], Véronique Verge[8], Laurent Weinmann[9], Camille Debord[10], Patrick Cohen[11], Agathe Maillon[12], Sandrine Girard[13], Frédérique Dubois[14], Véronique Baccini[15], Pierre Lemaire[16], Yaël Berda-Haddad[17], Thomas Tassin[18], Anne-Camille Faure[19], Soufiane Azdad[20], Samy Dahmani[20], Lauriane Armand[21], Agathe Delaune[21], Naama Bak[22], Paul Steffen[22], Valérie Bardet[1], Thomas Boyer[23]

Nature Communications (envisagé)

Quels outils sont actuellement disponibles?

Raabin-WBC dataset : 40000 images

72 lames de sang, 2 photomicroscopes

Figure 2

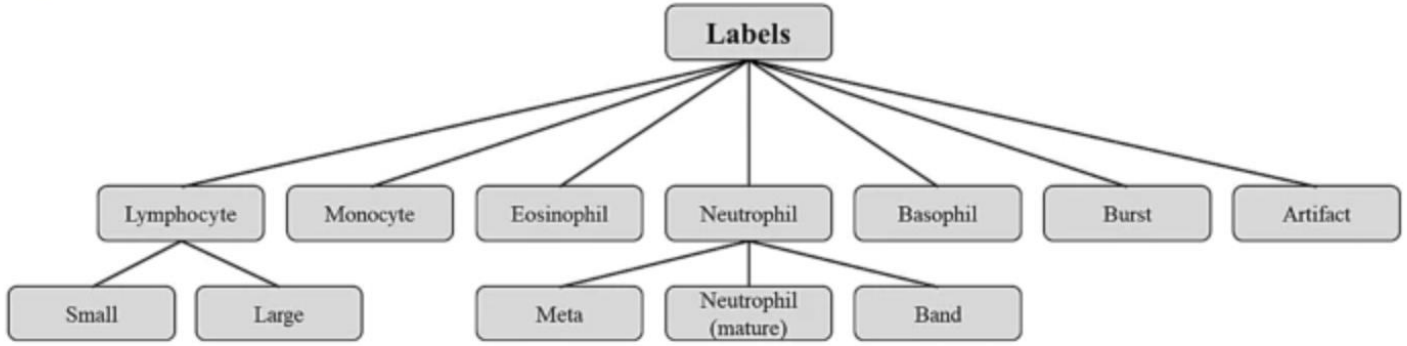
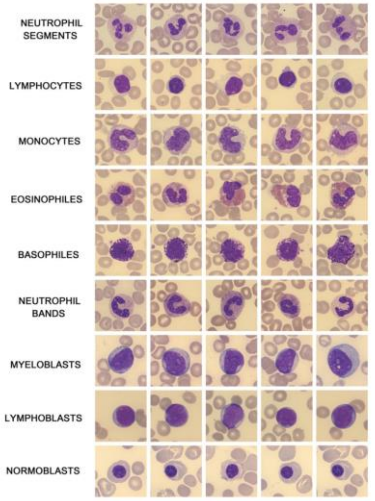


Diagram of labels in the Raabin-WBC dataset.

High Resolution WBC Dataset: 16027 images

Leucocytes normaux et pathologiques (9 classes)



Bodzas, Scientific data 2023

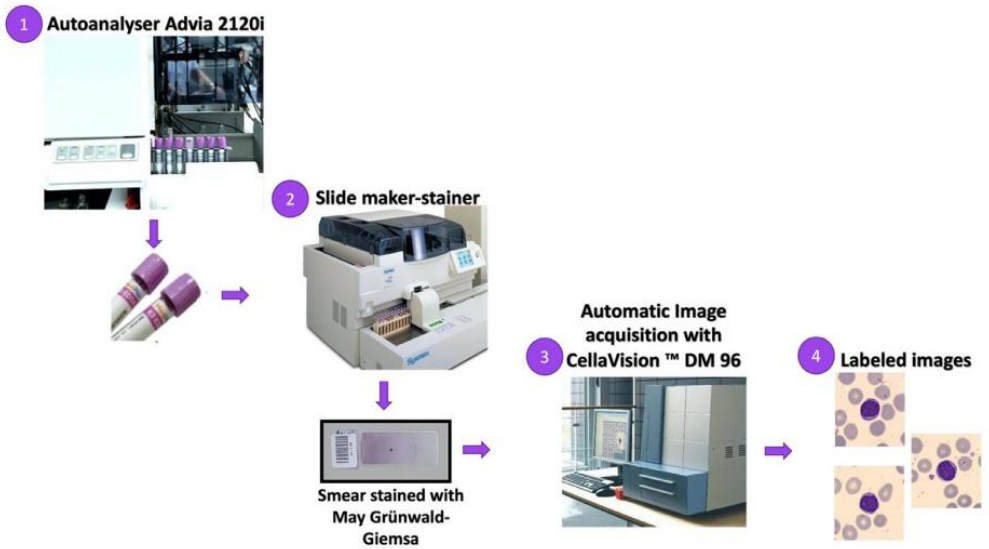
Quels outils sont actuellement disponibles?

Cancer Imaging Archive : 189784 images, 16 classes, microscope automatisé (myélome, LAM, lymphomes…)

Dataset de Barcelone : 17092 images, DM96, 8 classes (5 leucocytes normaux, granuleux immatures, érythroblastes, plaquettes géantes)

Globalement, pas assez de classes et pas assez de variabilité (datasets monocentriques)

Dataset multicentres +++



Acevedo, Data Brief 2020

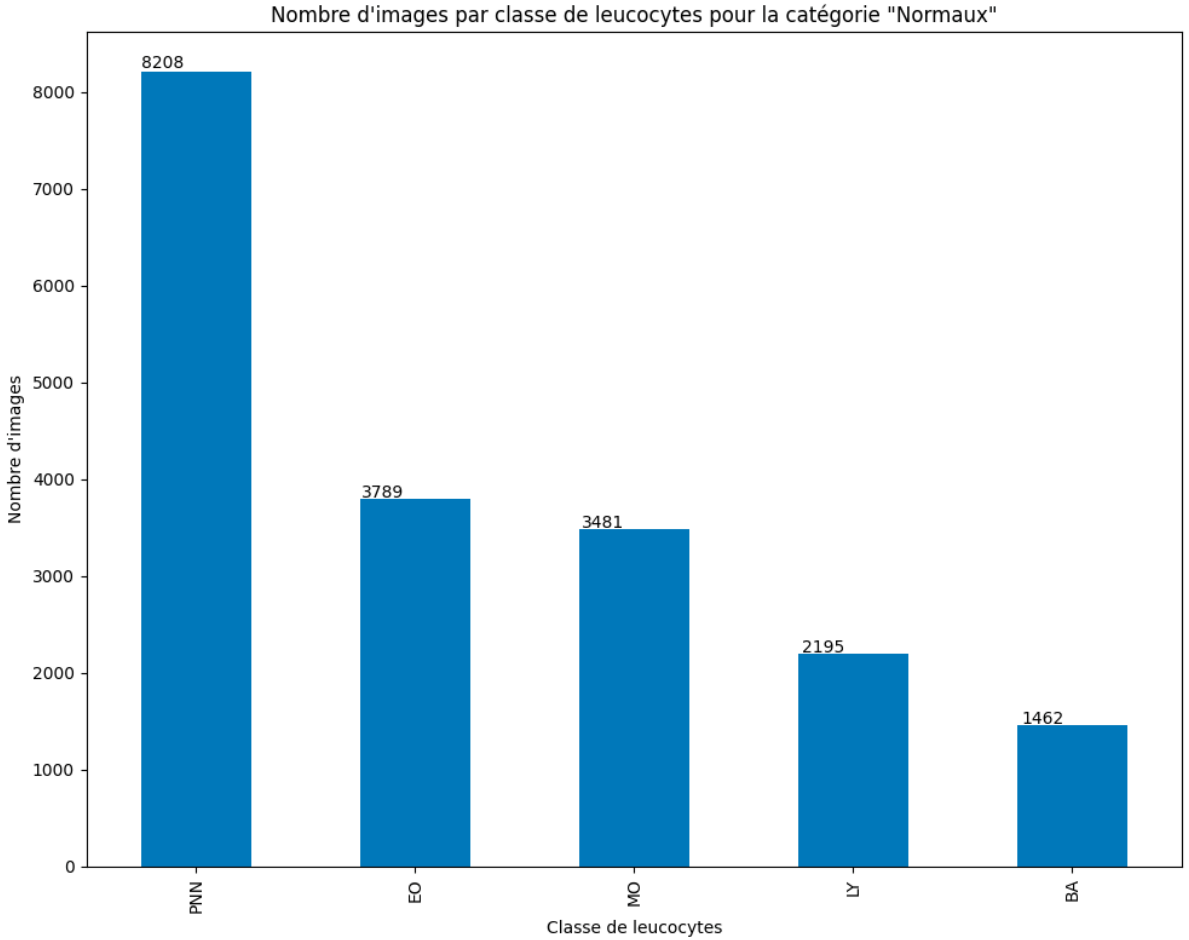
La base de données CytologIA

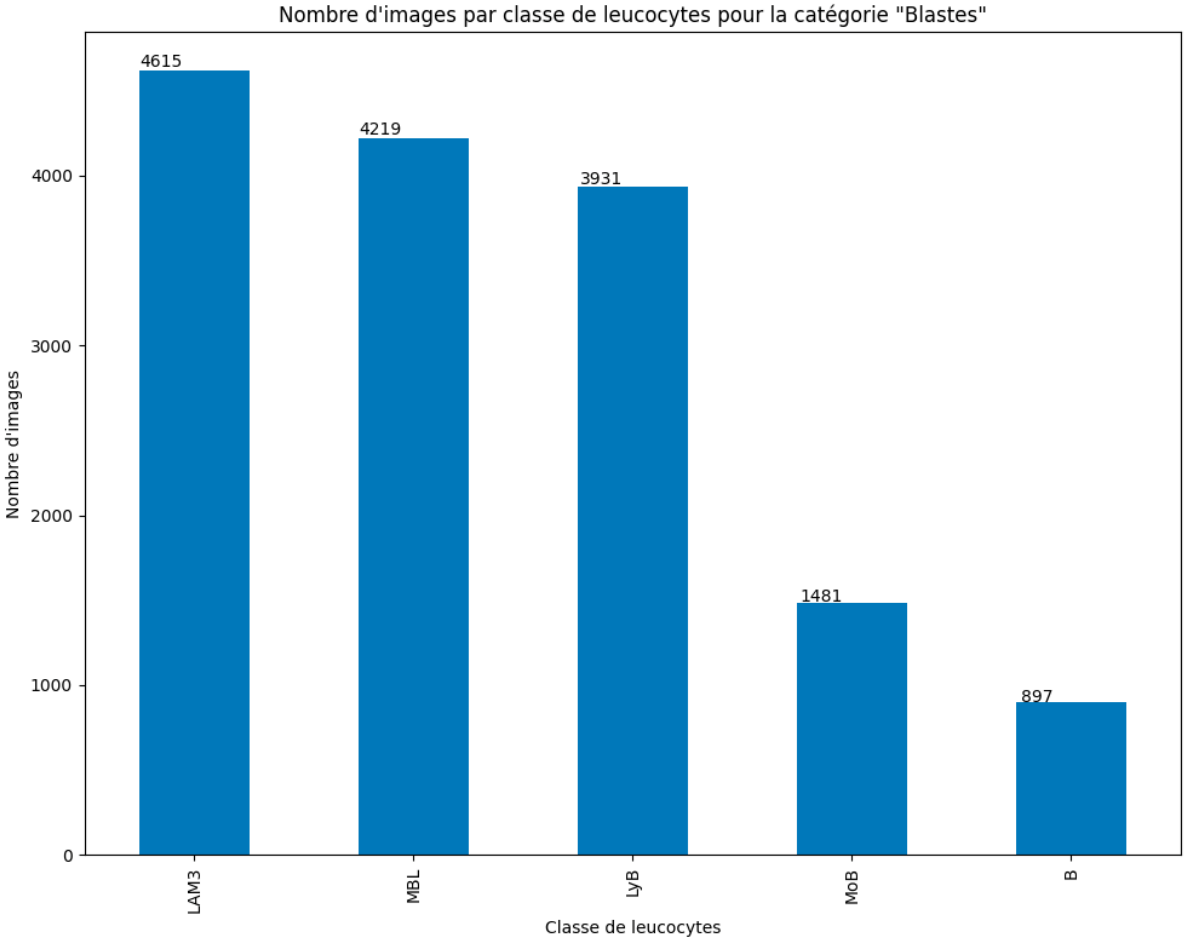
69168 images réparties en 23 classes

Labélisation des images par un biologiste de chaque centre, **relecture par 4 biologistes experts du GFHC**

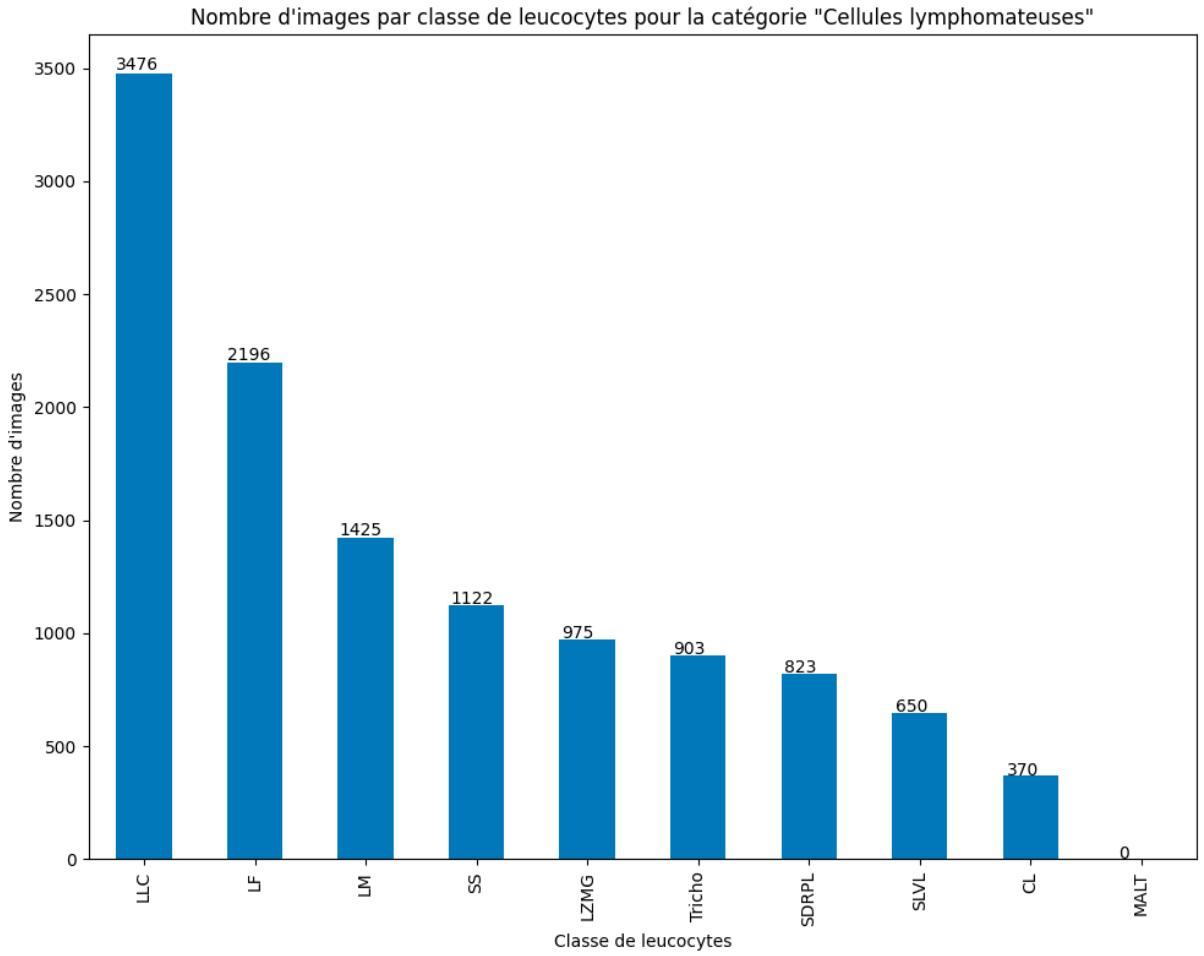
20 centres Francophones ont participé à l'élaboration de la base de données

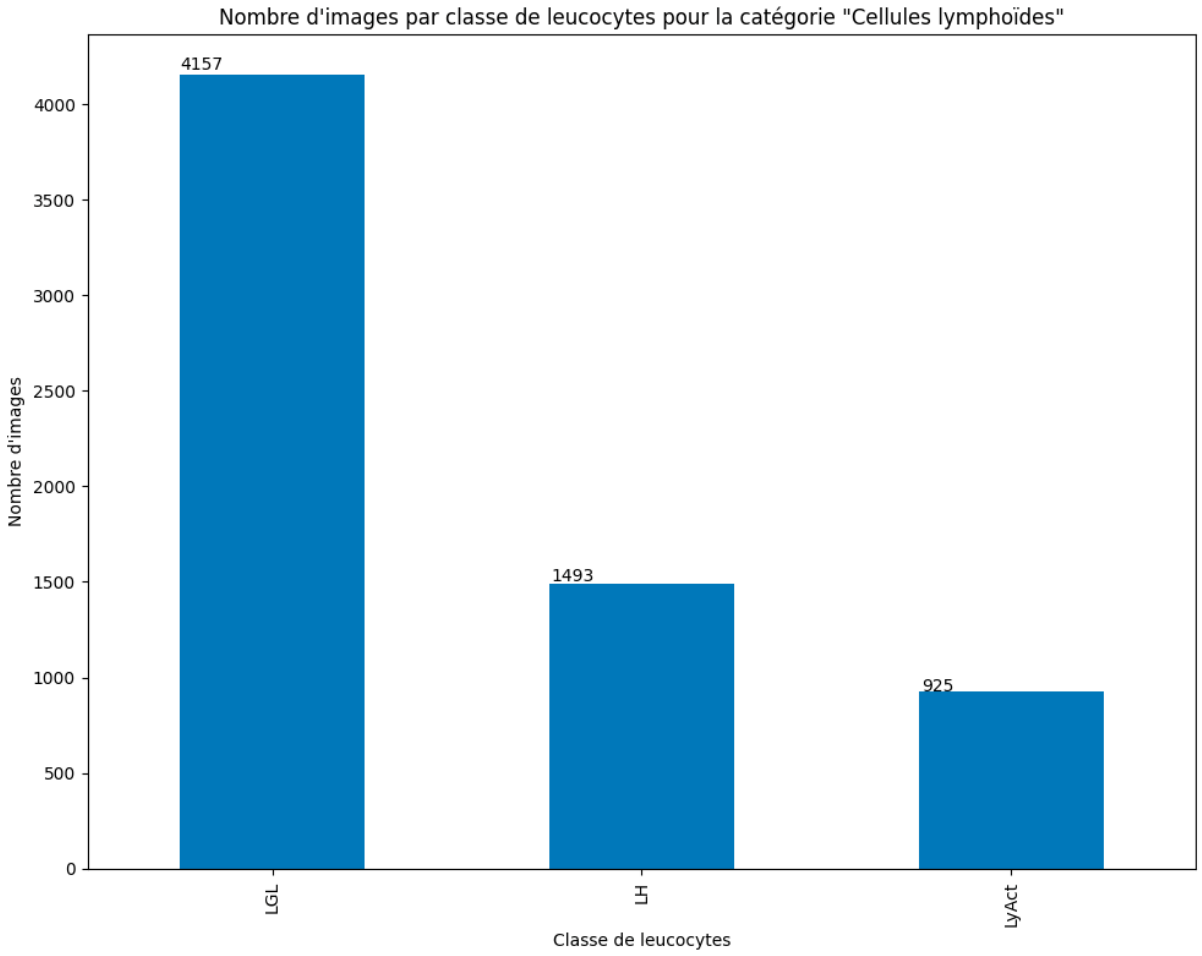
La base de données CytologIA



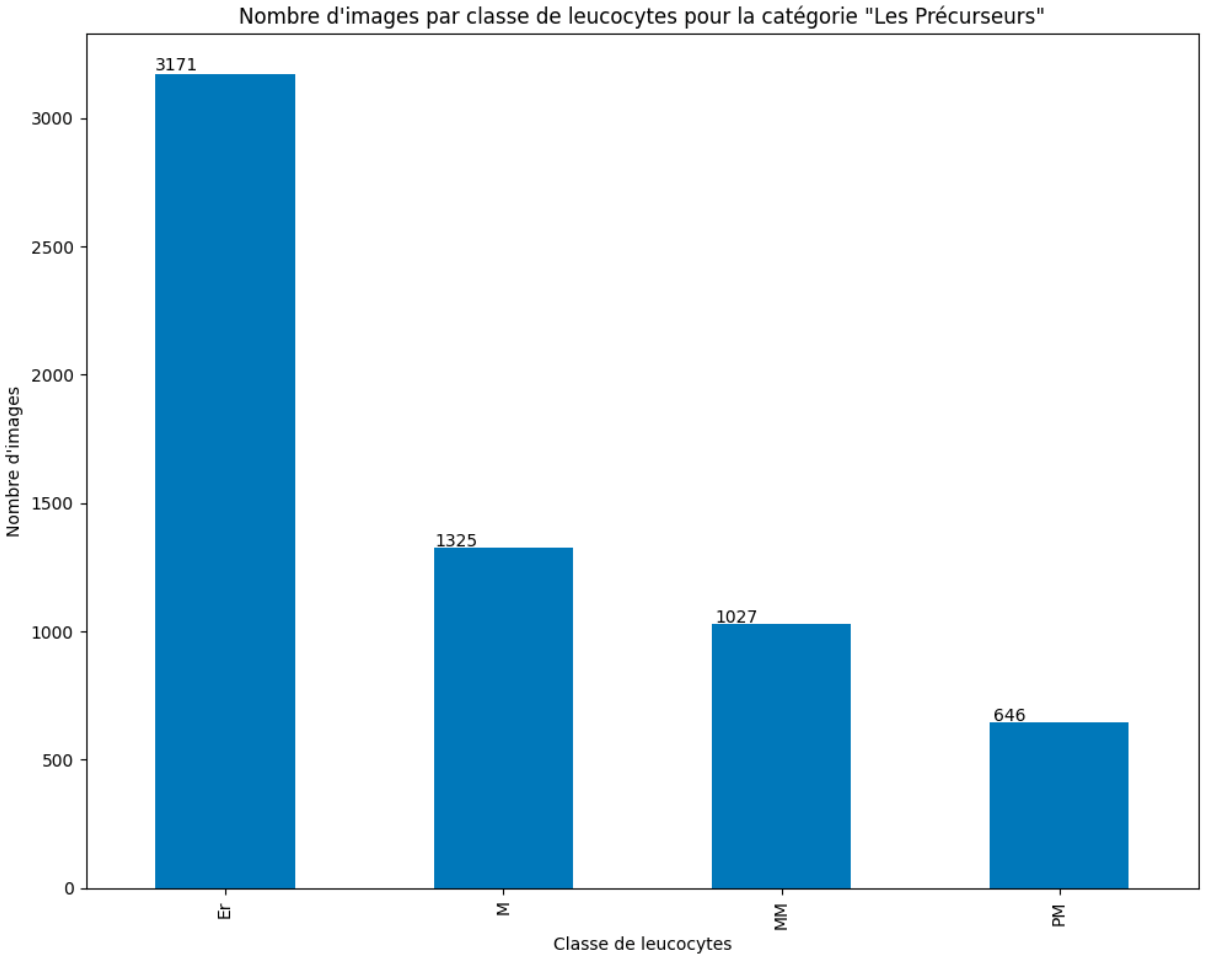


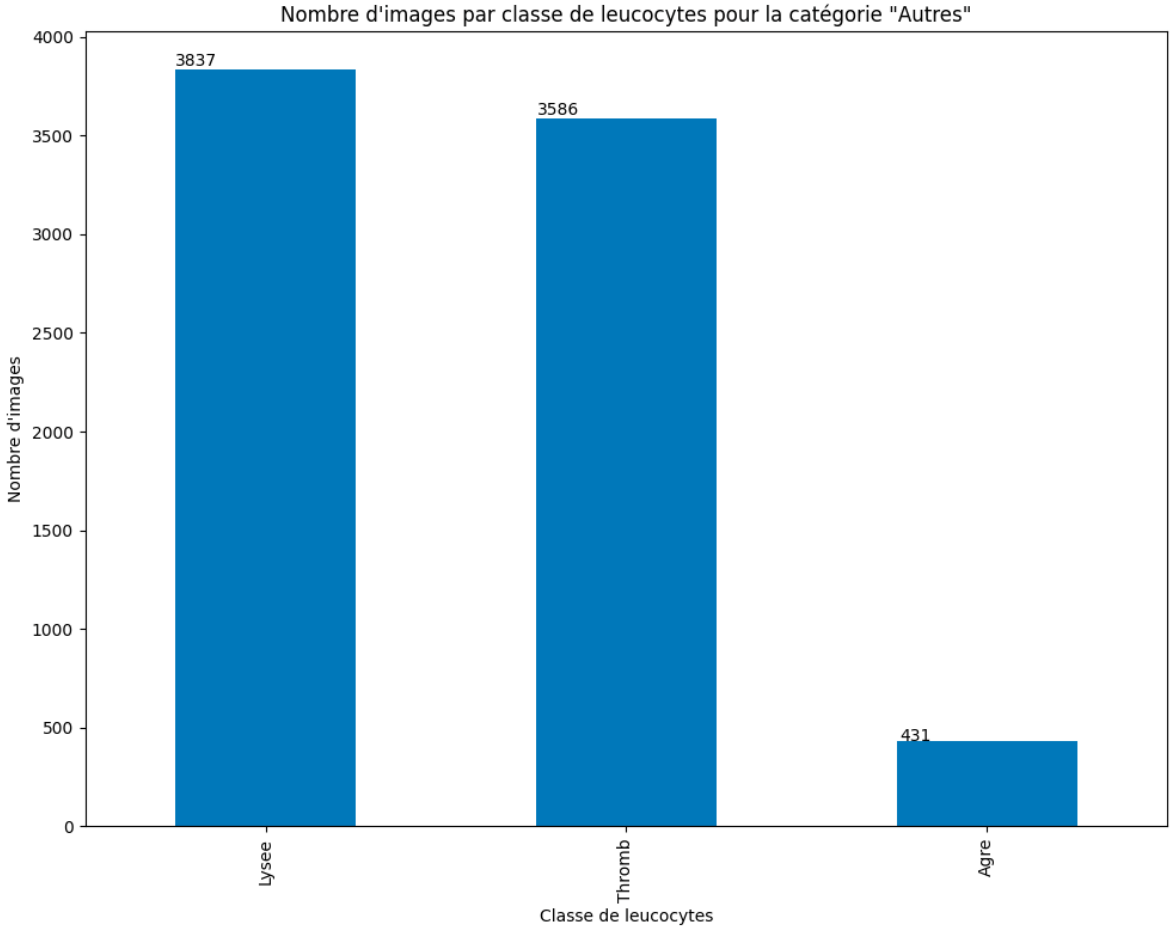
La base de données CytologIA





La base de données CytologIA





Et au-delà de l'outil pédagogique?

- **But du Data Challenge : égalité des chances d'un diagnostic hématologique précis sur le territoire**
- **Comment? Intégration de l'algorithme aux microscopes automatisés**
- **Performances peu satisfaisantes** des algorithmes des MA (mise en défaut sur lames pathologiques)
- Quid du marché actuellement? CellaVision
- Collaboration en cours

En conclusion

Data Challenge avec de belles perspectives : pédagogiques, technologiques

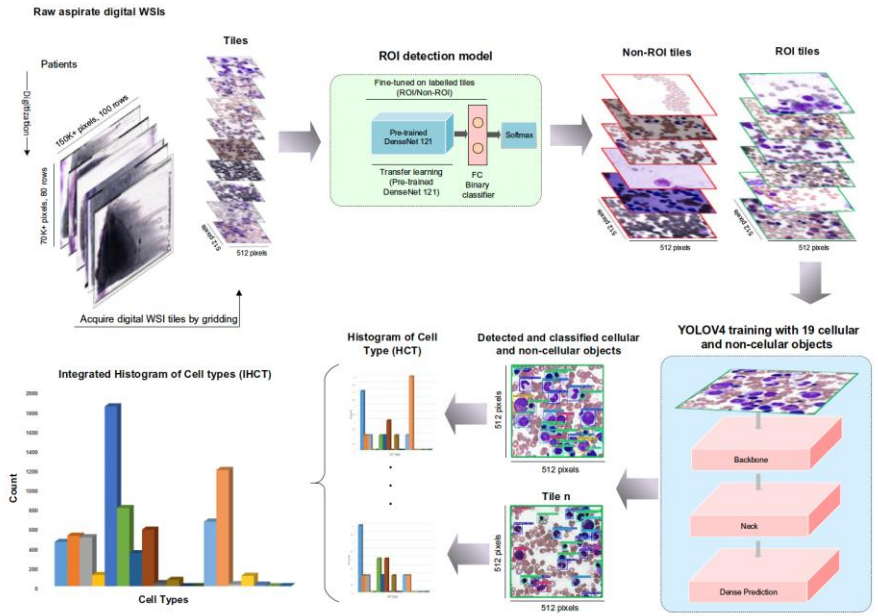
Base de données la plus variée et la plus complète à ce jour

Validation de l'algorithme sur un Dataset labellisé

Formation des futurs praticiens

Intégration dans modèles multi-omics avec d'autr

Perspectives : lames de myélogramme?



Tayebi, Commun Med 2022

HEALTH DATA HUB

PAUSE

15h40 - 16h00

Journée de l’open science en santé - Programme

15h40 - 16h00
Pause

16h00 - 17h00 **Présentation des résultats et remise des prix du Data Challenge CytologIA :**

- **3ème place**, Simon Thomine
- **2ème place**, Xueer Chen
- **1ère place**, Eric Ben Hamou

17h00 - 17h20 **L'open data, un catalyseur de l'engagement citoyen au service de la santé**, Augustin Courtier (Latitudes)

17h20 - 17h30 **Conclusion et remerciements**

17h30 - 19h00
Cocktail

Présentation des résultats et remise des prix du Data Challenge CytologIA

3ème place

HEALTH DATA HUB

Présentation des résultats et remise des prix du Data Challenge CytologIA - 3ème place

16h00 - 16h20

Simon Thomine

Ingénieur de recherche chez VitaDX

Simon Thomine est ingénieur de recherche en vision par ordinateur chez Vitadx, où il développe des solutions d'IA pour la pathologie, notamment pour la détection précoce du cancer de la vessie. Titulaire d'un doctorat en apprentissage profond, il est passionné par l'IA et l'enseignement de ses concepts. Ses travaux de thèse ont été publiés dans diverses conférences et journaux scientifiques.

CytologIA data challenge
3ème place
Simon Thomine (VitaDX)

4 Juin 2025

Journée Open Science en Santé

Sommaire

Qui suis-je ?

🎓Docteur en intelligence artificielle de l'université de technologie de Troyes

🤖 Ingénieur en deep learning chez VitaDX

🧬👨‍💻 Passioné par l'IA pour la pathologie et l'éducation sur le de deep learning

🌐 Site Web : https://simonthomine.github.io/
LinkedIn : www.linkedin.com/in/simon-thomine-287920170
GitHub : https://github.com/SimonThomine

A propos de VitaDX

IA et analyse d'images appliquées à la cytologie pour le diagnostic des cancers

VISIOCYT®

TECHNOLOGIE BREVETÉE

5 familles de brevets issus de la recherche française

VISIOCYT® Bladder

DISPONIBLE A LA PRESCRIPTION

Essai clinique multicentrique prospectif

Marquage CE IVDR

MEDIPATH Médecins pathologistes indépendants

Test disponible via un centre de pathologie partenaire

BIOGROUP biologie médicale

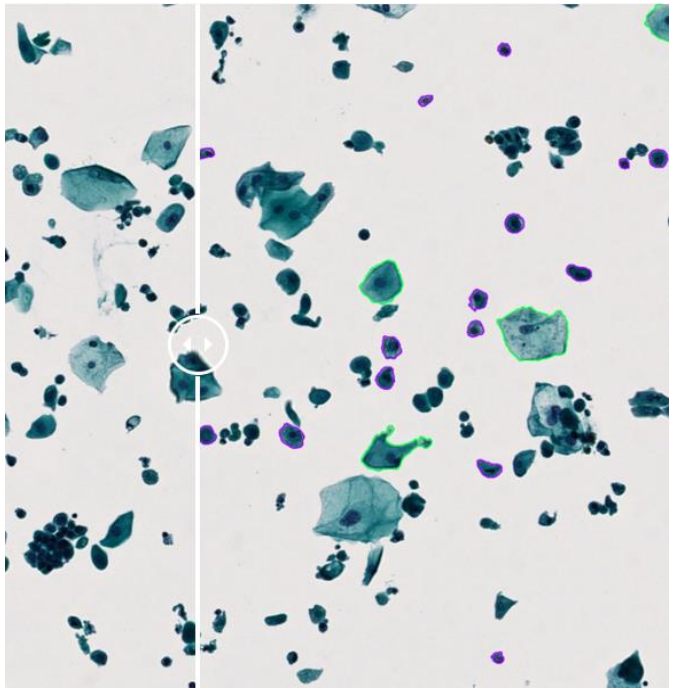
Recueil simplifié grâce à un partenariat avec un réseau de laboratoires de ville

IA et analyse d'image appliquées à la cytologie

VISIOCYT® Bladder

DIAGNOSTIC ET LA SURVEILLANCE DU CANCER DE LA VESSIE

CE-IVDR - Disponible à la prescription par les urologues



- **Test non invasif à partir d'un échantillon d'urine**
  En amont de l'endoscopie
- Algorithmes de machine learning et computer vision **prédisant le résultat de l'endoscopie / histologie**

	VISIOCYT®	Cytologie
Sensibilité Haut Grade	93,7%	62.8%
Sensibilité Bas Grade	66,7%	26.1%
Spécificité	61.8%	NA

- **Personnalisation de la prise en charge du patient**

Développer de nouvelles solutions

Cancer de la Vessie	Cancer de la Vessie	Cancer de la Thyroide	Autres pathologies
VISIOCYT® Bladder	VISIOCYT® Screening	VISIOCYT® Thyroid	VISIOCYT® ???
Pleinement opérationnel	***Développement en cours***	***Développement en cours***	***Réflexion en cours***
Essai clinique validé en 2021 CE IVDR	Reconfiguration de l'algorithme pour se concentrer sur la détection du cancer chez les patients asymptomatiques mais à haut risque	Lauréat appel à projet “Imagerie médicale” de France 2030 en collaboration avec Medipath	Diversification de l'activité de VitaDX à d'autres pathologies utilisant de la cytologie pour le diagnostic
Diagnostic et surveillance du cancer de la vessie	**Dépistage du cancer de la vessie**	**Diagnostic du cancer de la thyroïde**	**Sang, poumons, …**

Eléments du challenge

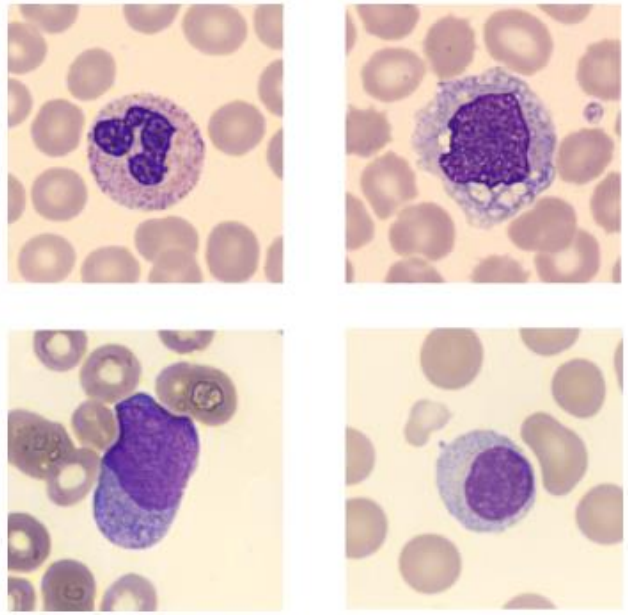
Problème de classification des globules blancs sur des patches issus de lames

Peu de datasets existants, sources différentes (microscope, scanner etc …), nombre de classes insuffisant

Dataset PBC (Acevedo at al.) : 17 092 images regroupées en 8 classes

Le dataset de la compétition CytologIA : 69 000 images et 23 classes distinctes

Ajout de l'aspect détection (pertinent si plusieurs cellules sur l'image)



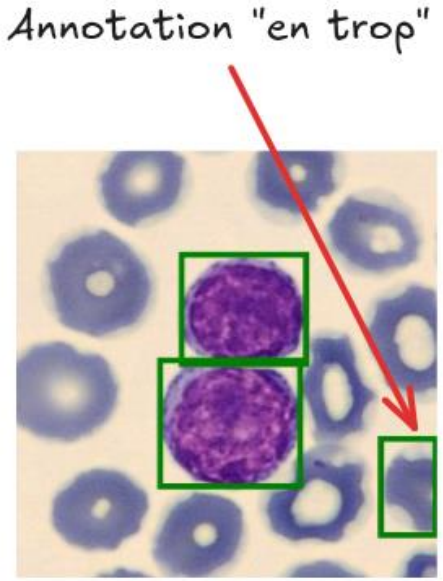
Challenge du dataset

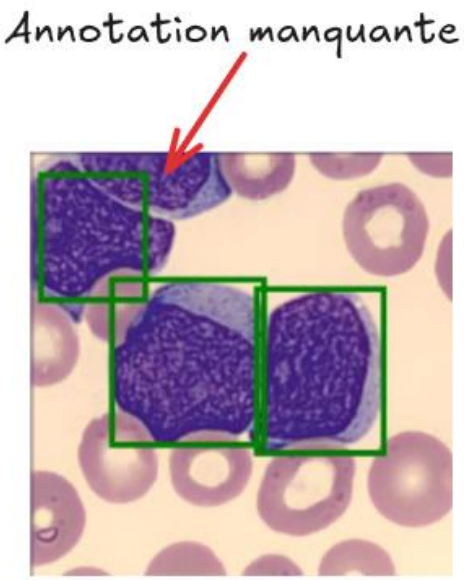
Comme souvent dans le domaine médical, le dataset est bruité

Bruit sur la classification : difficile à quantifier si on n'est pas expert du domaine

Bruit sur la détection : globules rouges annotés en globule blanc, globules blancs sur les bords non annotés

Un travail sur les données permettrait d'améliorer les performances





Spécifications

Durée du challenge : 1 mois et demi
Participants : 383 équipes

Evaluation Metrics

The performance of the submitted models will be assessed based on the following criteria:

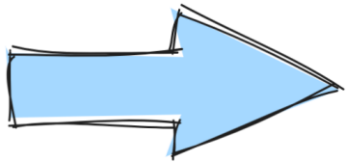
Bounding Box Accuracy:

Evaluated using **Generalized Intersection Over Union (GIOU)**, contributing **20%** to the final score.

Classification Performance:

Assessed using the **F1 Score**, contributing **80%** to the final score.

Précision de box englobante ne représente que 20% du score

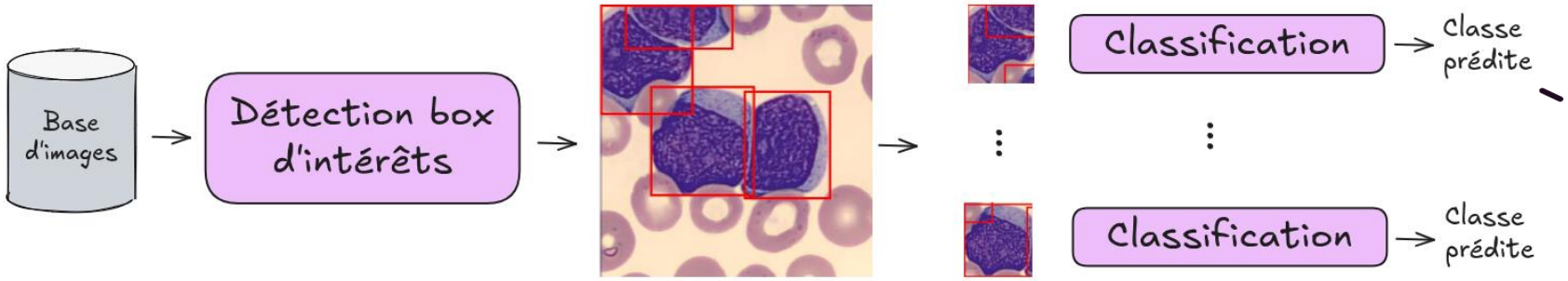


Logique car l'information précise n'est pas très importante par rapport à la classification

Il aurait également été possible de poser le problème comme de la "multi-label classification

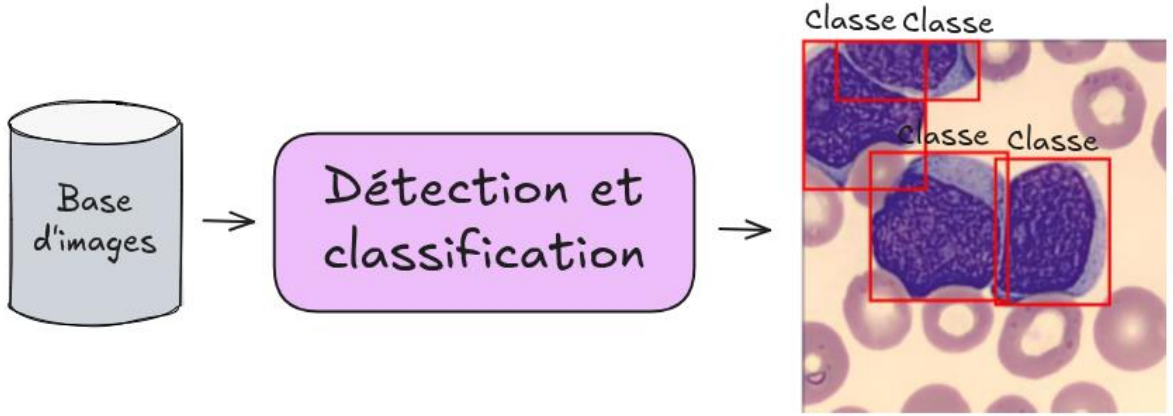
Choix des modèles

Détection puis classification



On peut prédire une classe dans la partie détection et combiner les classes des deux modèles

Détection seulement

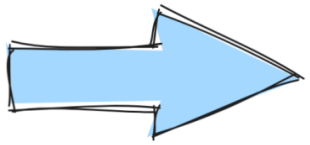


Cette méthode est plus rapide et plus simple à entraîner

Complexité et rapidité

Inference Timing:
Each image must be processed with an inference time of **less than 500 ms**. While GPU usage is permitted to accelerate processing, the solution must be compatible with regular GPUs typically available in standard setups, without requiring specialized hardware like high-end server-grade GPUs. Solutions optimized for **CPU-based** inference are preferred to ensure broader applicability, but CPU optimization is not mandatory.

Plutôt que d'utiliser une architecture très complexe et lente, on va plutôt chercher à optimiser les résultats sur une méthode plus simple

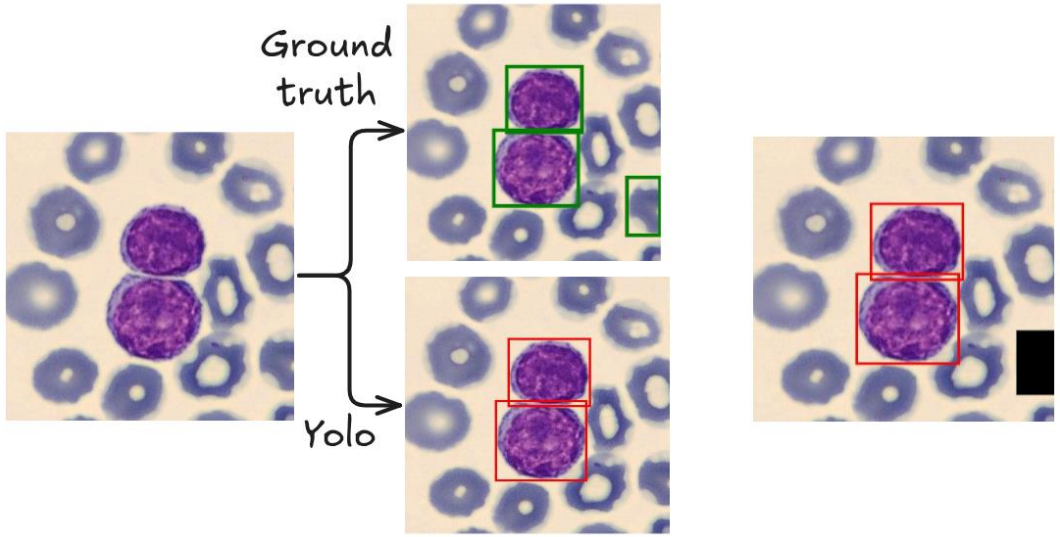


Notre choix s'est donc porté sur la seconde méthode

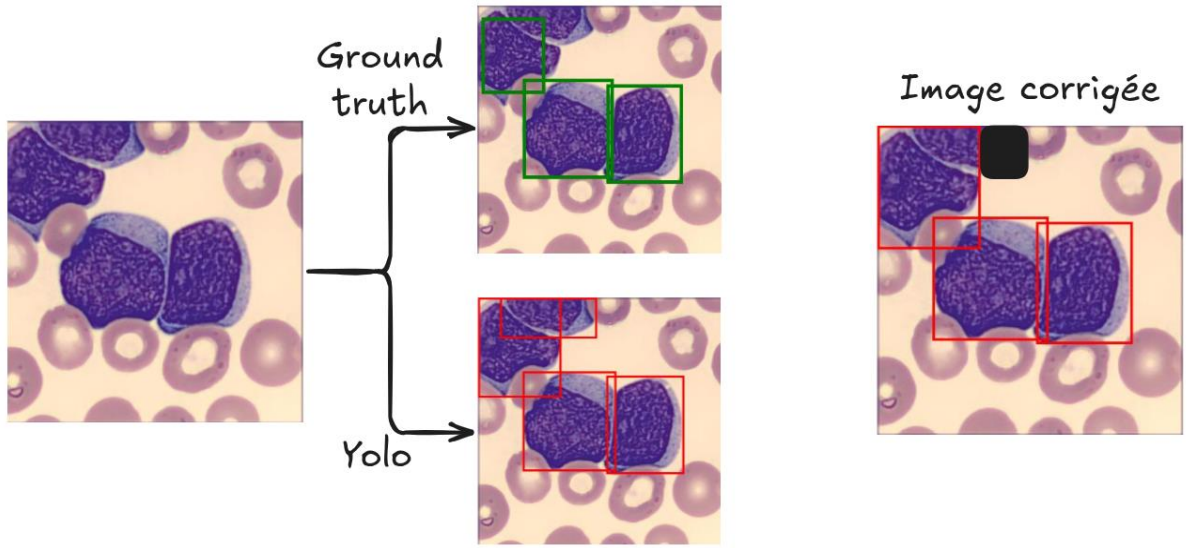
Approche centrée sur les données

Deux méthodes de filtrage des données

Une erreur d'annotation est présente (un globule rouge est annoté comme un globule blanc)

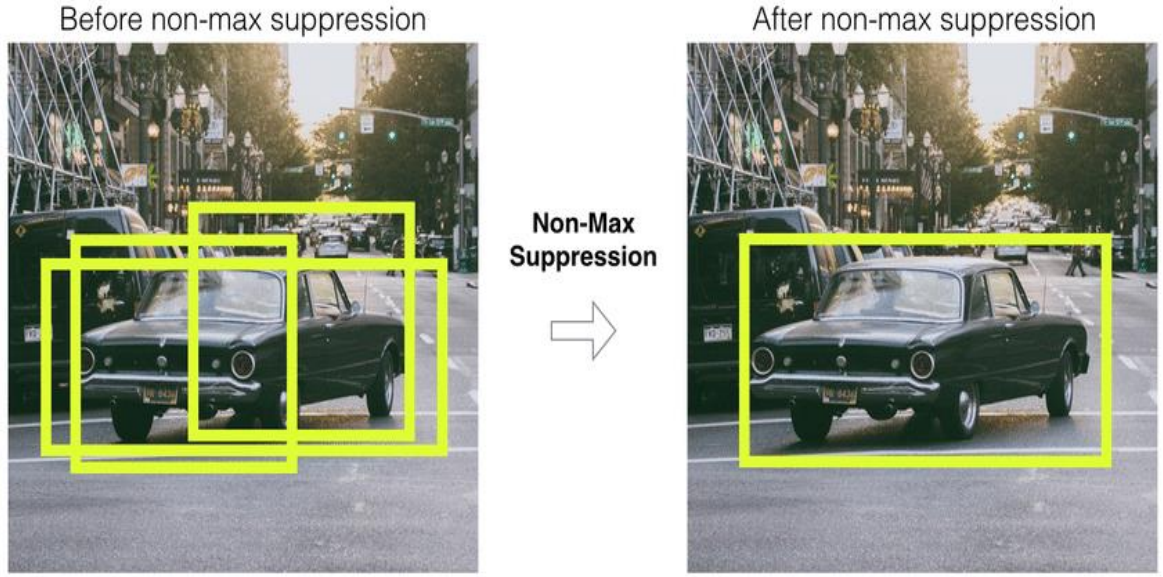


Les globules blancs sur les bords sont parfois non-annotés ce qui peut perturber l'entrainement du modèle

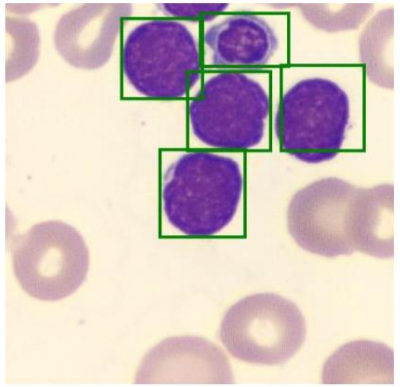


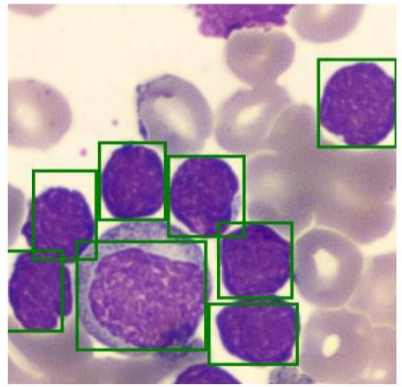
On se base sur un modèle yolo entrainé sur les données non filtrées

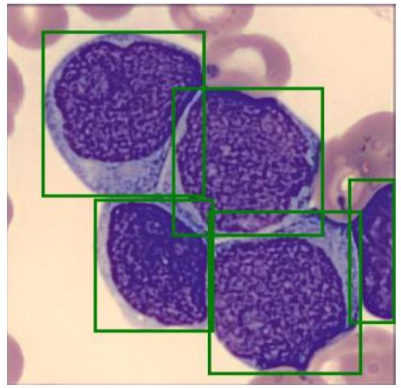
Non-maximum suppression

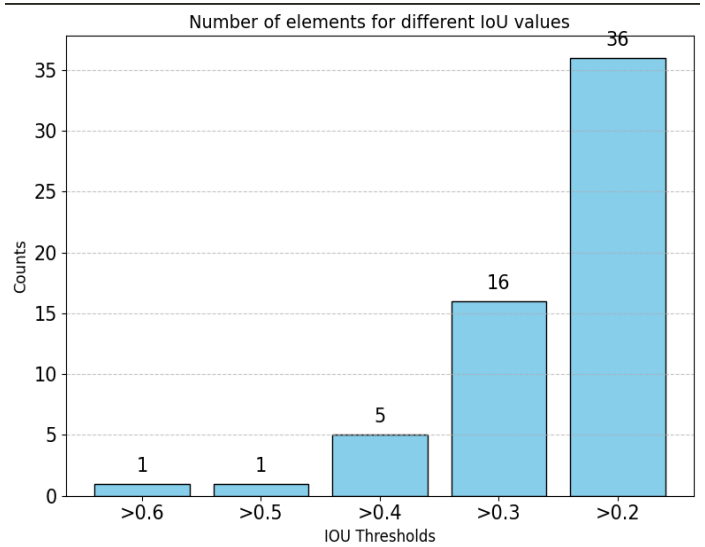


Sur une cytologie, les cellules sont rarement superposées les unes sur les autres



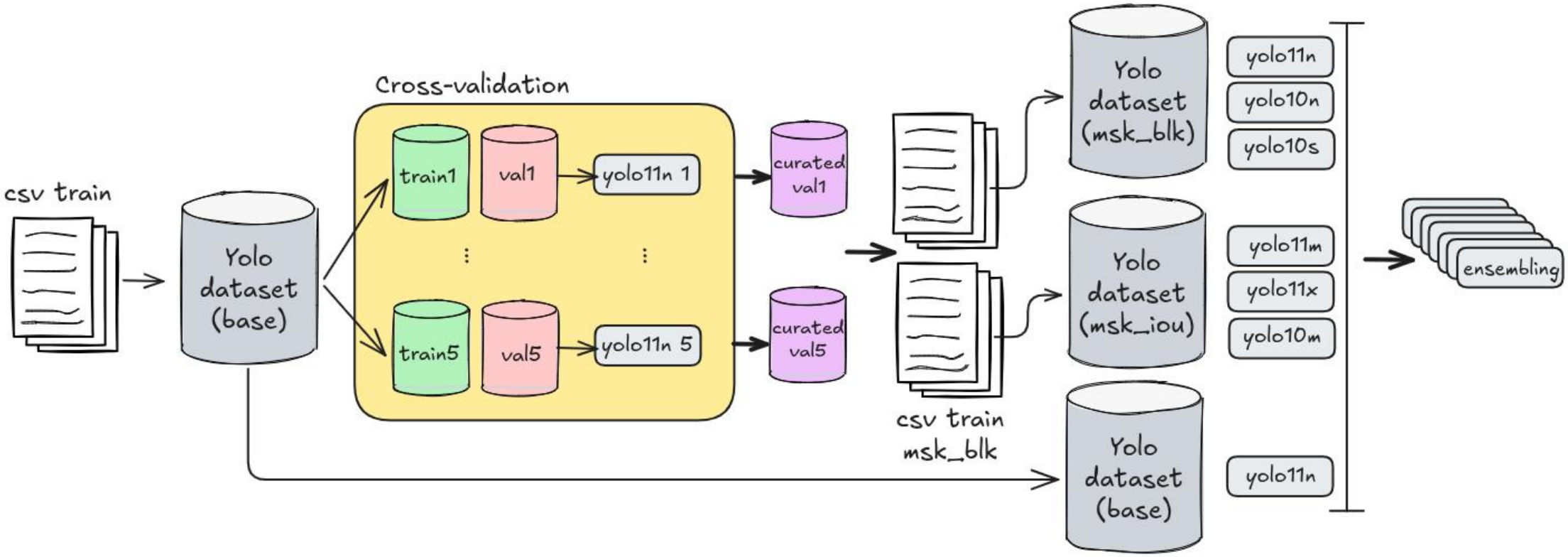






On choisit un seuil de non-maximum suppression assez faible

Architecture globale



L'ensembling est utilisé pour rivaliser avec les meilleures performances du challenges : un modèle simple sans ensembling est quand même très performant

https://github.com/SimonThomine/Cytologia-Data-Challenge

Résultats

Les différences de performance entre les différents concurrents sont très faibles (presque négligeables)

La différence notable entre les premiers du classement se fait sur le temps d'inférence

Publique

Rank	Members	Team	Score
2	2	Sheoran	0.94165777356
3	1	MPWARE	0.93930585557
3	1	xueer chen	0.93923078751
4	1	Simon Thomine (team)	0.93858266795
5	2	Roberta Becker	0.93674660980

Privé

Rank	Members	Team	Score
2	1	MPWARE	0.9371687102302637
3	1	xueer chen	0.9361097877099107
3	1	Simon Thomine (team)	0.9359152974124093
4	2	Roberta Becker	0.9342641403850803
5	2	Sheoran	0.933979747852665

Temps d'inférence

RTX 4080:

- **Best Ensemble** (7 YOLO models): **40ms** → **0.9385** on the public leaderboard
- **Single YOLOm Model** (`yolo11m384_iou.pt`): **5ms** → **0.9330** on the public leaderboard
- **Single YOLOs Model** (`yolov10n384_blk.pt`): **4ms** → **0.9320** on the public leaderboard

- On est bien en dessous des 500ms maximum du challenge
- Un modèle unique est beaucoup plus rapide et a des performances très proches

Remarques

Points positifs

- La plus grosse base de données en hémato sur le scanner Cellavision
- Base de données complète : 23 classes, cas avec plusieurs cellules sur une même image
- Support de Trustii réactif aux questions des participants

Points négatifs

- Nombre de cellules à détecter connue à l'avance (absurde en condition réelle)
- Score du leaderboard non transparent (F1 et IoU)
- Plateforme lente et mal pensée : forum inefficace ce qui limite la collaboration et les discussions
- Les annotations potentiellement générées par une IA et mal relues

Des questions ?

Merci pour votre attention

VitaDX : https://vitadx.com
Site Web : https://simonthomine.github.io/
LinkedIn : www.linkedin.com/in/simon-thomine-287920170
GitHub : https://github.com/SimonThomine

Presentation of the results and award ceremony of the CytologIA Data Challenge

2nd Place

HEALTH DATA HUB

Presentation of the results and award ceremony of the CytologIA Data Challenge - 2nd Place

16h20 - 16h40

Xueer Chen

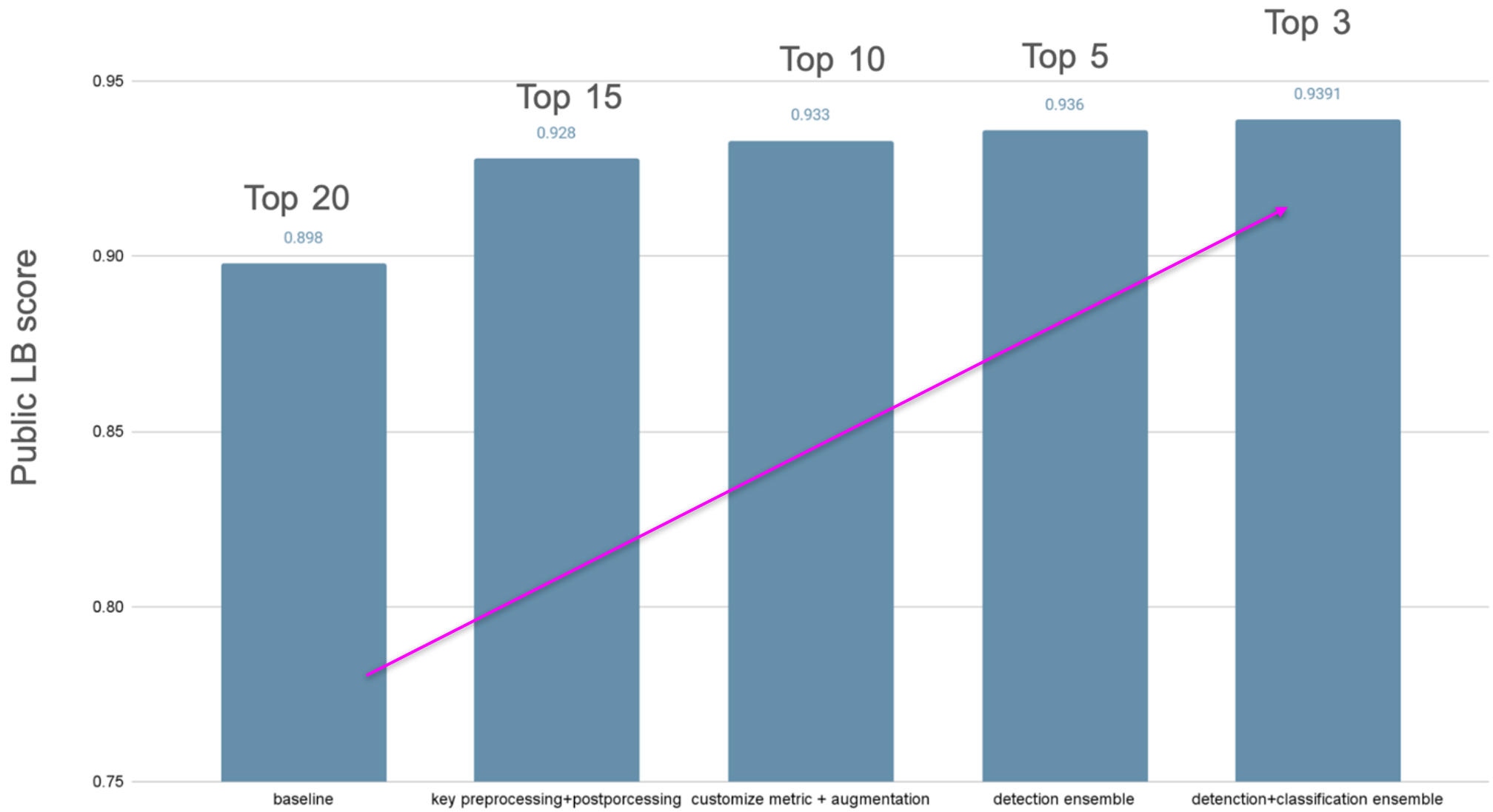
Senior Scientist - Computational Biology, Bristol Myers Squibb

Xueer Chen is a dedicated Senior Scientist at Bristol Myers Squibb, driving innovation in Translational Predictive Science. She is deeply engaged in pioneering work with single-cell multi-omics and architecting cutting-edge multi-modal foundation models that seamlessly integrate complex data like digital pathology and transcriptomics. Fueled by her extensive expertise in computational biology, cancer immunology, and generative AI, Dr. Chen is committed to pushing the boundaries of predictive science to unlock new therapeutic possibilities. Her exceptional machine learning capabilities are highlighted by her achievements as a Kaggle Competitions Master, consistently excelling in challenging data science competitions.

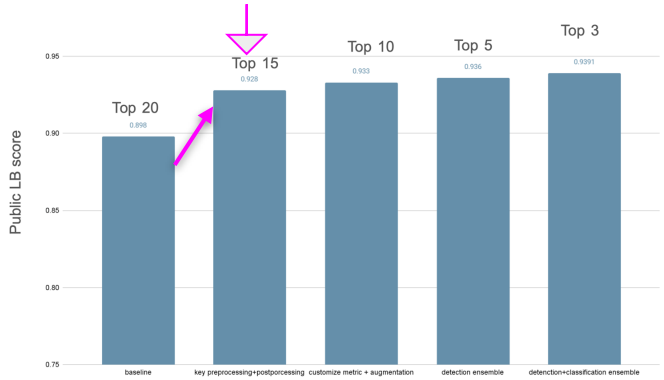
Ensemble Learning for Leukocyte Detection & Classification: CytologIA 2nd Place Solution

Xueer Chen

From Baseline to Top 3: Performance Evolution



Key Preprocessing and Postprocessing



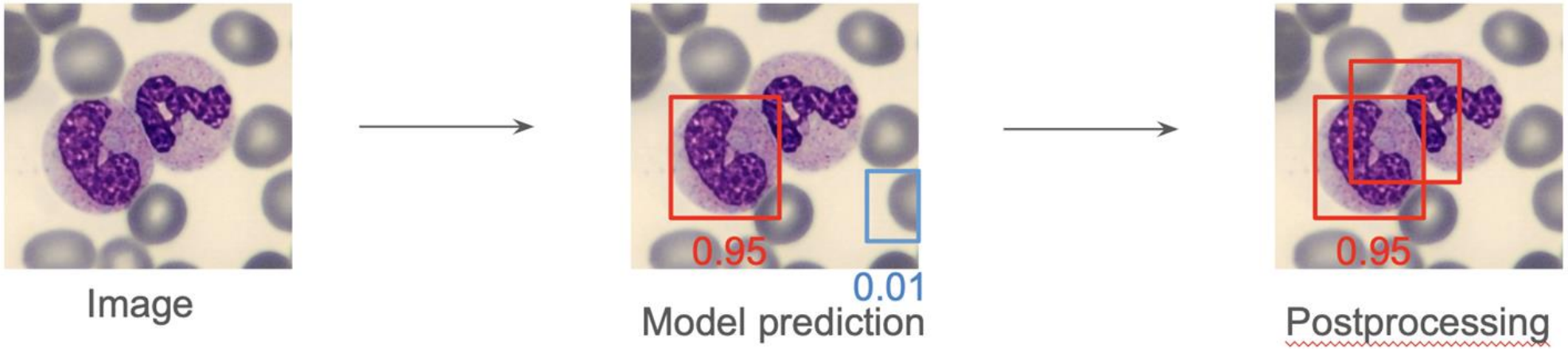
Preprocessing: fix corrupted jpeg images

train: *"WARNING: Ignoring corrupted image and/or label xxx.jpg: corrupted JPEG"*

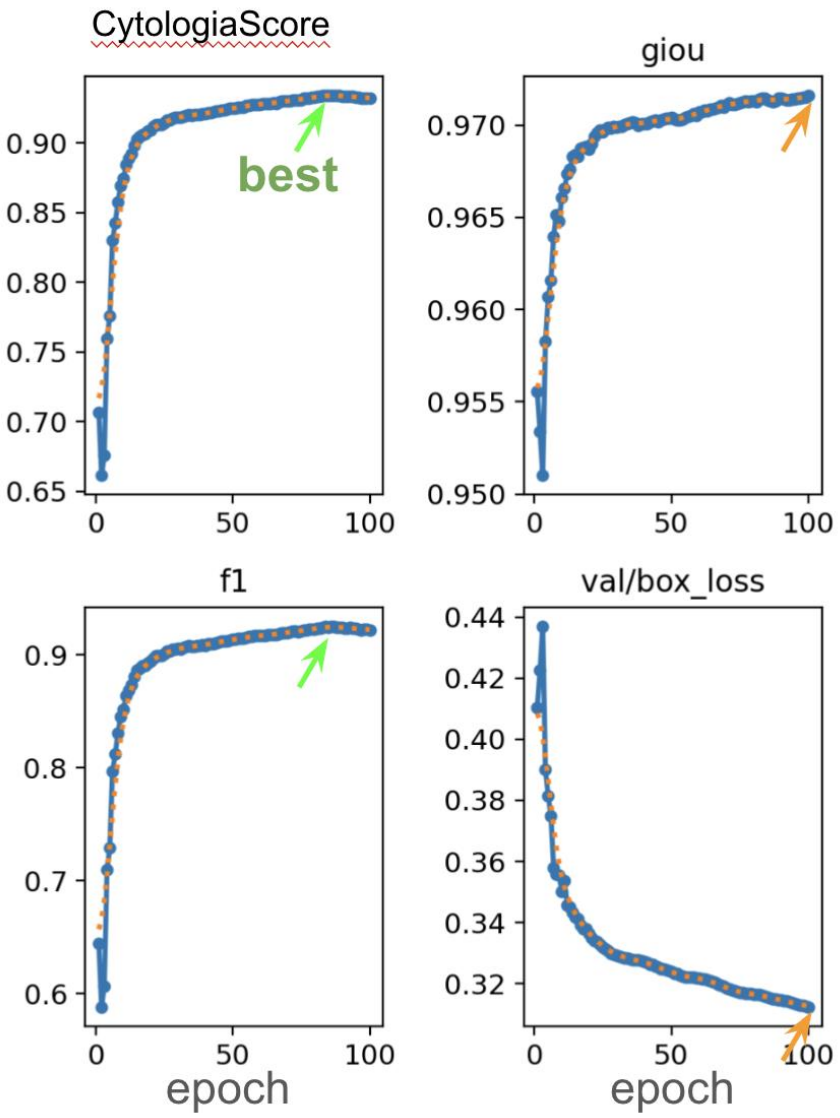
```
img = cv2.imread(img_path)
```

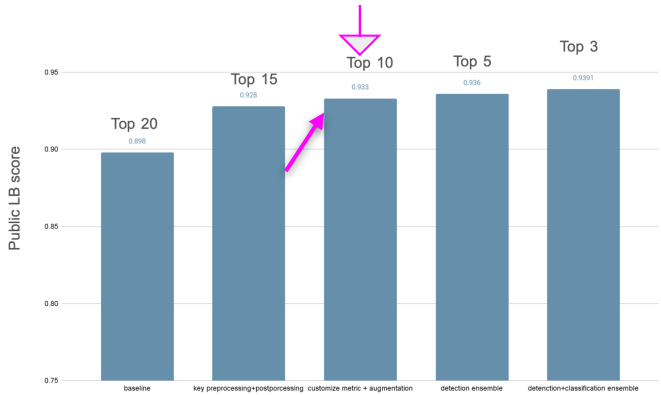
```
cv2.imwrite(img_path, img, [int(cv2.IMWRITE_JPEG_QUALITY), 100])
```

Postprocessing: duplicate top predictions with random jittering



Custom metric for optimal checkpoint selection





The Cytologia Metric:

```
CytologiaScore = 0.2 * Mean_GIoU + 0.8 * Macro_F1
```

Components of CytologiaMetric:

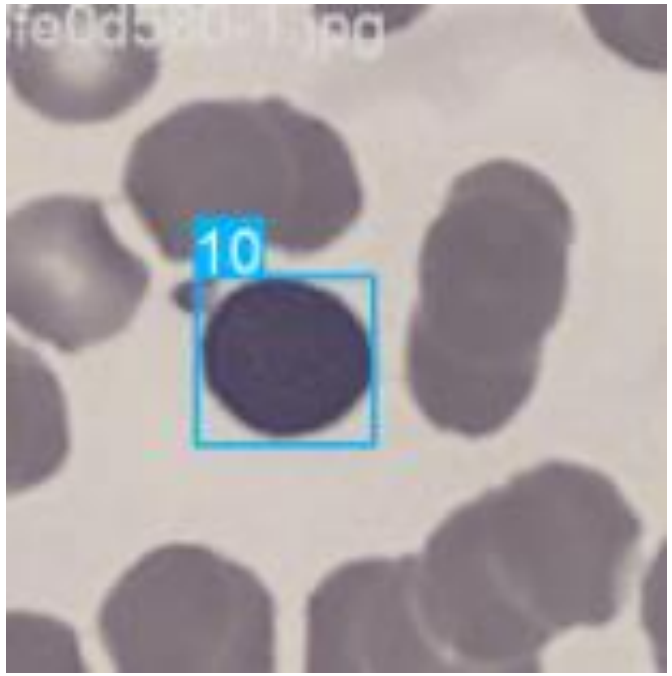
- Generalized Intersection over Union (GIoU):
  GIoU measures the overlap between predicted and ground truth bounding boxes.
- Macro F1 Score:
  The F1 score is the harmonic mean of precision and recall. The Macro F1 score calculates the F1 score for each class independently and then averages them. This is crucial for handling class imbalance and ensuring good performance across all object categories.

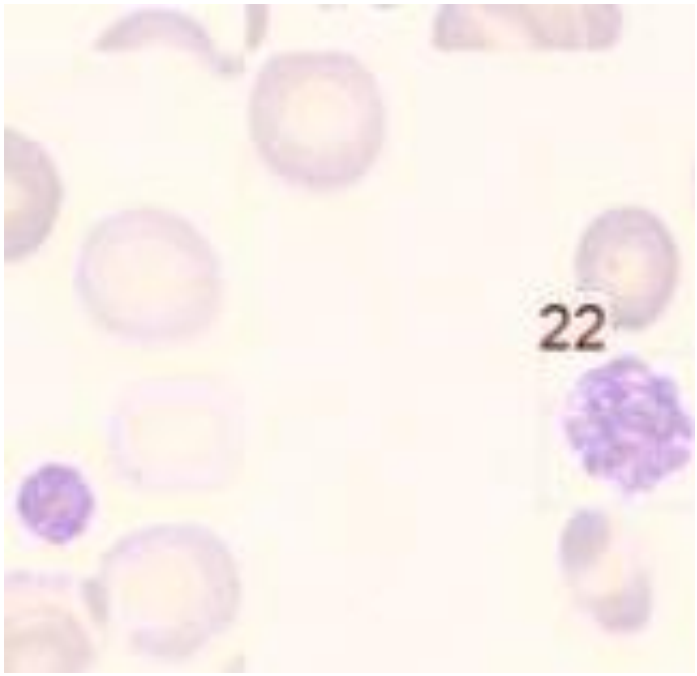
Image Augmentation: less is more

Turn off:

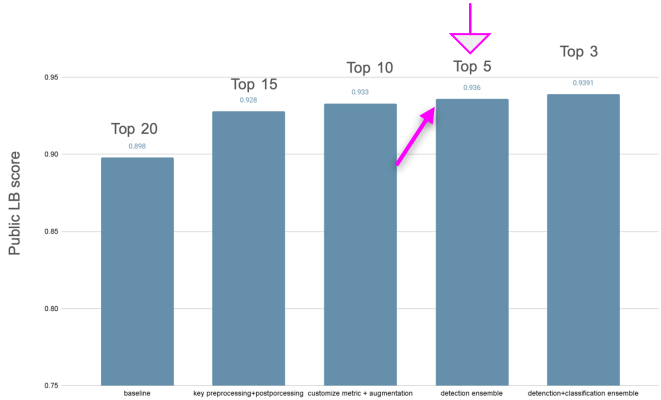
- Color space: hue, saturation and brightness

Examples of **bad augmentation** images if turned on hue, saturation and brightness





Stage1 - Detection Model Details



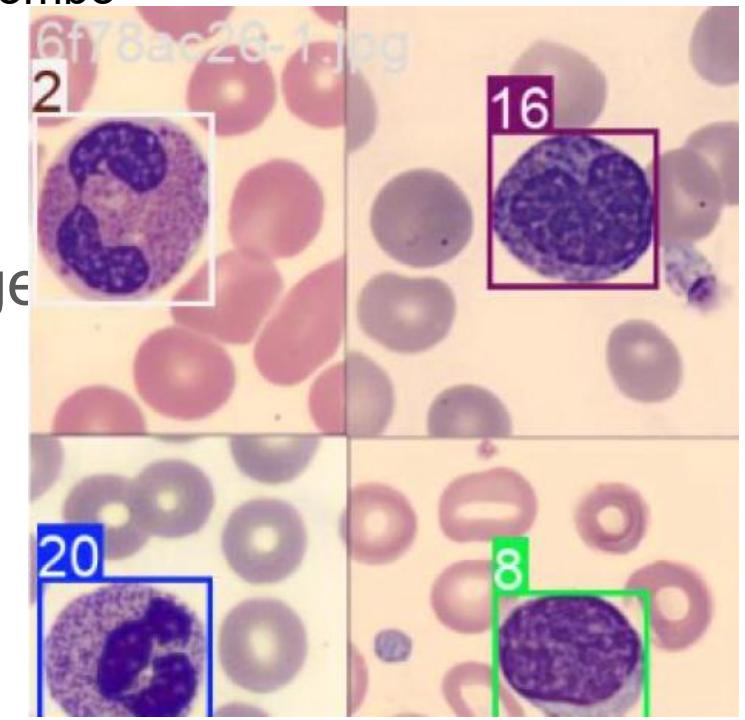
Architecture: YOLOv9c (SOTA object detection framework) fine-tuned on competition data

Training Strategy: 4-fold CV with data source-aware split

Custom validation metric: Replaced YOLO mAP (overfitting) with GIOU + F1 combo

Data Augmentation:

- Mosaic (shown with image):
  a. Combines four randomly selected and resized image
  b. Increases training diversity
  c. Helps prevent overfitting
- Random flip
- No HSV (less useful for medical images)



Improving Detection Ensemble with modified NMS

Problem: YOLO default doesn't provide class probability distribution

Fix: Modified non-maximum suppression function NMS to retain full class probabilities → allows soft ensembling

Better than voted-ensemble

Before NMS:

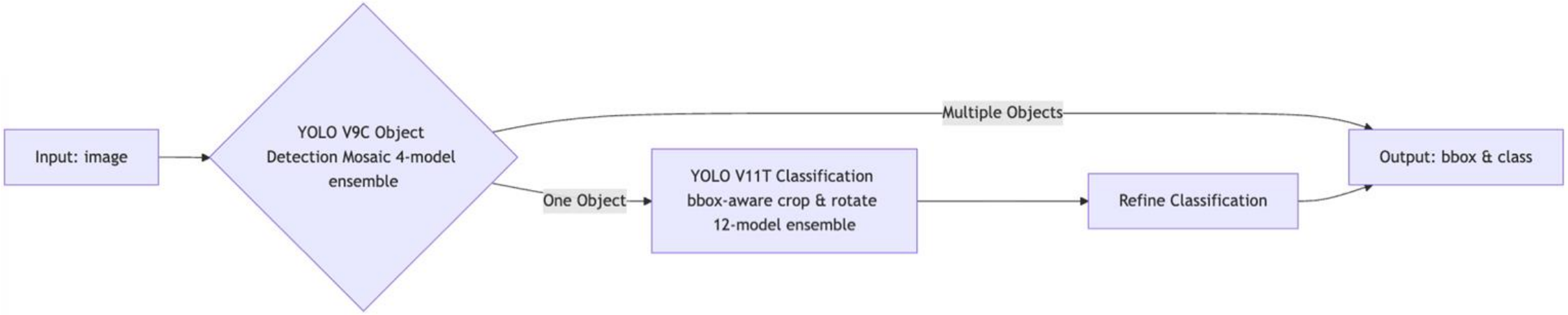
Box ID	Confidence	Overlap (IoU)	
Box A	0.95	–	
Box B	0.93	0.85 with A	← Removed
Box C	0.60	0.1 with A	← Kept

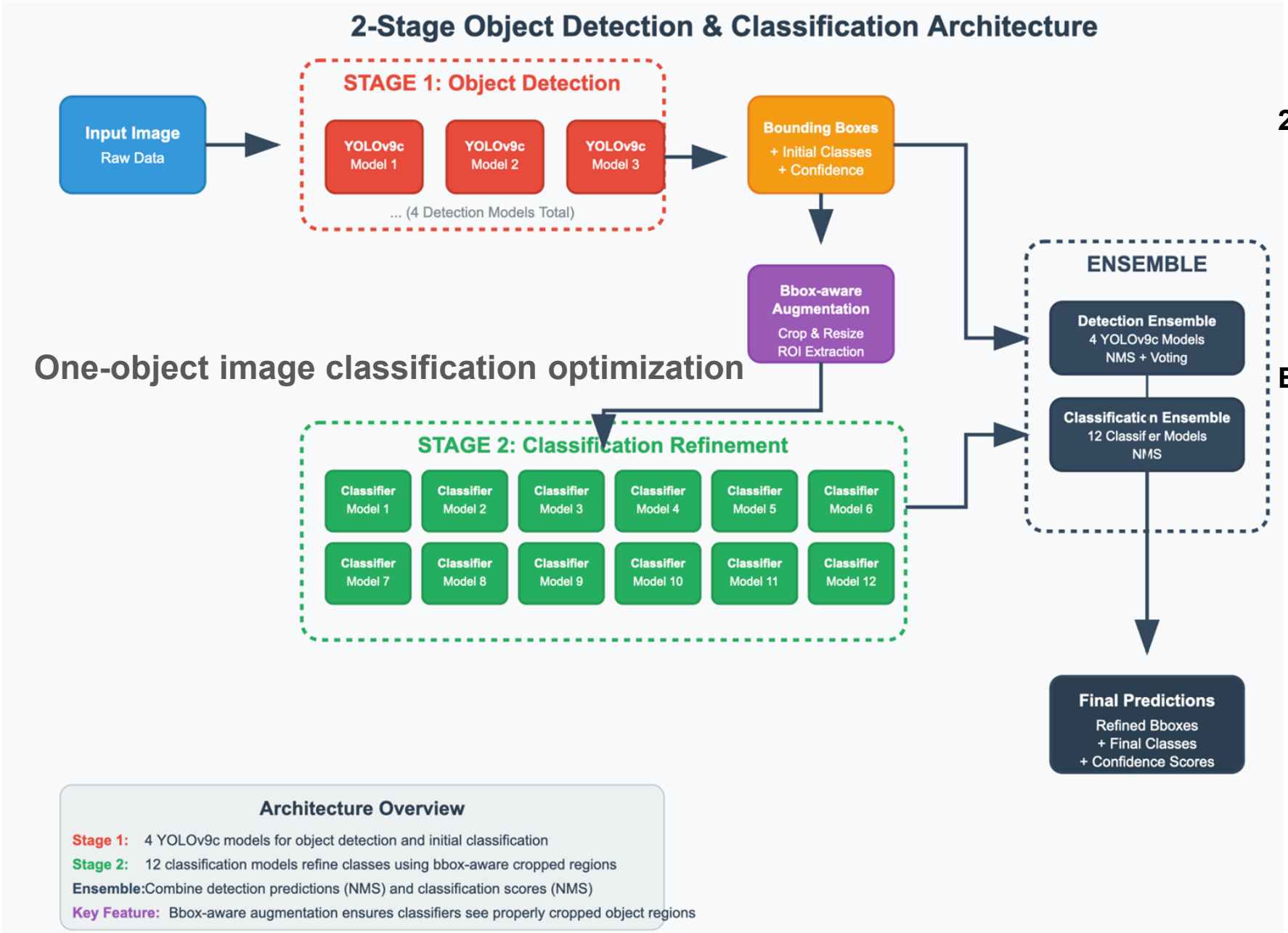
After NMS:

Keep Box A and Box C

Key Observation lead to 2-stage solution

1. 85% of images contain only one object
2. Bounding box prediction highly accurate for these images (97% GIoU)
3. Classification could be improved beyond 92% F1 score
4. Classification represents 80% of competition metric





One-object image classification optimization

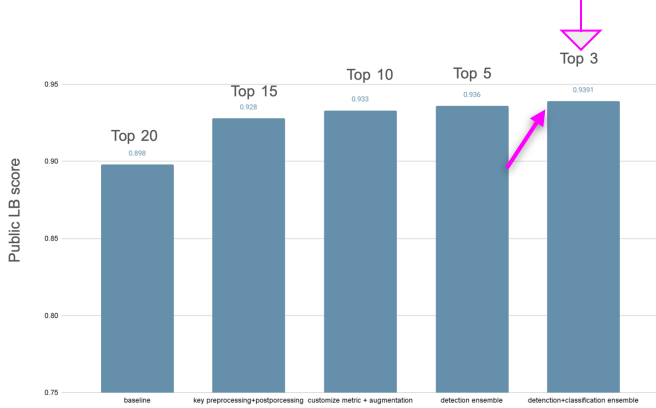
2-Stage Approach:

- Stage 1: YOLOv9c-based object detection
- Stage 2: Classification refinement using bbox-aware augmentation

Ensemble:

- 4 detection + 12 classification models
- Final inference time: ~300ms/image on T4 GPU

Stage2 - Classification Model Improvements



Purpose: Refine classification for single-object images

Model: YOLOv11t pretrained classification model

Training: Same 4-fold cross-validation strategy

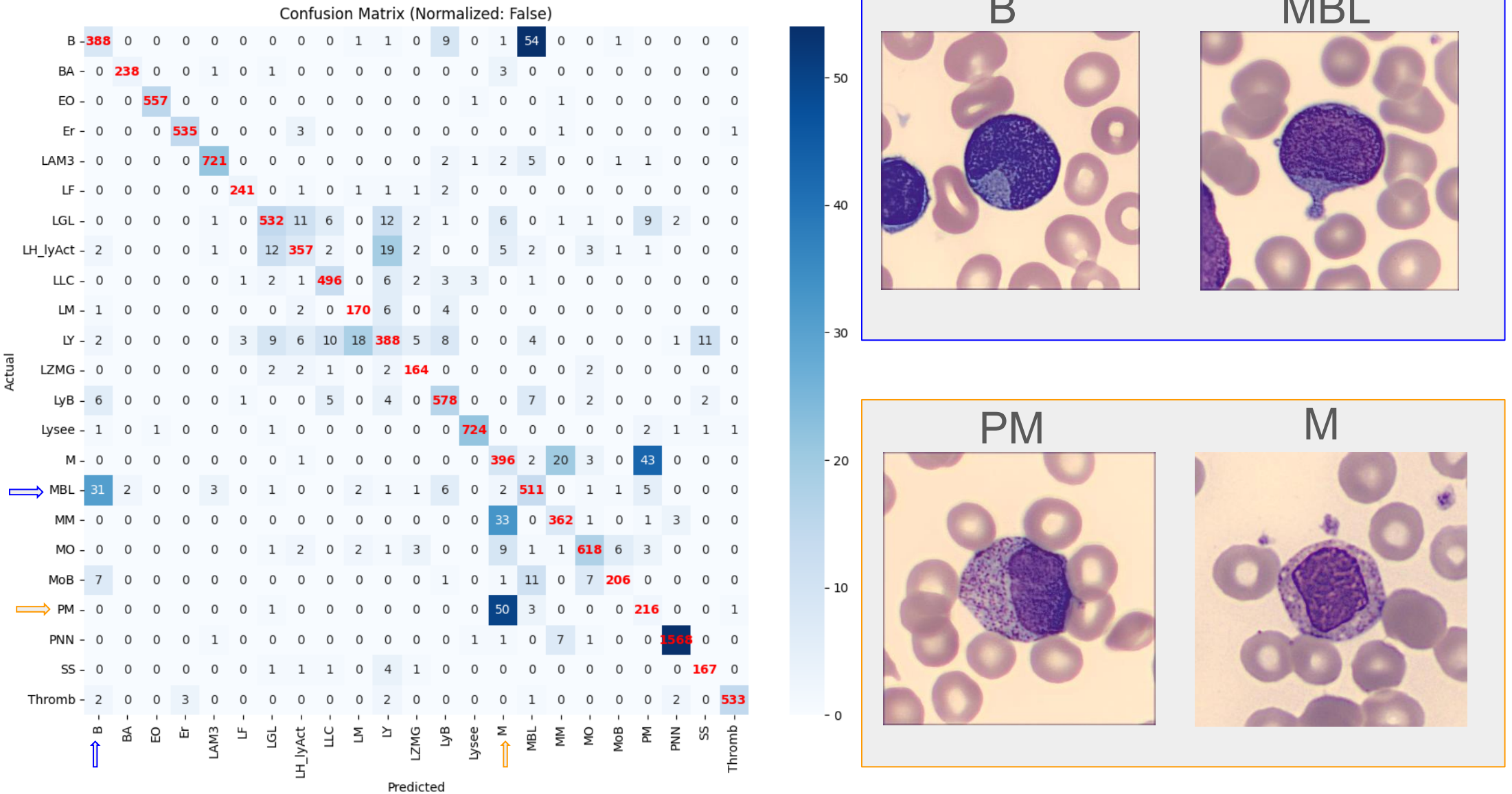
Metric: F1 score for checkpoint selection

- **Focal Loss**: Better optimization for F1 score with macro average
  - Addresses class imbalance
  - Focuses on difficult examples
- **Bounding Box-Aware Augmentation**:
  - Preserves regions of interest during augmentation
  - Prevents loss of critical cell information

$$FL(p_t) = -\alpha_t * (1 - p_t)^\gamma * \log(p_t)$$

- `p_t` is the model's estimated probability for the true class
- `α_t` is a weighting factor for class balance
- `γ` (gamma) is the focusing parameter

Confusing classes

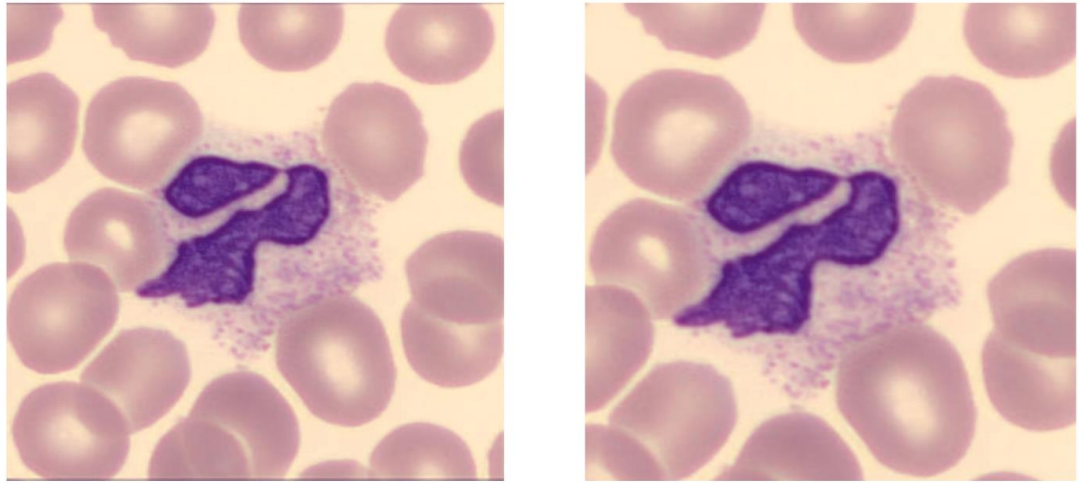


Bounding Box-Aware Augmentation

Left: Original image

Right: Cropped image preserving region of interest

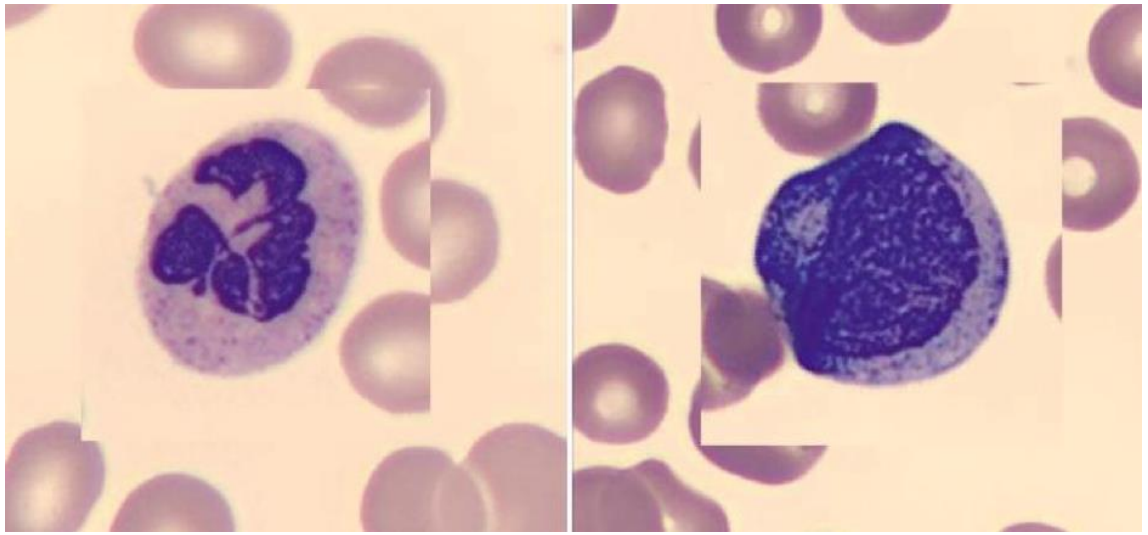
Cell region slightly enlarged and shifted



Crop bounding box region

Rotate and paste back onto original background

Enables learning different cell poses



Results & Achievements

2nd place in competition

High detection accuracy (>97% GIoU)

Improved classification through ensemble approach

Efficient inference (300ms per image)

Key Takeaways

- Ensemble of diverse models (detectors + classifiers) improves F1
- Custom metrics and bbox-aware augmentations are critical

Practical, reproducible pipeline deployable in cloud environments

Resources & Acknowledgments

GitHub: https://github.com/xueerchen1990/cytologia_2nd_place

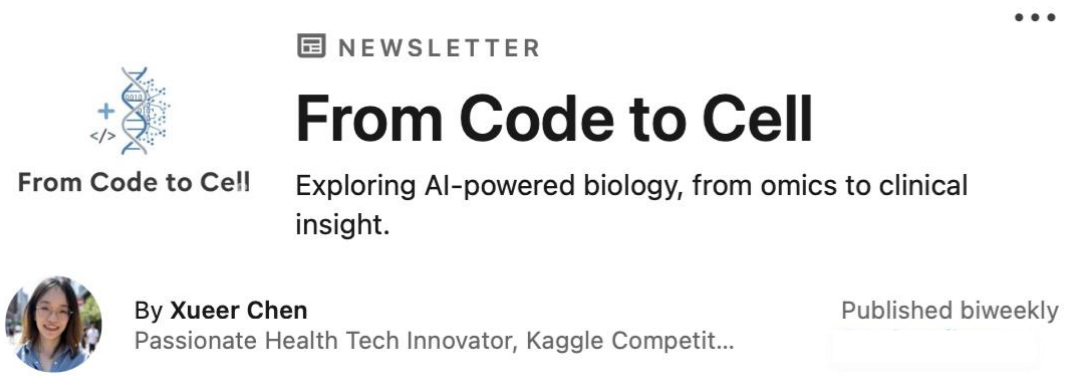
Google Colab (try with your own data): [Link]

Thank you! Trustii.io & partners

Where to find me:

Linkedin: https://www.linkedin.com/in/xueer-chen-02b51a2a/

My Health AI Newsletter on Linkedi



Présentation des résultats et remise des prix du Data Challenge CytologIA

1ère place

HEALTH DATA HUB

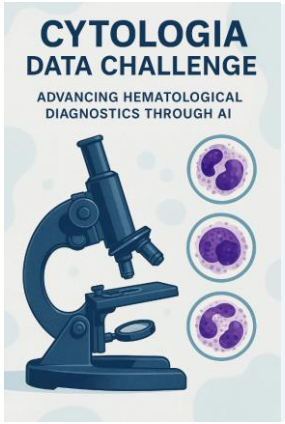
Présentation des résultats et remise des prix du Data Challenge CytologIA - 1ère place

16h40 - 17h00

Eric Ben Hamou

Senior Software Engineer &
Data Scientist

Avec 25 ans d'expérience, dont plus de 15 ans dans l'industrie aéronautique, j'ai également évolué dans les secteurs pharmaceutique, bancaire, télécoms et public. Je conçois et développe des solutions logicielles robustes et innovantes, déployées en production et utilisées au quotidien par des clients internationaux. Spécialisé en architecture logicielle et en science des données, j'ai contribué pendant 5 ans aux activités d'innovation et de recherche au sein d'un département R&D. Depuis près de 10 ans, je mets en œuvre des approches de machine learning en vision par ordinateur (médical/santé, environnement), séries temporelles (capteurs, énergie) et traitement du langage (rapports d'incidents, classification, LLM). Passionné par la résolution de problèmes concrets et curieux des évolutions technologiques, je participe régulièrement à des challenges Data/ML, dans une démarche d'apprentissage par la pratique.



White Blood Cell Detection & Classification

A Two-Stage Machine Learning Approach

1st place overview

Trustii.io

Agenda

Two-stage approach
Cross-Validation strategy
Object Detection
Classification
Ensemble and Inference
Takeaway and Next steps

Eric BEN HAMOU
Senior Software Engineer & Data Scientist

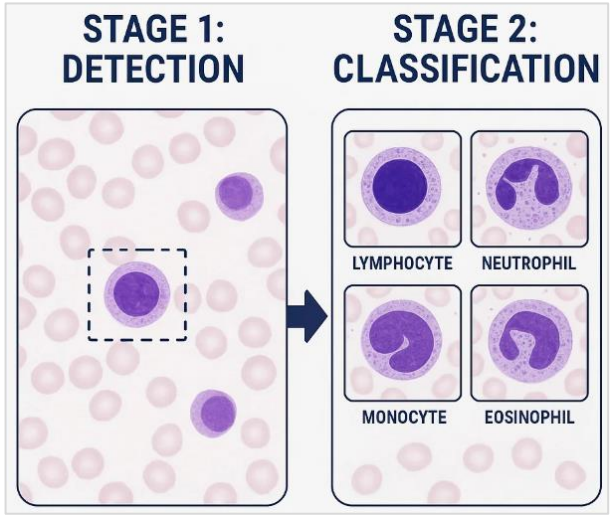
Kaggle Grandmaster x2 (competitions + discussions)
Fifth French reaching this level (2021)

Hugging Face Github

Two-Stage Approach

Solution:

- Stage 1 (Detection): Universal White Blood Cell localization (single-class) – Easy task
- Stage 2 (Classification): Specialized multi-class models – Difficult task



All-in-one models (e.g. YOLO family) are fast and effective for general tasks but could struggle with fine-grained classification like WBC subtype differences.

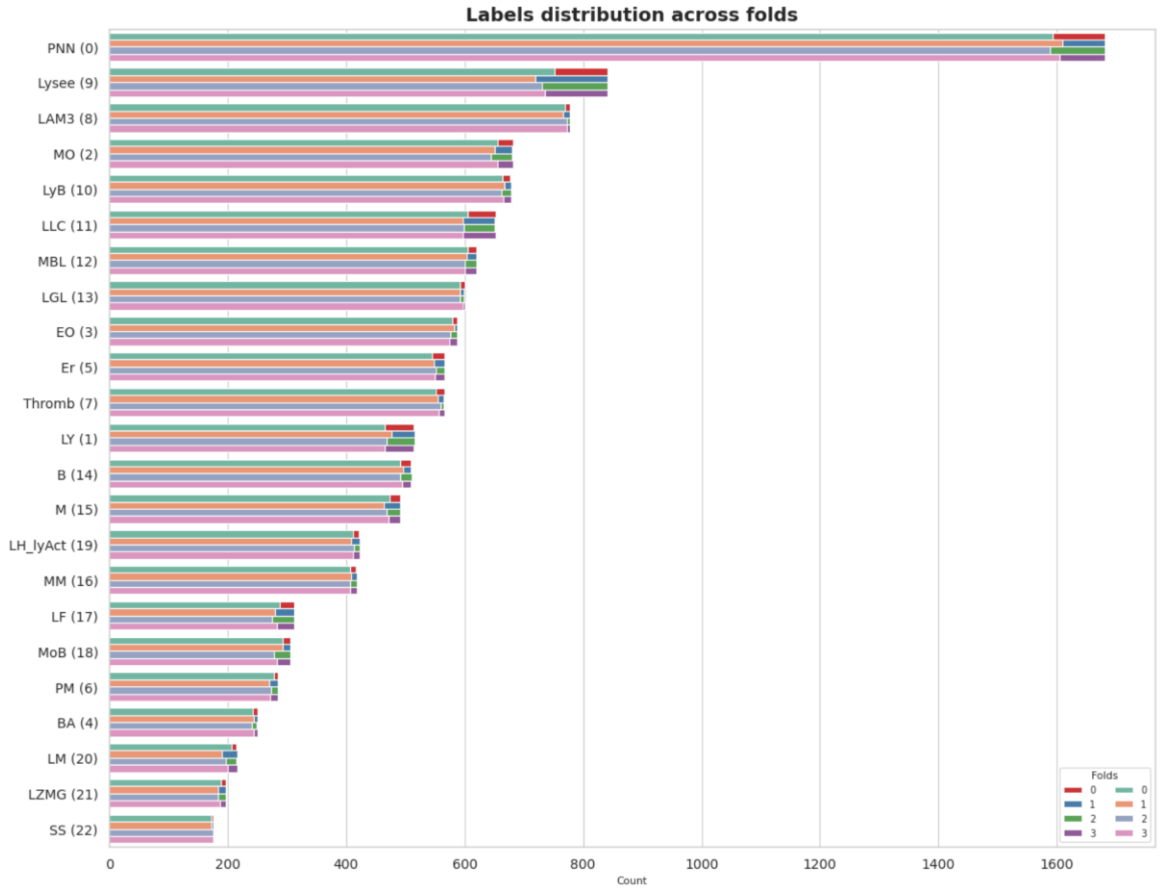
Other motivations:

- This approach ranked Top 3 (prize money) in a previous competition (paper in progress)
- Many competitors will use all-in-one models as it’s the state of the art
- Time limited (6 weeks) - Modular design allows for rapid experimentation and upgrades
- Resources limited

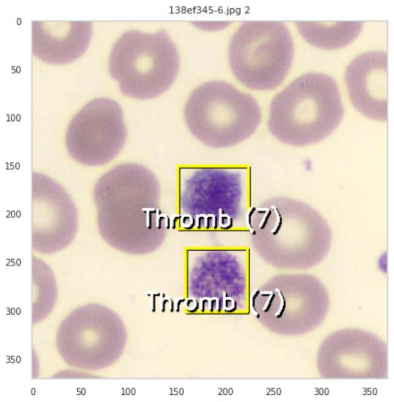
Cross-Validation Strategy

Key points:

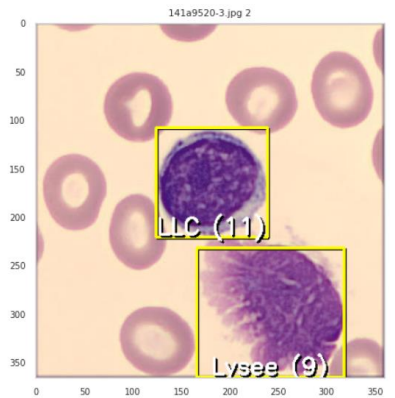
- Remove duplicate images from train to prevent leaks (154 images removed)
- Multi-labels stratification (label, image mean/std) to ensure balanced representation of complex cases
- 4 folds (75%/25%)



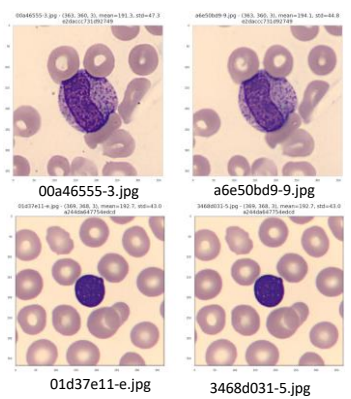
Total and unique (images) labels per fold



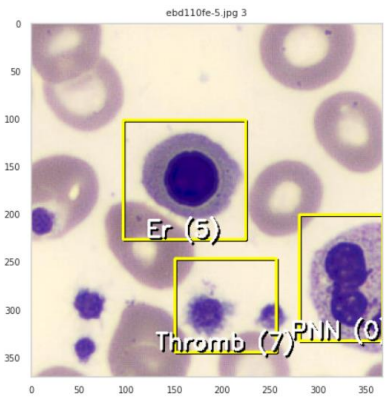
Single label



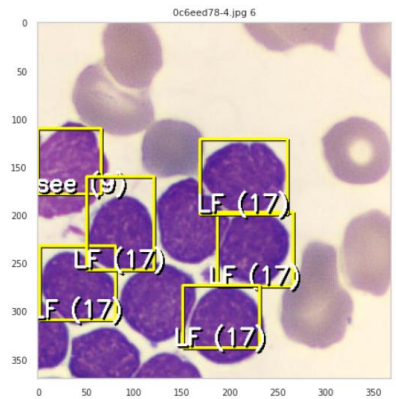
Two labels



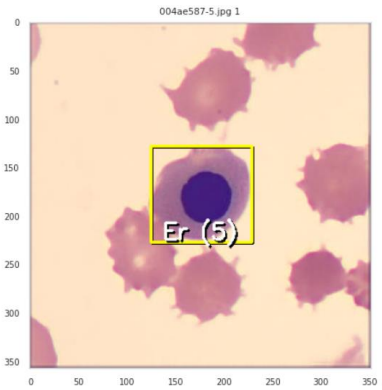
Duplicated images



Three labels



Missing labels



Single label and instance

Object Detection

Key points:

- Fix bounding boxes boundaries (1000 boxes out of the image across 927 images)
- Generate background images (around 3.5k) to reduce False Positives
- Remove noisy images (28 hard samples with bad bounding boxes)
- YOLOX modified to enable with GIoU loss (YOLOX preferred to YOLO for licensing)
- Two detectors (GIoU-maximizing vs. clean boxes). Challenge metric is 0.20xGIoU + 0.80xF1
- Training: Augmentations (MixUp, Mosaic, H/V Flip), 150/110 epochs

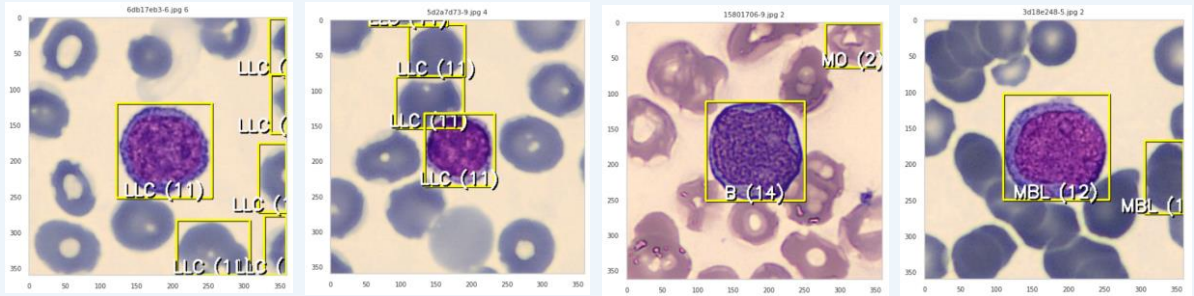
1

Label	Images	Boxes
0	48263	52586
Total	48263	52586

Object detector (YOLOX-S 512) mAP@0.95 = 0.943

Maximize GIoU

Find noisy labels with OOF confidence



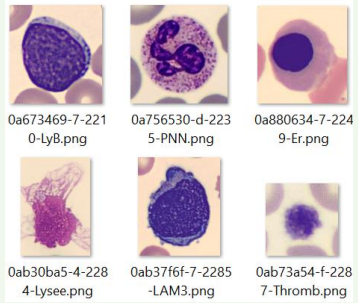
Noisy labels **Noisy labels** **Noisy label** **Noisy label**

2

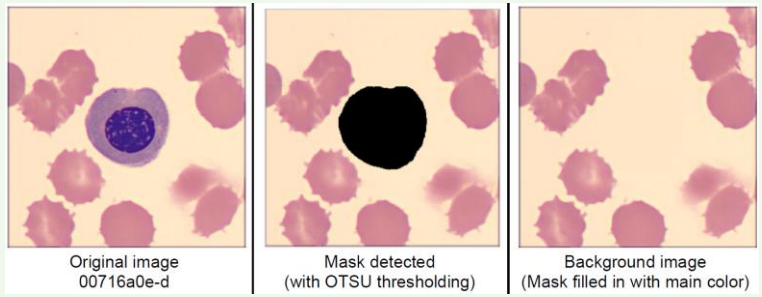
Object detector (YOLOX-S 640) mAP@0.95 = 0.943

Noisy images removed
Background images added

Match OOF with ground truth based on GIoU



Cleaned boxes for classifiers

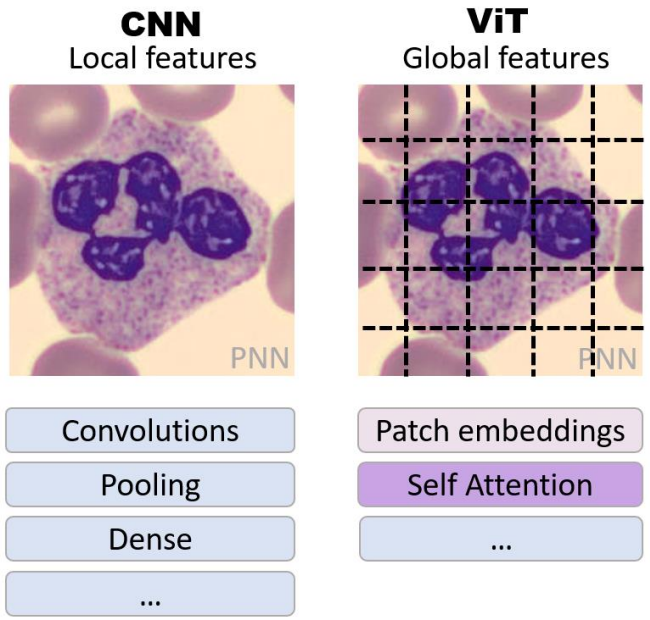


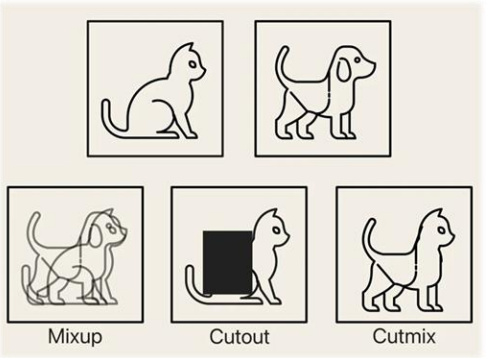
Background image generation

Classification

Key points:

- Detected boxes + 16px margin
- Diverse image sizes (512, 384, 224)
- Diverse model architectures:
  - **CNN**: EfficientNetV2
  - **Transformers**: ViT, NextViT, TinyViT
  - Foundation: DinoBloom (DinoV2)
- Some models with additional background class (24 classes)
- Multi-classes models (from cropped bounding boxes OOF)
- Multi-labels models (from full images)
- Training:
  - Augmentations: (H/V Flip, Rotate90, Rotate/Shift/Scale, Brightness/Contrast, Noise, Cut Out)
  - Augmentations: CutMix and **MixUp**
  - AdamW, CosineAnnealing LR
  - Loss: CrossEntropy and Focal
  - Precision: 16-mixed
  - Batch size: 32, Epochs: 32
  - Fine Tuning





Classification

Model	Image size	Loss	CV	Public LB
Multi-classes ViT large (23 classes)	224x224	CE	0.9061	0.9265
Multi-classes ViT large with background (24 classes)	224x224	CE	0.9037	0.9254
Multi-classes DinoV2/DinoBloom (23 classes) - 5 layers	224x224	CE	0.9133	0.9277
Multi-classes NextViT (23 classes)	384x384	CE	0.9165	0.9279
Multi-classes TinyViT (23 classes)	512x512	CE	0.9181	0.9291
Multi-classes EfficientNetV2m (23 classes)	512x512	CE	0.9182	0.9323
Multi-labels EfficientNetV2m with background (24 classes)	512x512	Focal	0.9279	NA

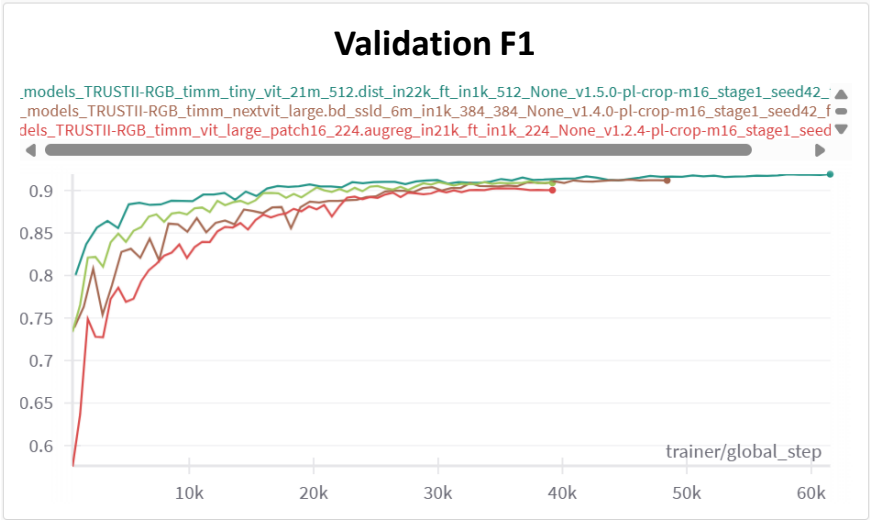
Tools and frameworks:

timm

PyTorch Lightning

Weights & Biases

Albumentations

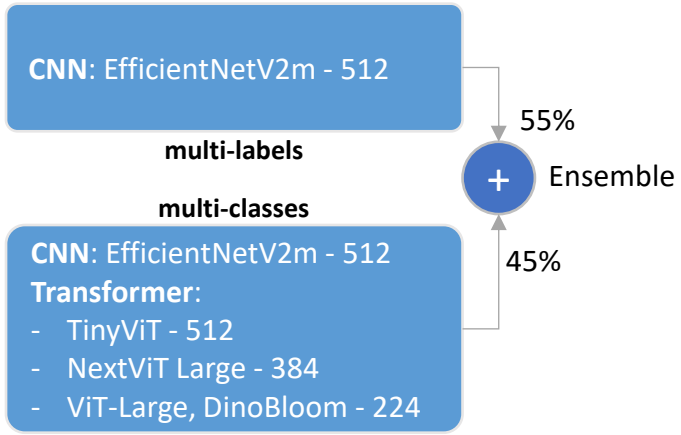


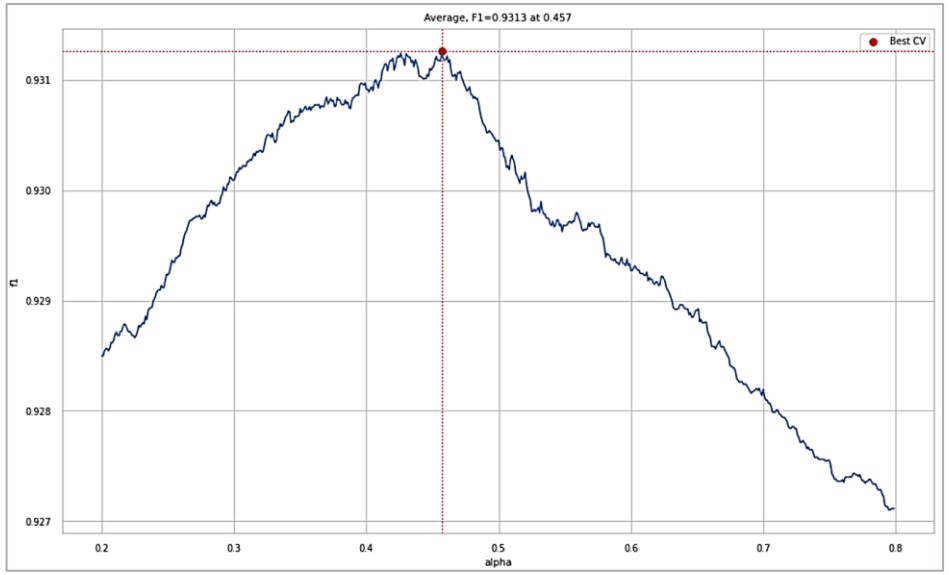
Training examples

Ensemble and Inference

Key points:

- Select models with best CV
- Limit number of models and TTA to be compliant with the 500ms runtime requirement per image
- 85% of images with single WBC instance in test (86% in train).
  - ➢ Ensemble multi-classes + multi-labels models
- 15% of images with multiple WBC instances
  - ➢ Ensemble multi-classes models
- Find best weights to maximize CV score



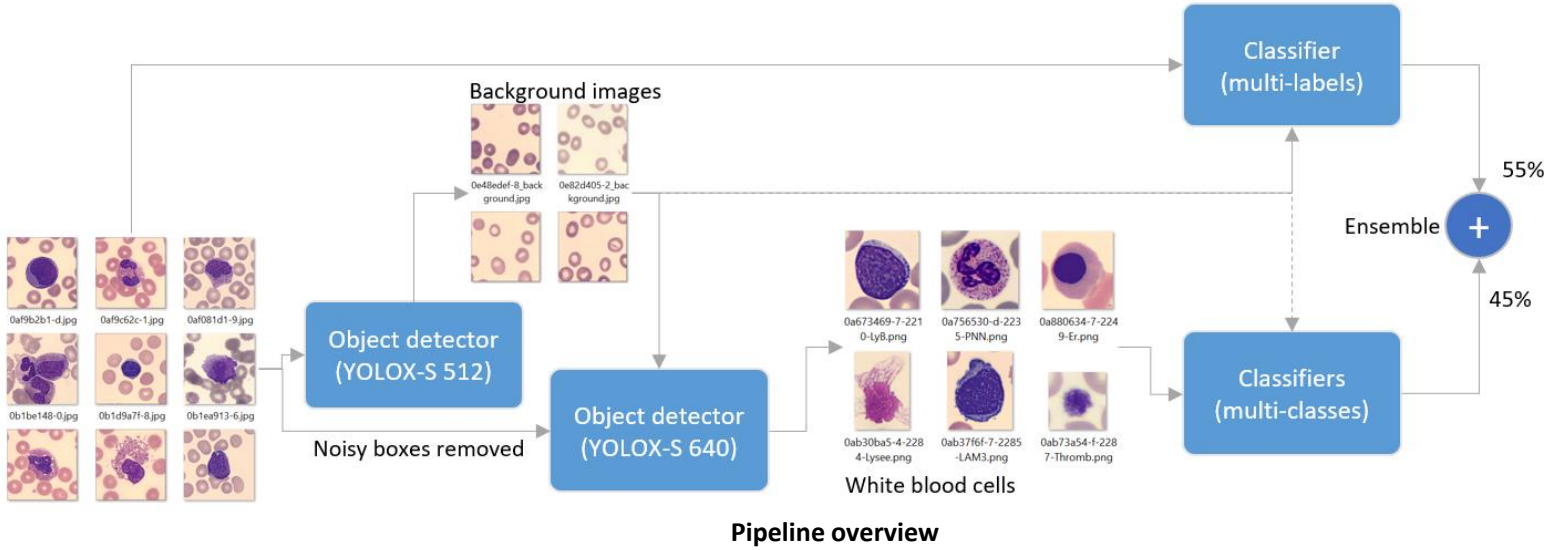


F1 score maximization

- **Best CV:** F1=**0.9313**, Public LB=0.9378, Private LB=**0.9371** WIN
- Best LB: F1=0.9286, Public LB=0.9393, Private LB=0.9348

Takeaway

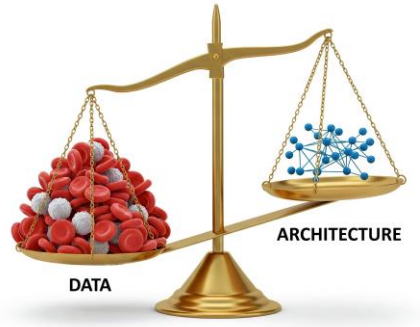
- Focused Expertise with two stages
- Remove hard noisy samples
- Diverse models ensemble
- Trust your CV



Pipeline overview

Next steps

- Data is more important than architecture
- Additional datasets:
  - PBC (Acevedo), BCCD, BCD, LISC, Raabin-WBC
  - AML Hehr/Matek, Elsafty …
- More augmentation (uncropped images or Jigsaw puzzle)
- Train expert models (when we've labels)
- Train foundation/generalized models (Self supervised / DiNO)



L'open data, un catalyseur de l'engagement citoyen au service de la santé

HEALTH DATA HUB

L'open data, un catalyseur de l'engagement citoyen au service de la santé

17h00 - 17h20

Augustin Courtier

Co-fondateur de l'association Latitudes

Augustin Courtier a co-fondé l'association Latitudes qui forme 21 000 personnes par an à la citoyenneté numérique. En particulier, l'Open Data University permet à 2 000 étudiants et étudiantes par an de s'engager sur des causes sociales et environnementales grâce à la data. Augustin annoncera à la Journée Open Science le défi dédié à la santé qui aura lieu en 2025/2026.

Latitudes

L'Open Data University

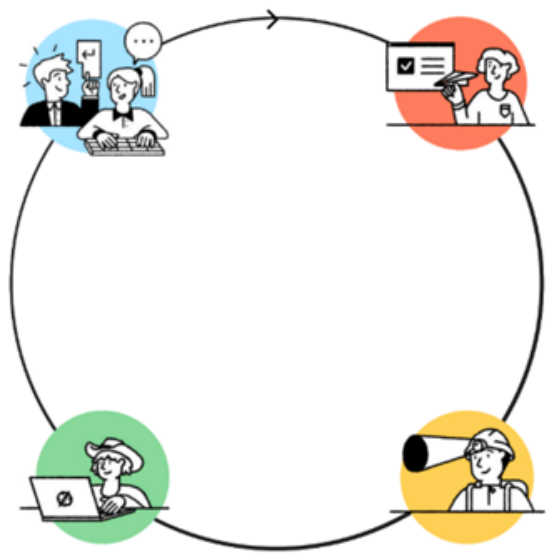
pour l'Enseignement Supérieur

Latitudes est une association de loi 1901 créée en 2017 qui fédère un mouvement de citoyens, citoyennes et organisations du numérique qui veulent agir pour un numérique plus vertueux sur le plan social et environnemental.

Un cercle vertueux au service de l'engagement citoyen et l'open data

1. Les producteurs de données

2. Les établissements



3. Les étudiants et étudiantes

4. Les entreprises et professionnels data

L'ambition de l'Open Data University

Des réutilisations qui contribuent à l'émergence de services et à l'ouverture de nouvelles données

Des établissements qui sèment la graine de l'engagement citoyen dès les études

Des étudiants et étudiantes qui décident d'avoir une carrière orientée vers l'intérêt général

Des professionnels de la data qui savent s'emparer du potentiel de l'open data dans leur organisation

Zoom sur la saison 3

🔢 Quelques chiffres

+2127 étudiantes et étudiants
ont participé (ou participent encore !) à la saison 3

+53 établissements
ont participé (ou participent encore !) à la saison 3

249 groupes
ont travaillé sur les défis proposés cette saison

+104 réutilisations
ont été publiées (ou le seront encore !)

10 mentors mobilisés

🎯 13 défis disponibles

Dont **4 inédits cette saison** (avec des thématiques, technicités, niveaux de difficultés différents).

La répartition des sujets :

🏪 Commerces de centre-ville	44,5%
🗳️ Elections	10%
🚗 Véhicules électriques	8%
🌈 Changement climatique en France	7%
🎭 Offre culturelle	7%
⚡ Diagnostic de Performance Energétique	7%
🏙️ Revitalisation des petites villes	5%
📺 Les Français et Françaises face à l'info	4%
🚴 Infrastructures cyclables	3%
🧝 Diversité et inclusion	3%
🔥 Changement climatique et incendies	0,5%
🪫 Énergie en France	0,5%
🚸 Carte scolaire	0,5%

🔎 Zoom sur le défi Santé & Territoires

Fondation Roche

Niveau : intermédiaire

Ce défi propose de réaliser un **diagnostic territorial de santé publique**, afin de guider les collectivités locales dans la mise en place d'actions de prévention. Il se concentrera sur l'offre de soin, les déterminants sociaux de santé, et l'identification des populations vulnérables

Réalisations possibles : tableau de bord sur l'offre de soin et de prévention d'un territoire, datavisualisations des populations vulnérables, analyse des facteurs influençant les taux de mortalité, etc.

Conclusion et remerciements

HEALTH
DATA HUB

Suivez-nous sur les réseaux sociaux !