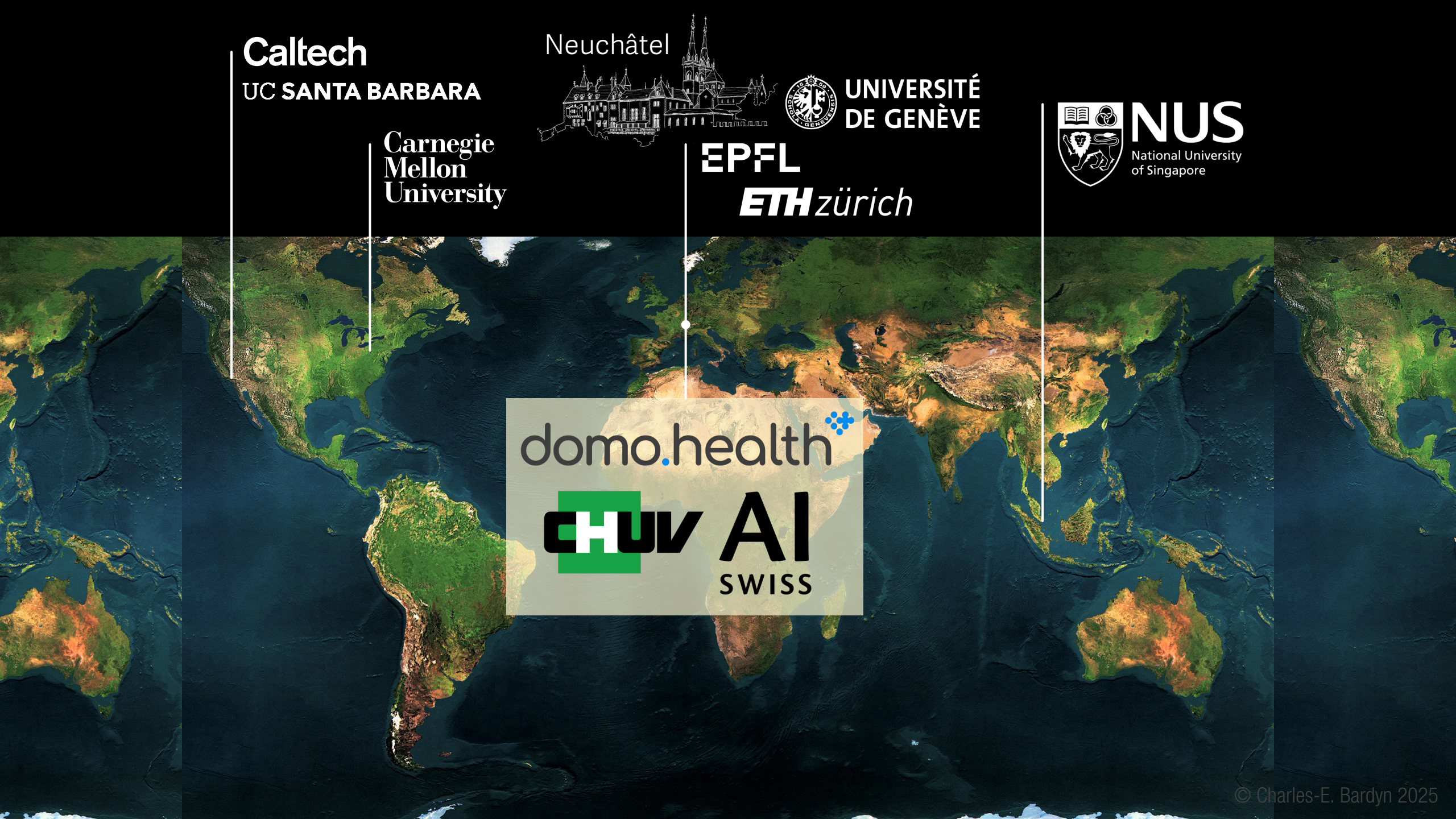
Charles-Edouard Bardyn
EHC | Symposium Digital Santé
16 janvier 2025


Caltech
UC SANTA BARBARA
Carnegie Mellon University
Neuchâtel
UNIVERSITÉ DE GENÈVE
EPFL
ETH zürich
NUS
National University of Singapore
domo.health
CHUV AI
SWISS
© Charles-E. Bardyn 2025

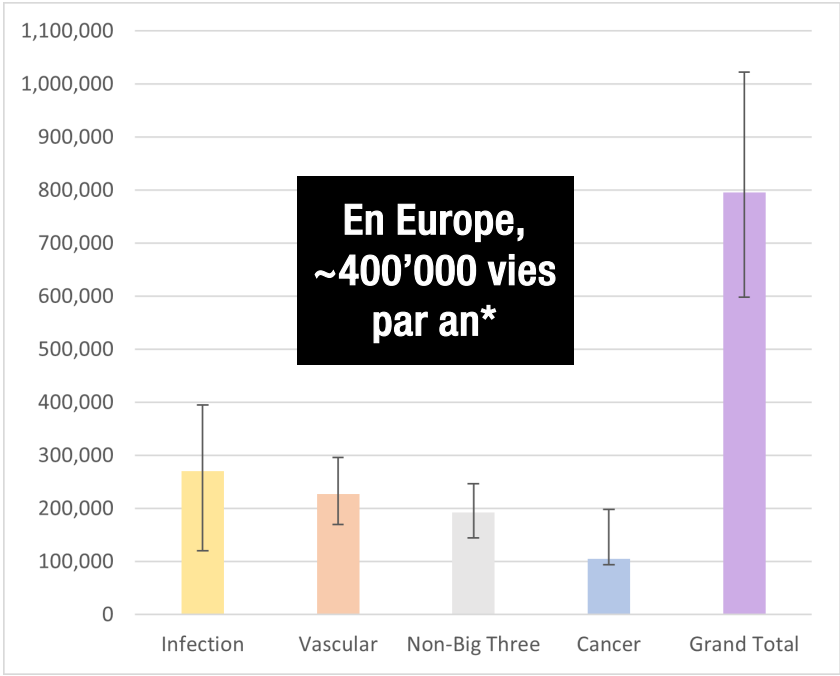
Aujourd'hui, la santé tue

ORIGINAL RESEARCH

Burden of serious harms from diagnostic error in the USA

David E Newman-Toker ,[1,2] Najlla Nassery,[3] Adam C Schaffer,[4,5] Chihwen Winnie Yu-Moe,[5] Gwendolyn D Clemens,[6] Zheyu Wang,[6,7] Yuxin Zhu,[1,6] Ali S. Saber Tehrani,[1] Mehdi Fanai,[1] Ahmed Hassoon,[1,2] Dana Siegal[8,9]

Original research



En Europe, ~400'000 vies par an*

Sans compter les erreurs moins graves et celles qu'on ne détecte pas encore

Figure 1 Annual population incidence of serious misdiagnosis-related harms from vascular events, infections, cancers and all non-'Big Three' others. The estimated grand total annual US incidence for serious harms (combining 'Big Three' harms with other non-'Big Three' harms) is 795 000 (probabilistic plausible range (PPR) 598 000–1 023 000). Whiskers denote PPRs from the Monte Carlo analysis.

*Deloitte & MedTech Europe, 2020

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2021-014130).

For numbered affiliations see end of article.

Correspondence to
Dr David E Newman-Toker, Department of Neurology, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA; toker@jhu.edu

Received 18 August 2021
Accepted 24 June 2023
Published Online First 17 July 2023

were 11.1% and 4.4%, respectively. Extrapolating to all diseases (including non-'Big Three' dangerous disease

events, infections and cancers.

HOW THIS STUDY MIGHT AFFECT

Aujourd'hui, la santé coûte

En Europe, par an* :

~200 milliards d'euros, ~1.8 milliards d'heures de professionels de santé

*Deloitte & MedTech Europe, 2020

Excess Medical Care Spending: The Categories, Magnitude, and Opportunity Costs of Wasteful Spending in the United States

Landmark reports from reputable sources have concluded that the United States wastes hundreds of billions of dollars every year on medical care that does not improve health outcomes. While

creating opportunities for these funds to improve public health.

To this end, we performed a review and crosswalk analysis of the literature to retrieve comprehensive estimates of wasteful medical care spending. We abstracted each source's definitions, categories of waste, and associated dollar amounts. We synthesized and reclassified waste into 6 categories: clinical inefficiencies,

TABLE 2—Range Estimates of Wasteful US Medical Care Spending by Category of Waste and Corresponding Analogous Expenditures, 2019

Category of Waste	2019 Range of Estimates (US$ Billions)	Median Estimate (US$ Billions)	Median Estimate (US$ Per Capita)	Examples of Opportunity Cost of Waste[a]
Clinical inefficiencies	27–378	202	609	Triple the annual National Institutes of Health Research Budget ($117.6 billion) and annual US biopharmaceutical sector research and development ($71.4 billion)[1]
Missed prevention opportunities	29–590	310	934	Total annual direct and indirect costs of diagnosed diabetes in the United States ($245 billion)[13] and annual estimated costs of the American Housing and Economic Mobility Act ($50 billion)
Overuse	66–835	451	1359	Annual estimated costs associated with switching to 100% renewable energy in the United States ($423.9 billion)[14]
Administrative waste	117–461	281	847	Repeal of the estate tax ($64 billion) and a 10% tax reduction to households earning less than $200 000 ($174 billion)
Excessive prices	96–241	169	509	Universal child care ($42 billion), paid family leave ($28 billion), and double the budget of the Supplemental Nutrition Assistance Program ($68 billion)
Fraud and abuse	59–312	185	557	Free annual tuition across all public US colleges and universities ($79 billion) and free annual universal pre-K ($26 billion)

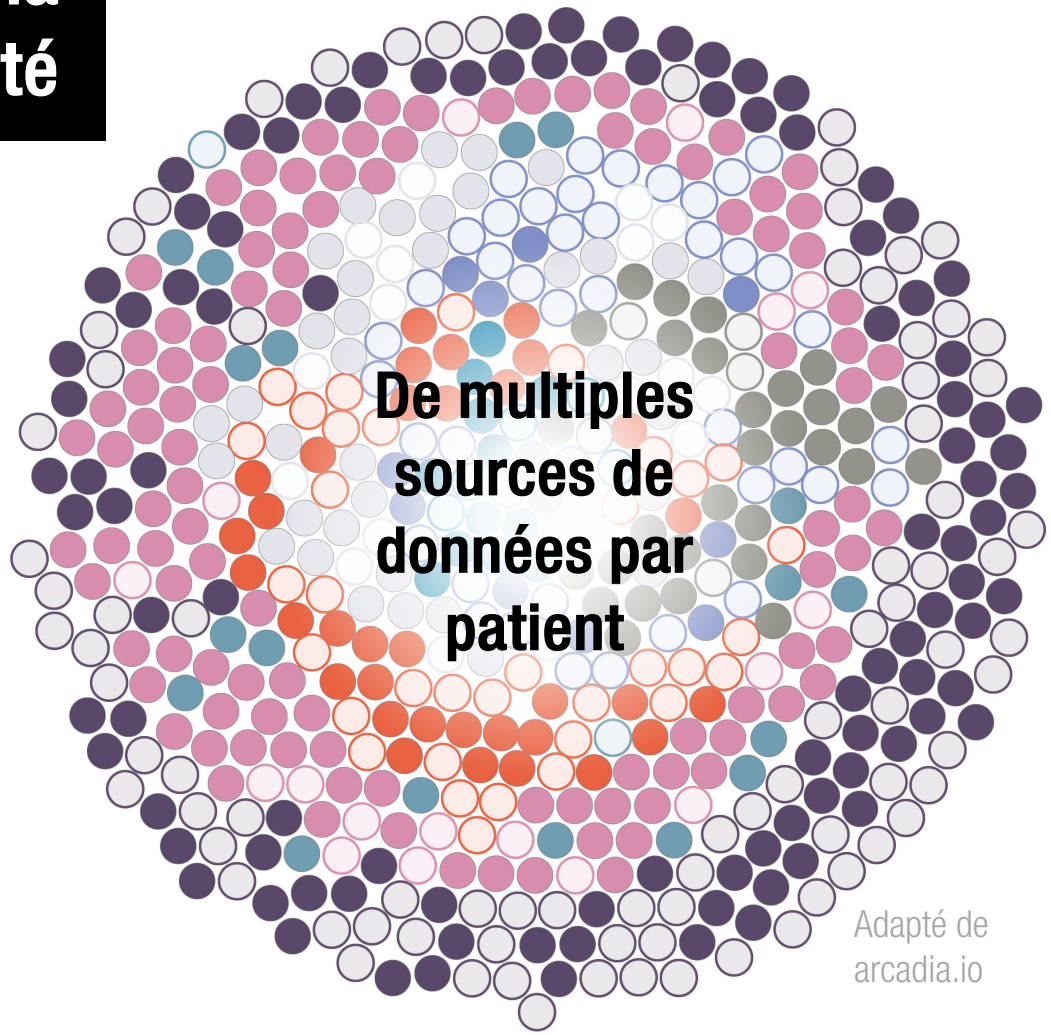
Note. All amounts shown in table are in constant 2019 dollars, adjusted for inflation using the Consumer Price Index medical price index growth rate.
[a]What median estimate amount could cover if addressed and reinvested.

Sciences, Engineering, and Medicine [NASEM]) Roundtable on Value and Science-Driven Health Care convened a 4-part workshop series to examine the major causes of excess

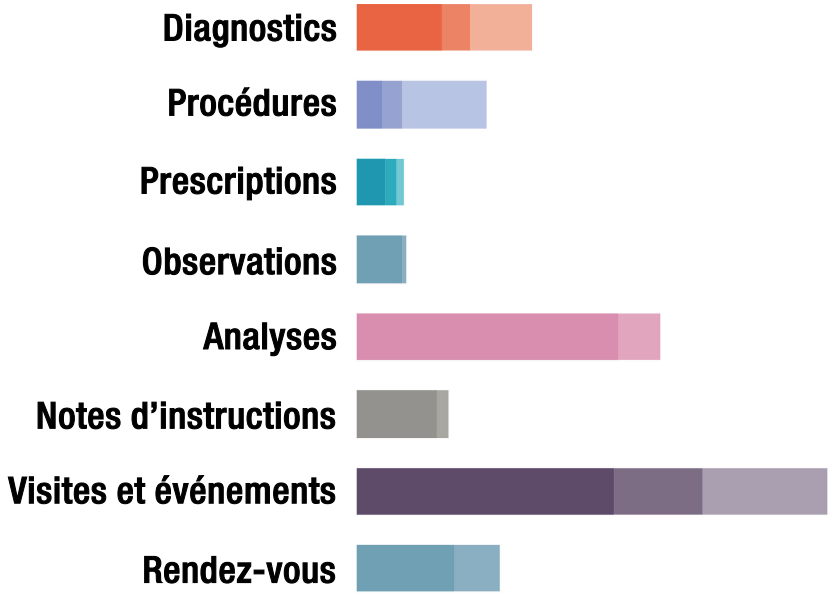
sufficient to completely erase outstanding student loans in the United States in a single year. Indeed, an underexplored aspect of waste is what other high-priority national needs are not adequately funded

We examined the opportunity cost of wasteful spending by identifying topical alternative public health priorities that are roughly equivalent in cost to wasteful medical care

La faute à la complexité



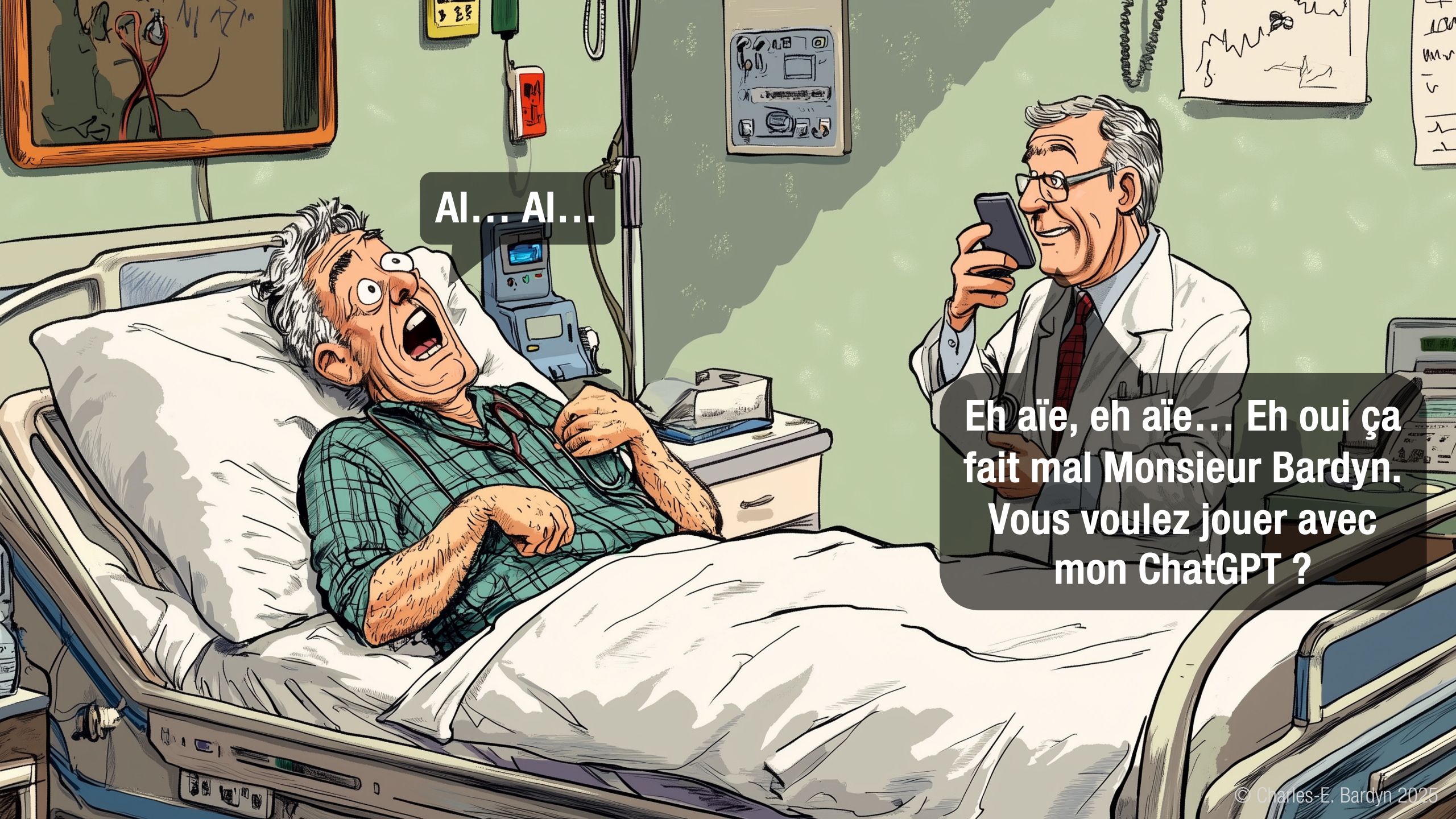
Adapté de arcadia.io



La majorité des professionels se basent sur des résumés…

…et la majorité* des décisions ne se basent sur aucune donnée publiée

* ~80 % selon Darst et al. (2010)



IA au sens large = méthodes de recherche dans l'espace des possibles

Programmes qui distinguent un chien d'un chat
(Kaggle, 2013)

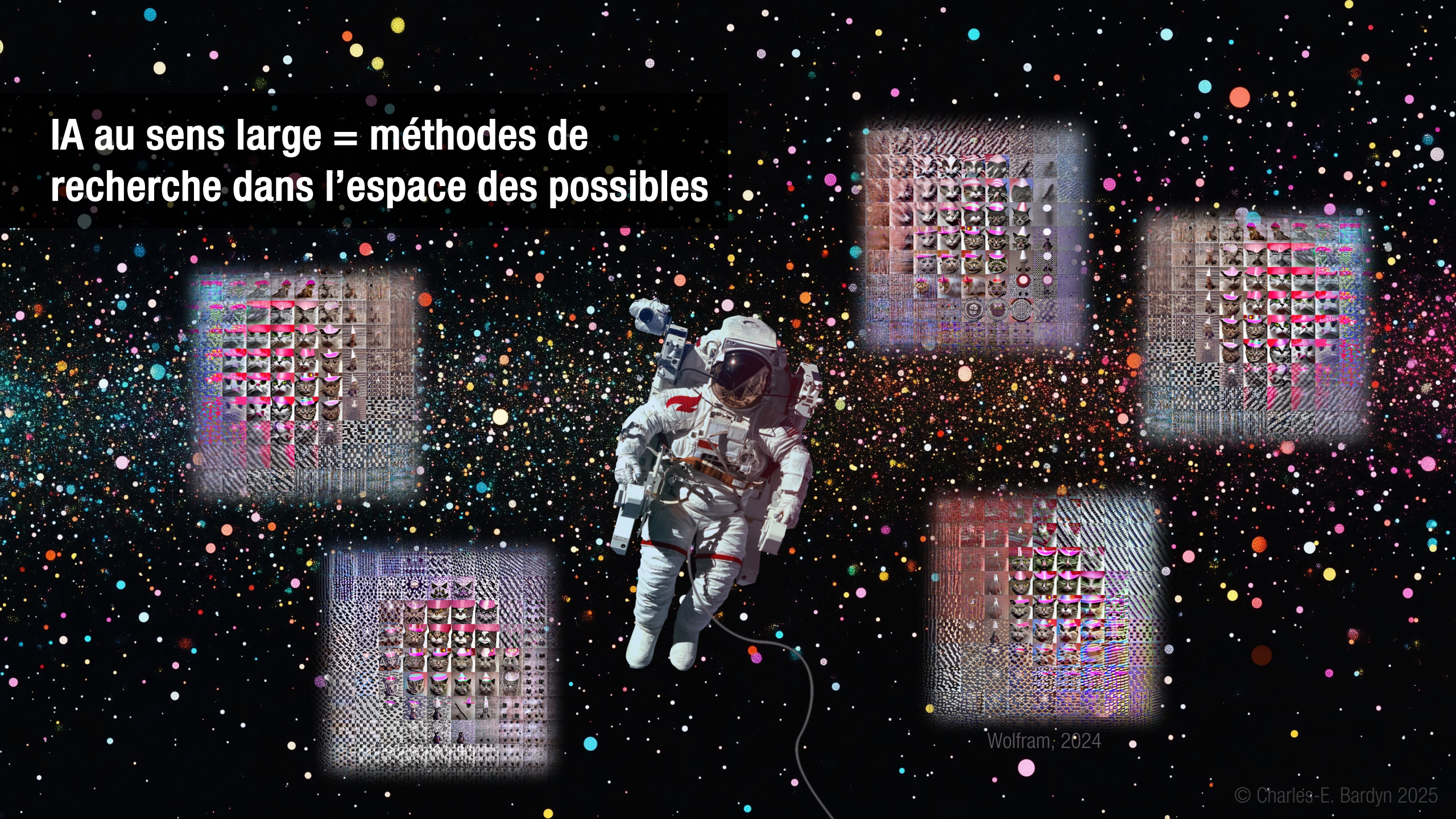
Programmes qui prédisent la structure de molécules
(Prix Nobel, 2024)

Programmes qui prédisent des trajectoires de santé
(CLMBR-T-base, Wornow et al. 2023)

Programmes qui détectent des crises d'épilepsie
(Bardyn et al., 2025)

Programmes qui :
- détectent des exoplanètes,
- prédisent des changements climatiques,
- etc. etc.

IA au sens large = méthodes de recherche dans l'espace des possibles

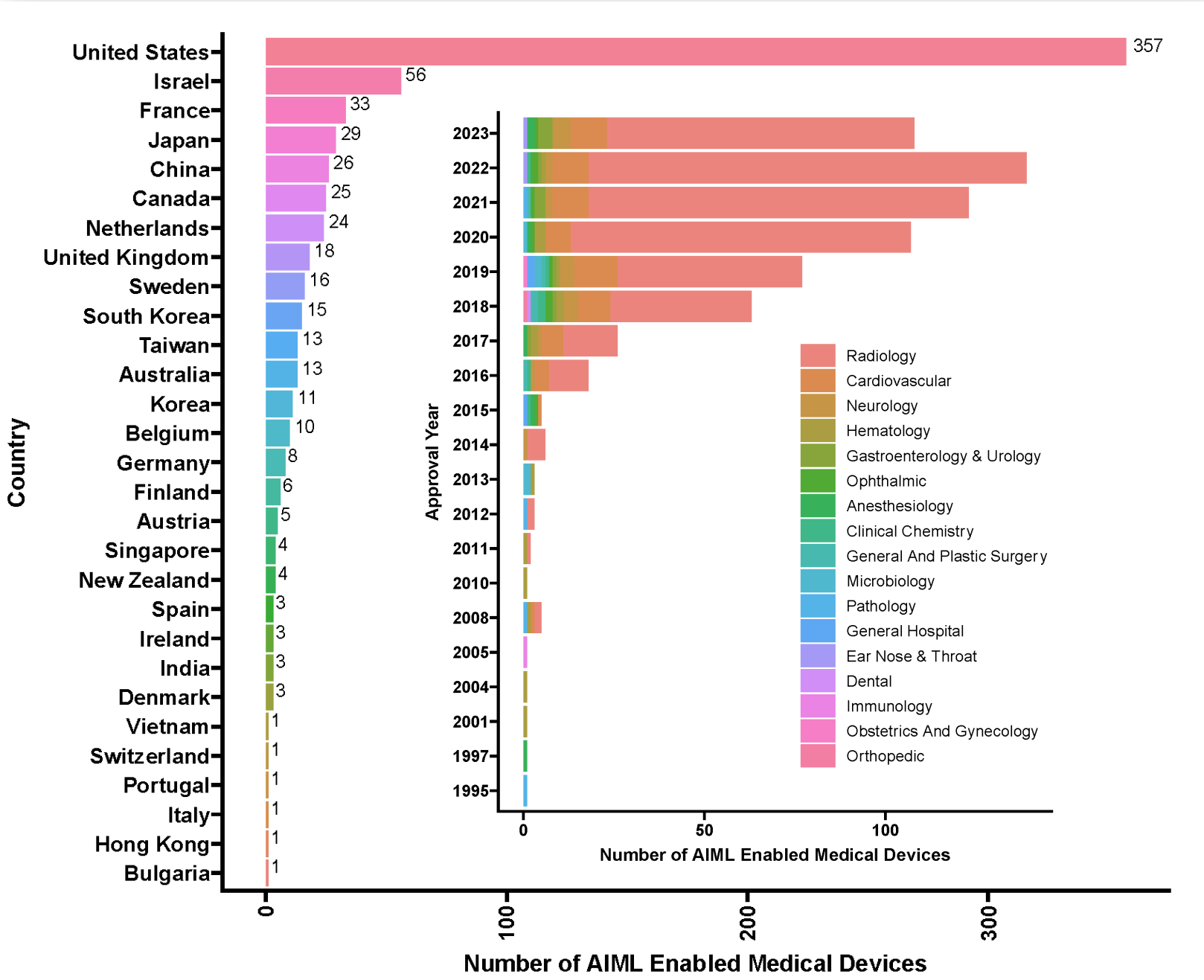


Wolfram, 2024

Les modèles de deep learning classique

> 1’000 algorithmes d’IA approuvés par la FDA

Entraînés sur des tâches spécifiques facilement validables (comparé au reste)



Joshi et al. (2024)

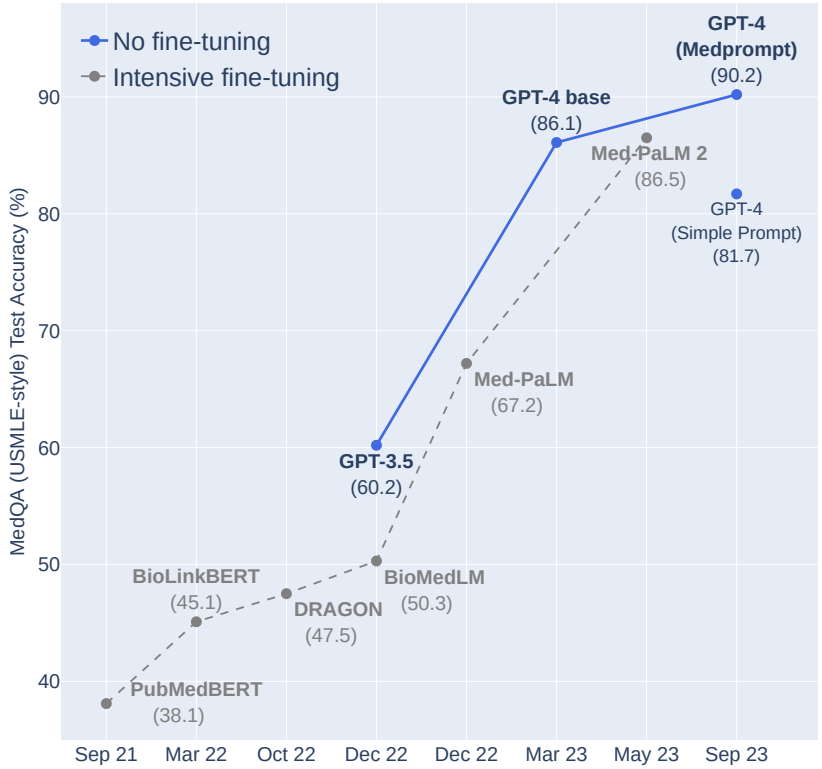
Les modèles “fondation”

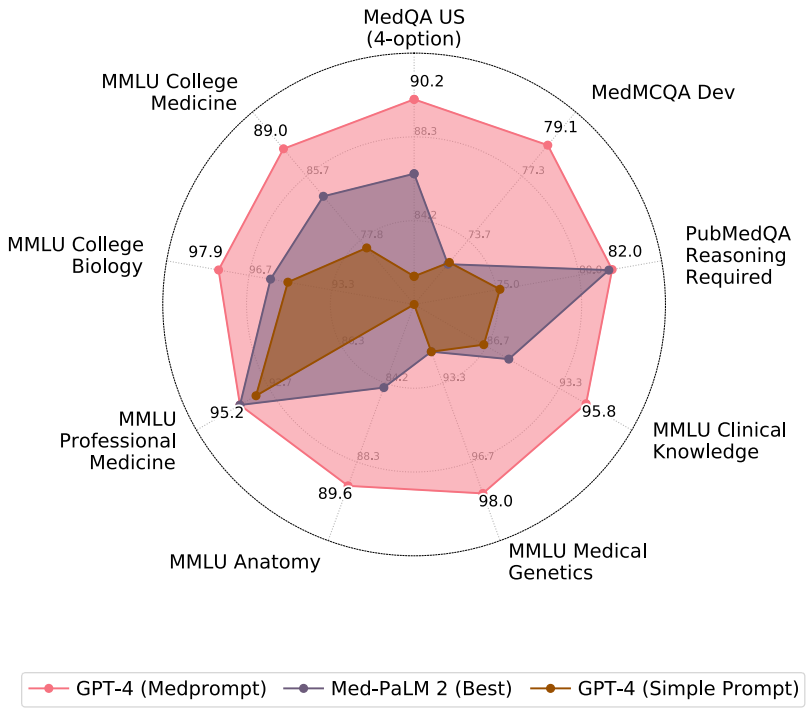
Les modèles fondation

Font émerger des comportements complexes insoupçonnés

Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine

Harsha Nori*‡, Yin Tat Lee*, Sheng Zhang*, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney†, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz‡





models, and surpasses a score of 90% for the first time. Moving beyond medical challenge problems, we show the power of Medprompt to generalize to other domains and provide evidence for the broad applicability of the approach via studies of the strategy on competency exams

Les modèles fondation

S'adaptent mieux à de nouvelles données

npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01166-w

A multi-center study on the adaptability of a shared foundation model for electronic health records

Lin Lawrence Guo[1,7], Jason Fries[2,7], Ethan Steinberg[2], Scott Lanyon Fleming[2], Keith Morse[3], Catherine Aftandilian[4], Jose Posada[5], Nigam Shah[2,8] & Lillian Sung[1,6,8]

Foundation models are transforming artificial intelligence (AI) in healthcare by providing modular components adaptable for various downstream tasks, making AI development more scalable and cost-effective. Foundation models for structured electronic health records (EHR), trained on coded medical records from millions of patients, demonstrated benefits including increased performance with fewer training labels, and improved robustness to distribution shifts. However, questions remain on the feasibility of sharing these models across hospitals and their performance in local tasks. This multi-center study examined the adaptability of a publicly accessible structured EHR foundation model (FM_{SM}), trained on 2.57 M patient records from Stanford Medicine. Experiments used EHR data from The Hospital for Sick Children (SickKids) and Medical Information Mart for Intensive Care (MIMIC-IV). We assessed both adaptability via continued pretraining on local data, and task adaptability compared to baselines of locally training models from scratch, including a local foundation model. Evaluations on 8 clinical prediction tasks showed that adapting the off-the-shelf FM_{SM} matched the performance of gradient boosting machines (GBM) locally trained on all data while providing a 13% improvement in settings with few task-specific training labels. Continued pretraining on local data showed FM_{SM} required fewer than 1% of training examples to match the fully trained GBM's performance, and was 60 to 90% more sample-efficient than training local foundation models from scratch. Our findings demonstrate that adapting EHR foundation models across hospitals provides improved prediction performance at less cost, underscoring the utility of base foundation models as modular components to streamline the development of healthcare AI.

Les modèles “fondation”

Servent de blocs pour d’autres niveaux d’émergence (raisonnements, agents, conscience ?)

Pense étape par étape
(Wei et al., 2022)

Génère de nouvelles réponses à partir des meilleures réponses passées
(Lim et al., 2024)

Pense plusieurs fois étape par étape et choisis ta meilleure réponse
(Wang et al., 2023)

Evalue ta réponse, puis améliore-là
(Wang et al., 2024)

Débats avec d’autres modèles pour améliorer ta réponse
(Du et al., 2023)

Cherche sur le web

Ecris et fais tourner du code pour répondre

Etc. etc.

Les modèles fondation

Reste à vérifier les bénéfices sur le terrain

A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs)

Authors: Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah

0. Key Points

- **Question:** How are healthcare applications of large language models (LLMs) currently evaluated?
- **Findings:** Studies rarely used real patient care data for LLM evaluation. Administrative tasks such as generating provider billing codes and writing prescriptions were understudied. Natural Language Processing (NLP)/Natural Language Understanding (NLU) tasks like summarization, conversational dialogue, and translation were infrequently explored. Accuracy was the predominant dimension of evaluation, while fairness, bias and toxicity assessments were neglected. Evaluations in specialized fields, such as nuclear medicine and medical genetics were rare.
- **Meaning:** Current LLM assessments in healthcare remain shallow and fragmented. To draw concrete insights on their performance, evaluations need to use real patient care data across a broad range of healthcare and NLP/NLU tasks and medical specialties with standardized dimensions of evaluation.

1. Abstract

Importance: Large Language Models (LLMs) can assist in a wide range of healthcare-related activities. Current approaches to evaluating LLMs make it difficult to identify the most impactful LLM application areas.

Objective: To summarize the current evaluation of LLMs in healthcare in terms of 5 components: evaluation data type, healthcare task, Natural Language Processing (NLP)/Natural Language Understanding (NLU) task, dimension of evaluation, and medical specialty.

Aujourd’hui :

95% des études n’utilisent pas de vraies données patient

84% se focalisent sur la tâche de répondre à des questions

1 étude sur 519 seulement analyse l'impact financier / la rentabilité

La biologie est inefficace pour créer des machines capables d'apprendre

Aujourd'hui, quelques lignes de code suffisent

Les modèles fondation et les architectures dérivées (par ex. agentiques) promettent un monde d'innovations

En santé, la plupart des établissements n'ont ni les talents ni l'infrastructure nécessaires

Très peu sont capables de valider des IA, encore moins d'assurer la surveillance humaine nécessaire

Validation, surveillance, infrastructure, maintenance, formation, etc. sont à considérer **en premier** pour éviter les « pilotitis »

Depuis peu, on sait enfin couper le cordon de l'IA

Recherches ouvertes dans l'espace des raisonnements possibles

Recherches ouvertes dans l'espace des actions possibles

Recherches ouvertes guidées par l'intuition des modèles foundation existants

OMNI-EPIC (Faldor et al. 2024)
ADAS (Hu et al. 2024)
PAE (Zhou et al. 2024)
Etc.