THE UNIVERSITY OF MELBOURNE

MDHS AI Education Series

Panel Discussion Webinar:

Leveraging AI for Big Data and Healthcare Analytics

"Discussing how AI facilitates management and analysis of large healthcare datasets, machine learning for data interpretation, and public health

Facilitated by Associate Professor Cathy Quinlan

Panel members (left to right):

Professor Tony Papenfuss, Deputy Director, Walter and Eliza Hall Institute of Medical Research

Dr Chris Longhurst, Chief Clinical & Innovation Officer, UC San Diego Health

Professor Jim Buttery, Professor of Centre for Health Analytics

Visit our Website or scan to watch the recording:

Enabling molecular insights into disease:

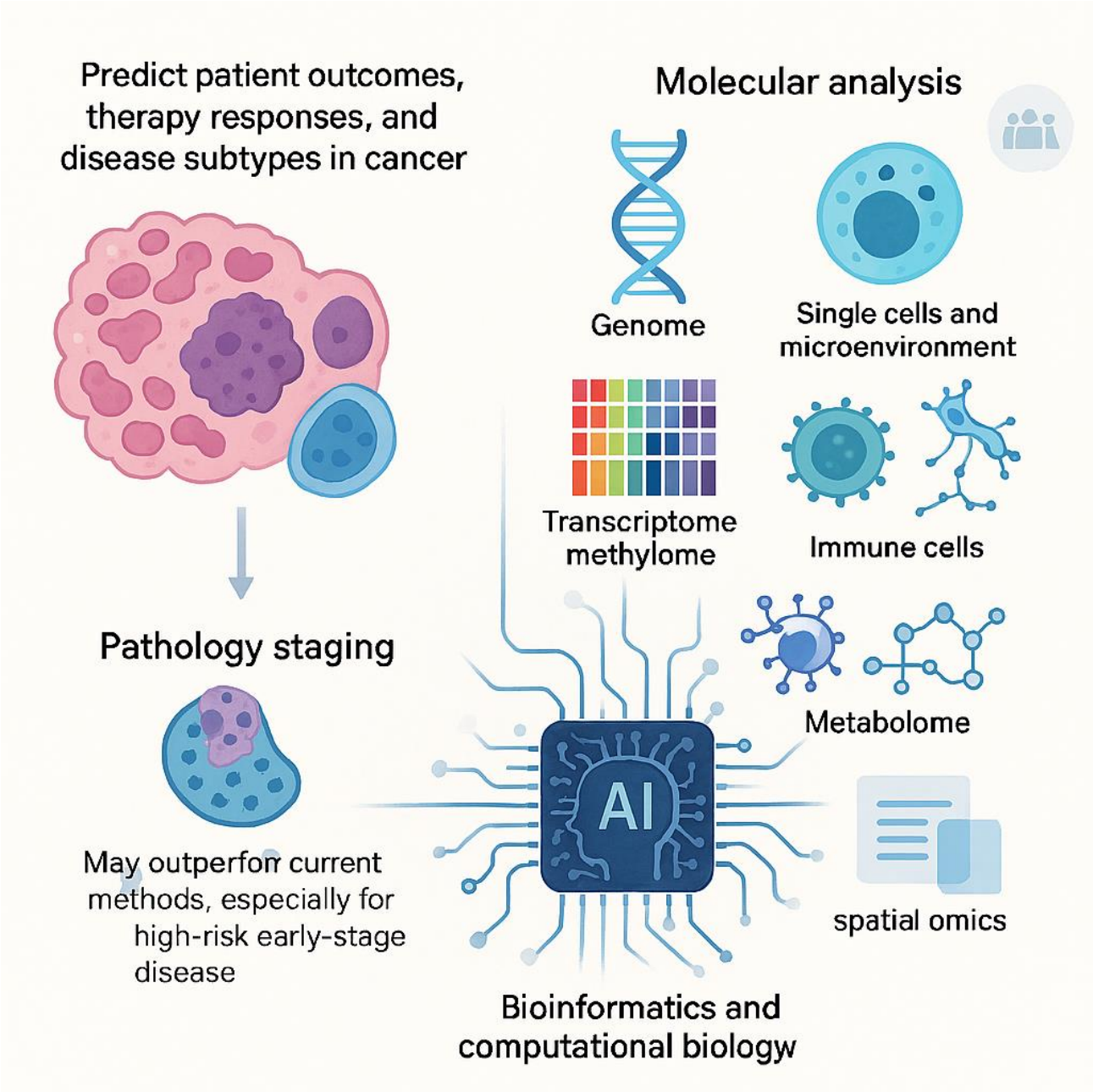
Using AI in bioinformatics and computational biology

Professor Tony Papenfuss
Deputy Director, WEHI

Molecular subtyping of disease

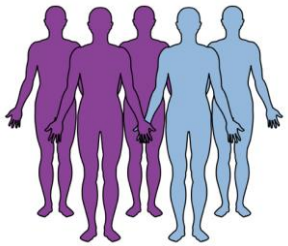
Can we identify molecular features that predict a patient outcome / response to therapy / disease subtype?

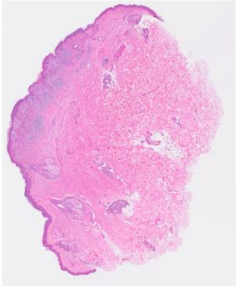
Can molecular features refine pathology staging?

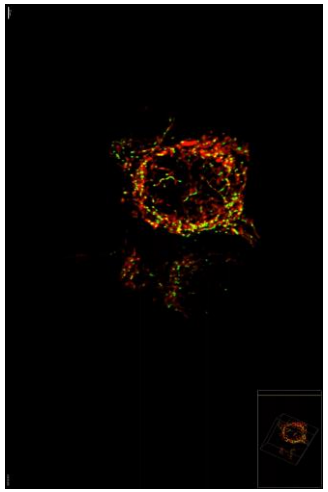


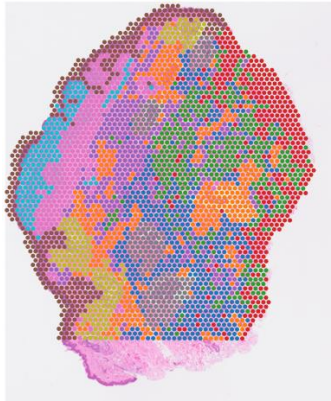
Generated using Copilot

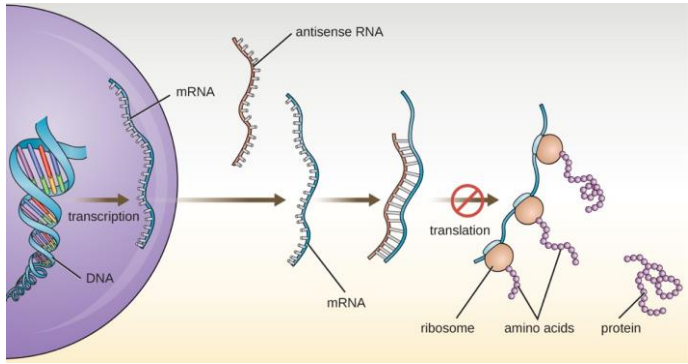
Biomedical data spans population to molecular scale











Population-scale
Medical records
Clinical cohorts

Organisms

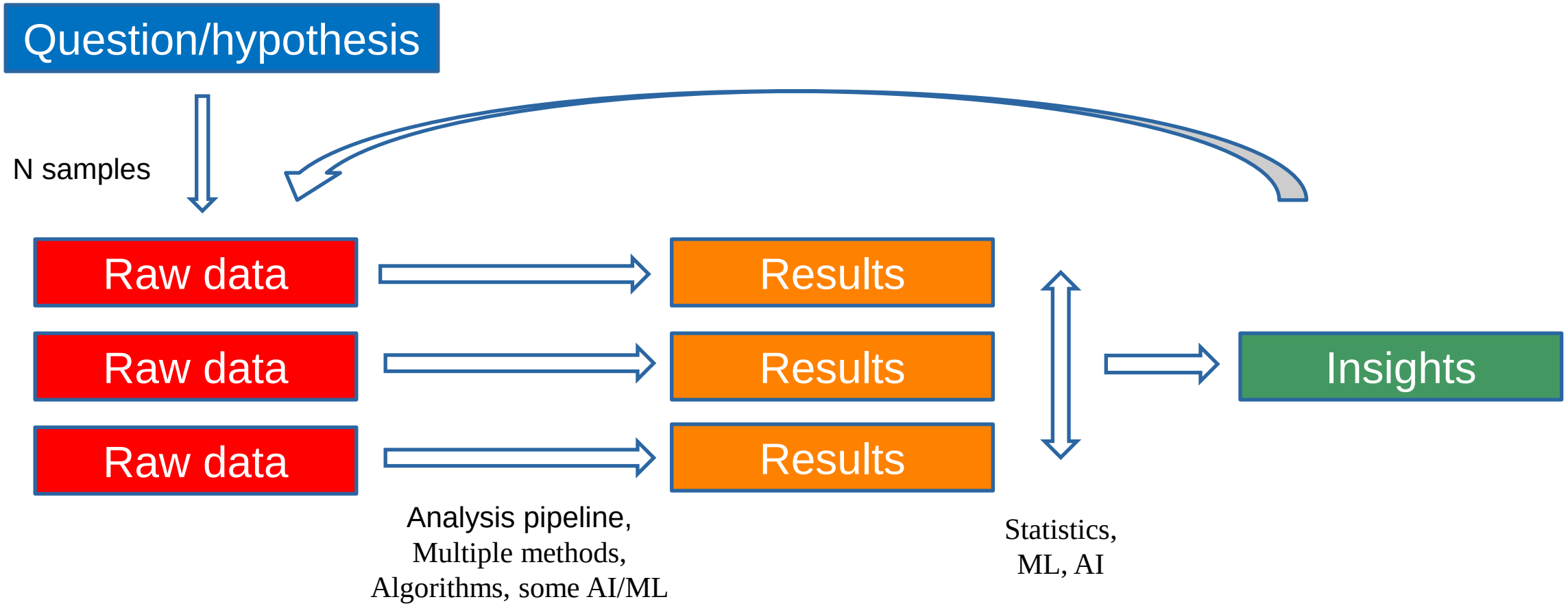
Tissues

Cells

Spatial omics

Molecular data (omics)
Genomes, genes, proteins, drugs, sequence, structure, …

Life history of a research question—today



e.g. Cancer genome → Mutations → Driver genes
Biological processes

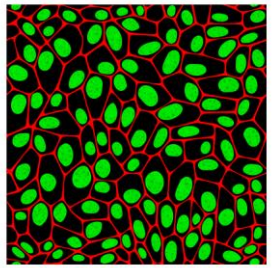
Data size range also varies

Small
N~1-few

Medium-large
N~100s-1000s
+ high quality annotations

Big
N~10K-100K
+ low quality annotation

E.g.



1 image

1K prostate cancer multiomes

10K donors
67M cells

AI opportunities

Apply ML/DL models

Augment data

Apply Foundation models

Train Foundation models

Bring trained models to data

e.g. CNNs

Big public data

>500K participants
>100K genomes

76K WGS, 730K WES

10K donors, 67M cells

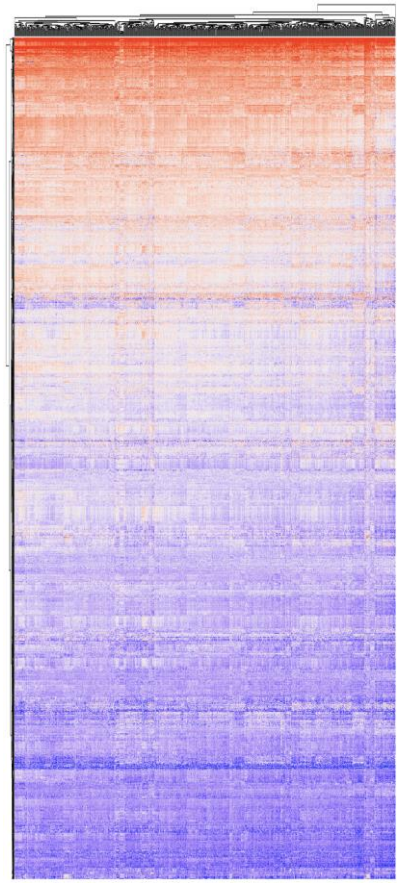
WEHI CFHAC: >139K MS

UK 100K Genomes: 10.5K cancer genomes

Lower quality clinical annotation

Can gene expression in primary prostate cancer predict metastatic potential?

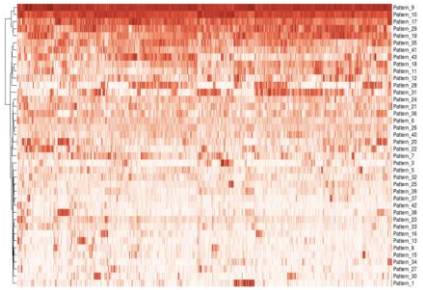
Gene expression



1500 patients
20000 genes

AI

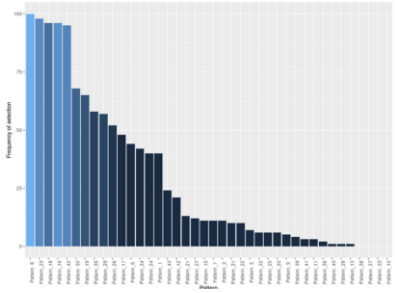
Find recurrent patterns



1000 patients
40 patterns

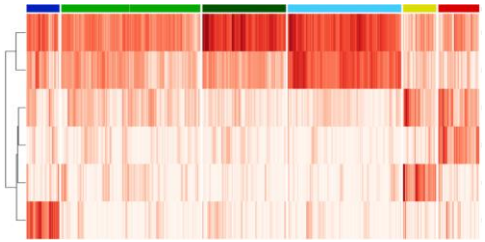
ML, AI or statistics

Pattern importance

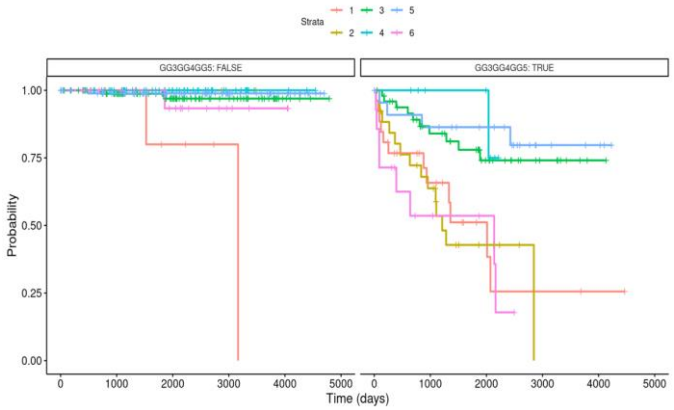


Clustering

Subtypes with different metastatic programs

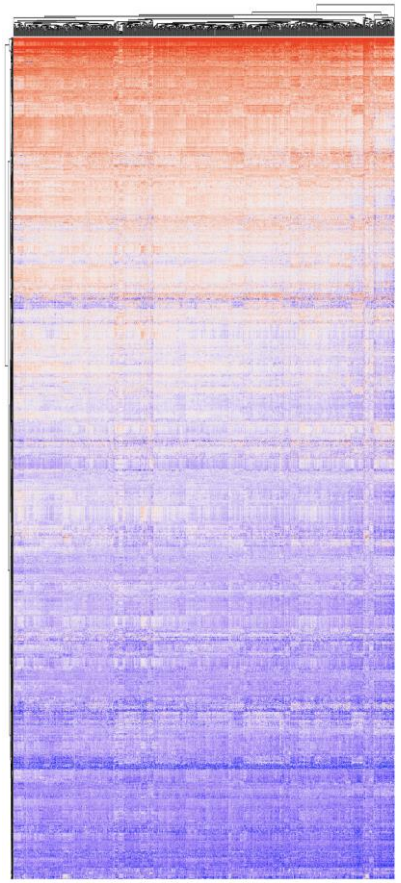


1000 patients
6 subtypes



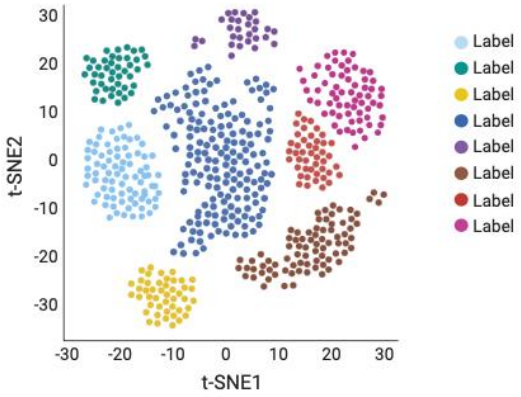
Big public data to prostate cancer & AI

Gene expression



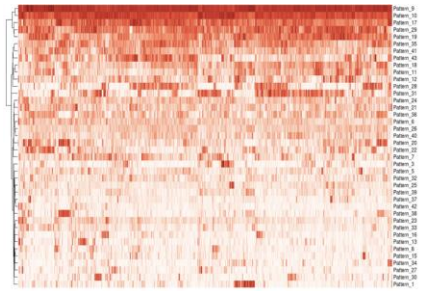
1000 patients
20000 genes

+ big public single cell data



AI

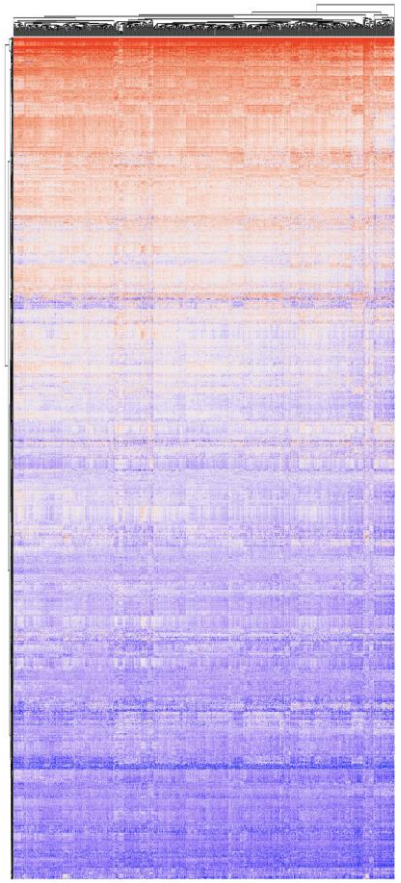
Find recurrent patterns



1000 patients
40 patterns

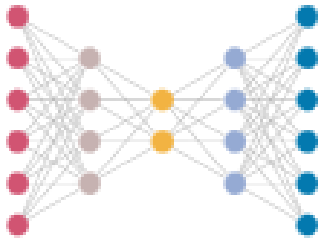
Foundation models & mechanistically informed deep learning

Gene expression



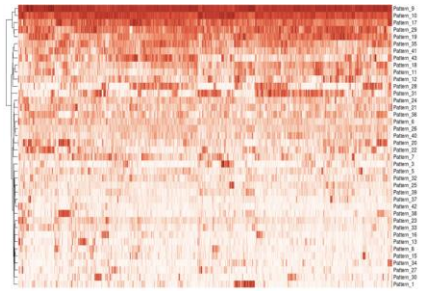
1000 patients
20000 genes

+ foundation single cell foundation model



AI

Find recurrent patterns

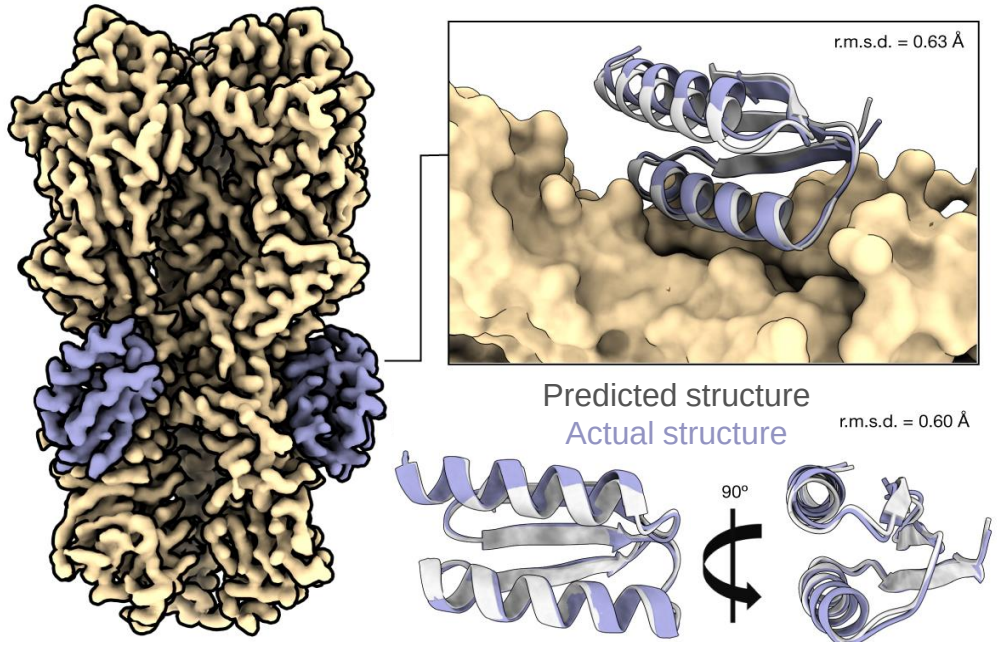


1000 patients
40 patterns

Bring models trained on big public data to your data

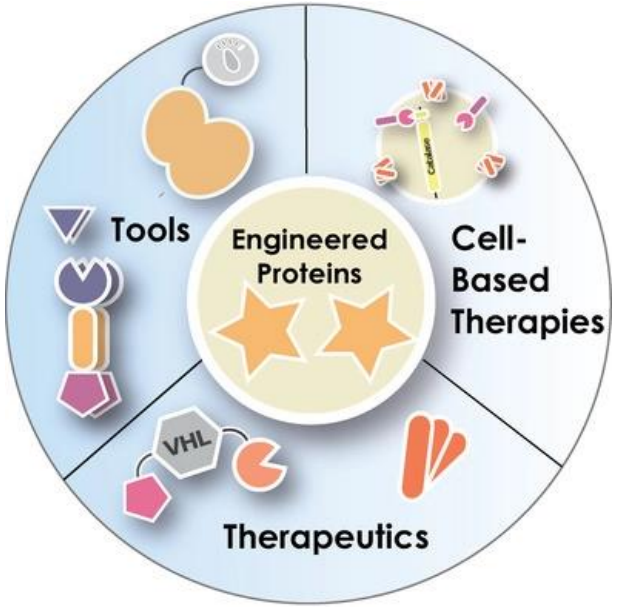
Creating new proteins using AI

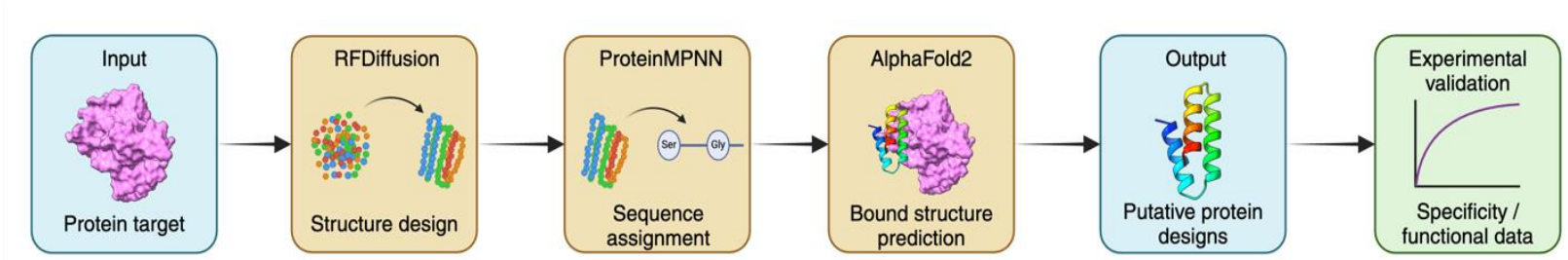
Team: Mr Dylan Silke, Dr Josh Hardy, Prof Isabella Lucet, Dr Marija Dramicanin, Dr Julie Iskander



Binder against influenza spike protein

Applications of de novo binders





Other biomedical applications of AI

Omics/ multiomics

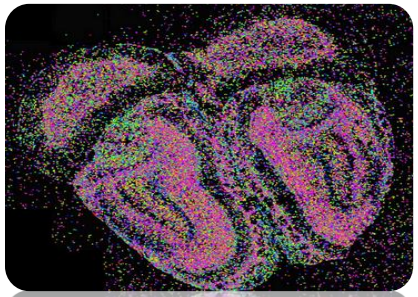
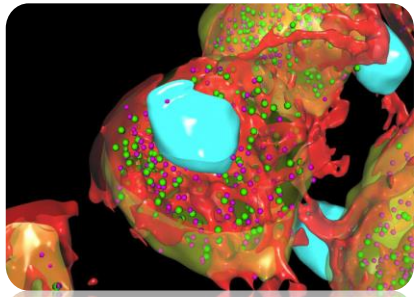


Image analysis



Drug discovery

Text

Clinical

Summary: AI is molecular data analysis and integration

- AI and machine learning is **enabling analyses** that accelerate or enable discoveries not otherwise possible
- **Creativity is needed** to identify how and where to apply AI—not just big data
- **Big public datasets** provide resources for training robust AI models, but quality?
- Foundation models facilitate advances in drug discovery, protein design, and image analysis
- Combining high-quality annotations with big data unlocks new opportunities in precision medicine

WEHI

brighter together

WEHI_research WEHIresearch WEHImovies WEHI_research Walter and Eliza Hall Institute

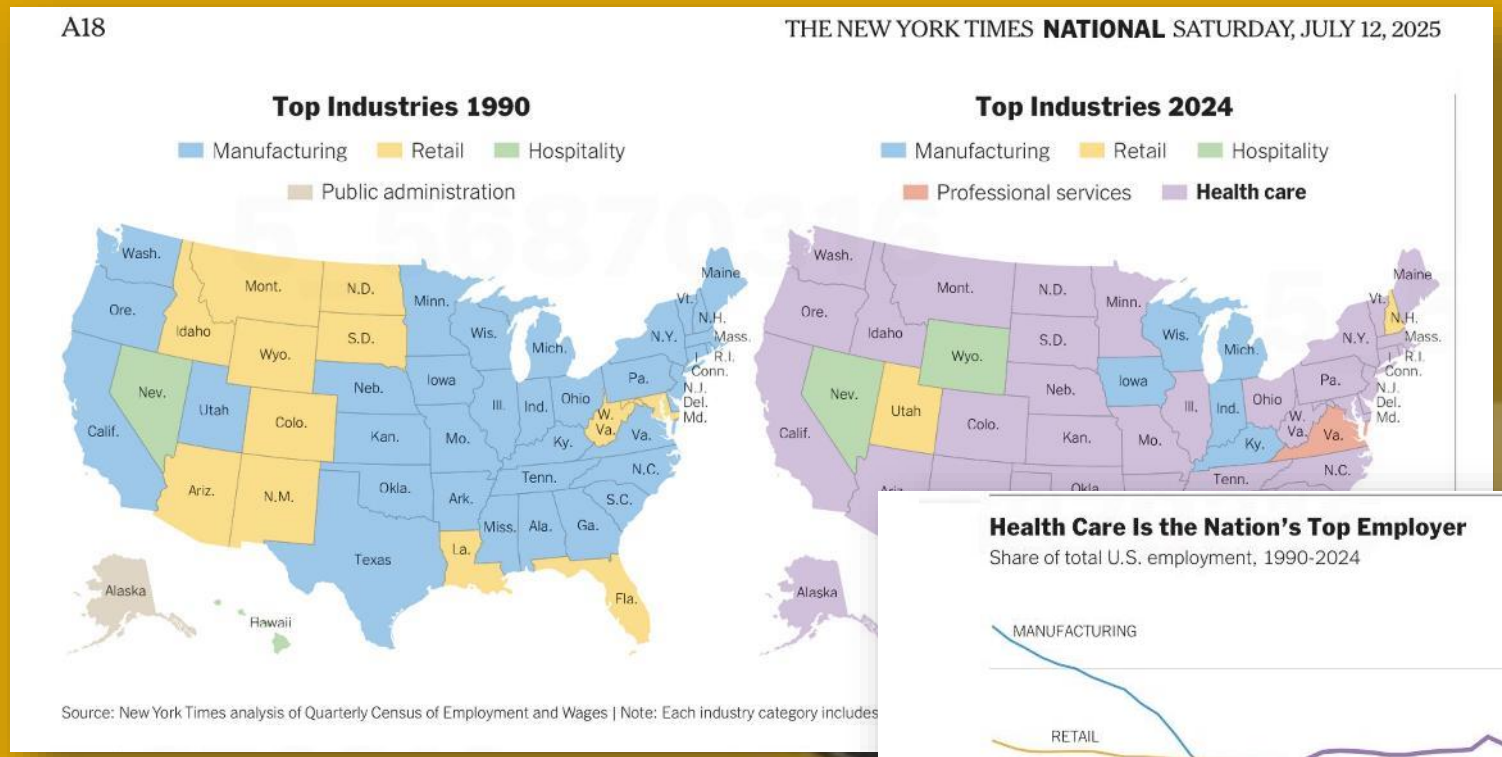
Healthcare Delivery in the AI Era

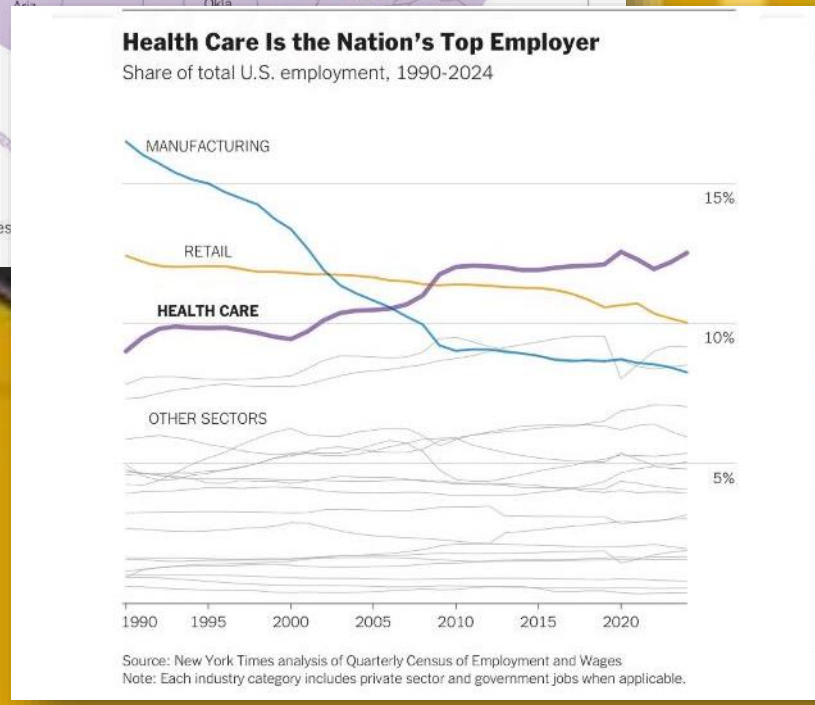
Christopher Longhurst, MD, MS, FACMI

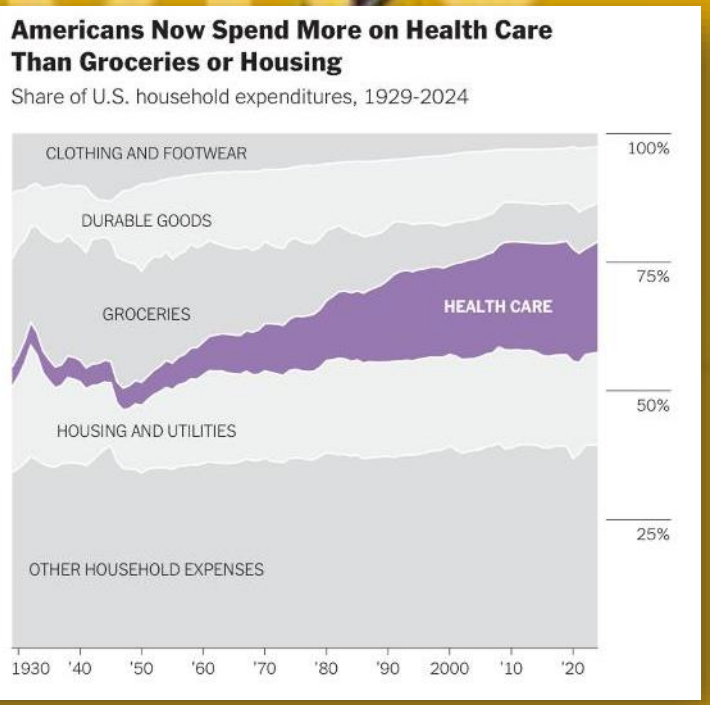
Chief Clinical & Innovation Officer, UC San Diego Health
Clinical Professor and Associate Dean, UCSD School of Medicine
Executive Director, Joan & Irwin Jacobs Center for Health Innovation @ UC San Diego

UC San Diego Health

Healthcare spending consistently outstrips GDP growth







https://jamanetwork.com/journals/jama/fullarticle/2683126

VIEWPOINT

Resolving the Productivity Paradox of Health Information Technology

A Time for Optimism

Robert M. Wachter, MD
Department of Medicine, University of California, San Francisco.

Michael D. Howell, MD, MPH
Google, LLC, Mountain View, California.

During the past decade, the US health care system has gone digital. In 2008, fewer than 1 in 10 US hospitals had an electronic health record (EHR) system; today, fewer than 1 in 10 does not. The increase in use of an EHR system in ambulatory practices has been similarly steep.

Although the evidence that health care digitization has led to improvements in quality and safety is generally positive,[1] these have been context-dependent improvements and relatively small. Moreover, rates of physician self-reported burnout are high, partly because little useful intelligence is delivered back to physicians despite all their time spent performing data entry. In addition, the presence of the computer often separates the physician from the patient, rather than enhancing communication.

Perhaps most surprising, gains in efficiency and productivity have not materialized. This represents a clear example of the productivity paradox of information technology—the observation that productivity does not immediately increase with investment in information technology.[2]

The good news from other industries is that the productivity paradox eventually resolves. In health care, the combination of widespread digitization, advances in machine learning and artificial intelligence, the opening of the health information technology business environment to both established technology companies and start-ups focused on solving specific health care problems, and growing competencies in systems thinking, work redesign, and population health have put health care on the cusp of potentially substantial improvements in quality, safety, patient and clinician experience, and efficiency. In this Viewpoint, we describe several reasons for optimism.

to new public reportin
tems that began hold
clinicians accountab
health care organizati
ity and safety depart
provement techniqu
focused on higher rel

Even though di
these efforts, its cont
have been expected.
care, EHR decision s
delivered as poorly
high rates of false-po
tions that lead to limit
problems. Clinical no
more useful, often in
questionable veracity
data from EHRs for r
has been the capaci
tools and apps—pa
parties—back into EH

Why Might Now Be

The lesson from othe
ity paradox eventua
decades. This was tr
factories at the start o
zation of manufacturi
of this century. Resea
this resolution: impro
reimagining of the wo
similar improvements

The first reason i
digital business envi

JAMA | Special Communication | AI IN MEDICINE

Will Generative Artificial Intelligence Deliver on Its Promise in Health Care?

Robert M. Wachter, MD; Erik Brynjolfsson, PhD

Viewpoint page 17
CME at jamacmelookup.com

IMPORTANCE Since the introduction of ChatGPT in late 2022, generative artificial intelligence (genAI) has elicited enormous enthusiasm and serious concerns.

OBSERVATIONS History has shown that general purpose technologies often fail to deliver their promised benefits for many years ("the productivity paradox of information technology"). Health care has several attributes that make the successful deployment of new technologies even more difficult than in other industries; these have challenged prior efforts to implement AI and electronic health records. However, genAI has unique properties that may shorten the usual lag between implementation and productivity and/or quality gains in health care. Moreover, the health care ecosystem has evolved to make it more receptive to genAI, and many health care organizations are poised to implement the complementary innovations in culture, leadership, workforce, and workflow often needed for digital innovations to flourish.

CONCLUSIONS AND RELEVANCE The ability of genAI to rapidly improve and the capacity of organizations to implement complementary innovations that allow IT tools to reach their potential are more advanced than in the past; thus, genAI is capable of delivering meaningful improvements in health care more rapidly than was the case with previous technologies.

JAMA. 2024;331(1):65-69. doi:10.1001/jama.2023.25054
Published online November 30, 2023.

Author Affiliations: Department of Medicine, University of California, San Francisco (Wachter); Digital Economy Lab and Institute for Human-Centered AI, Stanford University, Palo Alto, California (Brynjolfsson).

Corresponding Author: Robert M. Wachter, MD, Department of Medicine, University of California, San Francisco, 505 Parnassus Ave, Room M994, Box 0120, San Francisco, CA 94143-0120 (robert.wachter@ucsf.edu).

Since they became publicly available in late 2022, generative artificial intelligence (genAI) tools such as ChatGPT have elicited enormous enthusiasm, as well as concern, in all sectors of the economy. But the potential impact of genAI in health care seems particularly noteworthy. In a field in which an estimated 30% of the $4.3 trillion spent each year in the US adds little to no value,[1] in which many tens of thousands of people die yearly from preventable mistakes,[2,3] and in which access to care is fragmented and inequities are commonplace, it is natural to be enthusiastic about the

The Productivity Paradox of Information Technology

Research on the productivity paradox has demonstrated that there are 2 fundamental reasons that technologies don't rapidly create their promised value.[6] The first is that early versions of many technologies are flawed; the tools that ultimately succeed are ones that have improved with successive iterations.

https://jamanetwork.com/journals/jama/fullarticle/2812615

November 2022

What would you say to a 12 yo girl with a 1-week history of worsening swelling in her legs and blood in her urine yesterday?

If I were Meg's doctor, I would take a compassionate and empathetic approach to address her concerns and provide comfort and support. Here's an example of what I might say:

"Hi Meg, I understand that you might be feeling worried or scared about the symptoms you've been experiencing lately. I want to reassure you that we're here to help you get better. Based on your symptoms and the tests we've done, it looks like you have a condition called post-streptococcal glomerulonephritis, which can happen after an infection like the sore throat you had recently. The good news is that most people with this condition recover fully, and we're going to closely monitor your progress and provide the appropriate treatment to help you feel better.

If you have any questions or concerns at any time, please don't hesitate to ask me or any of the other healthcare professionals involved in your care. We're here to support you and make sure you understand what's happening with your body. And don't worry about missing school - your health is the top priority right now, and we can work with your school to ensure you catch up on your studies when you're feeling better.

Remember, it's okay to feel scared or worried, but know that we're here to help and support you every step of the way. If you need anything or have any questions, just let us know."

JAMA Internal Medicine, April 2023

JAMA Internal Medicine | Original Investi

ChatGPT found to give better medical advice than real doctors in blind study: 'This will be a game changer'

While AI chatbot's answers scored higher for quality and empathy, some experts express caution and worry

By Melissa Rudy · Fox News

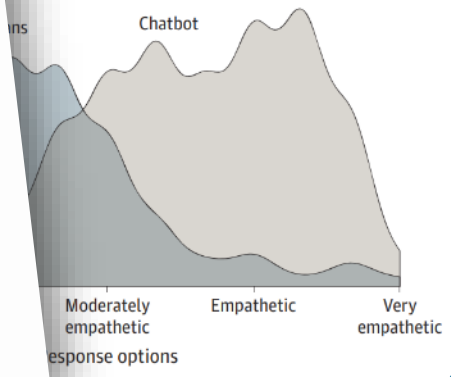
Published May 2, 2023 2:00am EDT

AI has better 'bedside manner' than some doctors, study finds

ChatGPT rated higher in quality and empathy of written advice, raising possibility of medical assistance role

A panel of healthcare professionals preferred ChatGPT's responses to medical questions over those of a doctor 79% of the time. Photograph: Ariel Skelley/Getty Images

ChatGPT appears to have a better 'bedside manner' than some doctors - at least when their written advice is rated for quality and empathy, a study has shown.



UC San Diego Health

Addressing a Problem – April 2023

Special Reports > Exclusives

Are Doctors Drowning in Inbox Overload?

— The slow simmer of non-stop messaging boils over during the pandemic

by Sophie Putka, Enterprise & Investigative Writer, MedPage Today September 22, 2021

EHR/HEALTH IT, INFORMATION, PHYSICIANS

MyChart Messages the Wild West of Patient Communication

OCTOBER 2, 2021 BY BRYAN VARTABEDIAN · READING TIME: 6 MINUTES

Share Share Tweet Share

What physicians, patients think of charging for MyChart messages

Laura Dyrda (Twitter) - Friday, November 18th, 2022

Save Post Tweet Share Listen Text Size Print Email

Cleveland Clinic's plan to charge for MyChart messages requiring clinician time and expertise has been met with mixed reviews from physicians and patients.

Cleveland Clinic began charging for messages requiring clinicians to make clinical assessments, medical decisions or medical history reviews that would take more than five minutes. The charges are often covered by insurers, but patients face up to $50 charges without coverage.

Medicare patients without secondary insurance may see $3 to $8 fees, according to Cleveland Clinic. The new fees went into effect Nov. 17. Some physicians and patients support the charges while others argue it will limit access to care.

"I don't think that it's fair to be charged just for communication, especially with all the other things

HEALTH TECH

HIMSS23: Epic taps Microsoft to integrate generative AI into EHRs with Stanford, UC San Diego as early adopters

By Heather Landi · Apr 17, 2023 11:00am

Microsoft | Epic | generative AI | electronic health records (EHRs)

AI Principles at UC San Diego Health

Artificial Intelligence

Our Statement

1. We believe that AI can enhance human health and well-being, and we are committed to developing and deploying AI solutions that are ethical, responsible, and beneficial for patients and society.
2. We respect the dignity, autonomy, and privacy of each patient, and we design and evaluate our AI systems with their needs, preferences, and feedback in mind.
3. We uphold the highest standards of scientific rigor, transparency, and accountability in our AI research and practice, and we adhere to the ethical codes and regulations of our profession and institution.
4. We foster a culture of collaboration, excellence, and innovation among our AI researchers, practitioners, partners, and stakeholders, and we seek to share our knowledge and expertise with the broader community.
5. We embrace the diversity of our patients, staff, and collaborators, and we strive to create an inclusive and supportive environment that values different perspectives, backgrounds, and experiences.

Effectiveness, Responsibility, and Accountability

AI products should be successful for their intended use, beneficial, valid, accurate, and reliable and those who develop and apply tools with responsibility and accountability

Ethics and Inclusiveness

AI products should strive to achieve health equity and fairness by design and operation

Human Factors

AI product design, development, and implementation should involve and prioritize the needs of the diverse population it serves

Promoting Human Well-Being, Safety, Privacy, and Common Good

AI products should protect human well-being, privacy, sustainability, and the environment

Transparency

AI products should be explainable, trustworthy, intelligible, and accountable

Protecting Human Autonomy and Dignity

AI products should empower individuals it serves

health innovation

Generated Draft Reply — Pioneering Feature

Hello Barbara,

Your CBC with Auto Diff result from 3/7/2023 is within the normal range and does not indicate any cause for concern. Your provider will contact you if any further action is needed. Your results are available for viewing in MyChart, under the "Test Results" tab.

If you have any further questions or concerns, please do not hesitate to contact us.

Best regards,

Learn more | Start with Draft | Start Blank Reply

Did you find the draft reply helpful?

This was helpful | This was not helpful

Dear [redacted]

Thank you for your message and for taking an active role in your health. It's great to hear that you're considering the Hepatitis A vaccine. Twinrix is indeed a combination vaccine that covers both Hepatitis A and B. However, since you've already had the Hepatitis B vaccine, you may not need the combination.

I see you have an upcoming appointment and I recommend reviewing at that appointment. It is a 2 shot series so you can do the 1st shot at the appt if you would like.

Take care,
Marlene May Millen, MD

Part of this message was generated automatically and was reviewed and edited by Marlene May Millen, MD.

UC San Diego Health

JAMA, April 2024

"Though the replies sound very robotic still, they're extremely helpful for generating the baseline response to what you'd want to say to a patient."

"I can't wait for them to get even better, to the point where they can mimic each physician's language/tone."

"I think AI responses have its place. [I] worry about inaccuracies that I may miss due to busy workload. I have been very impressed with [a] few of the responses."

"Great initiative which requires supervision. Hopefully there would be time when minimal supervision would be needed."

"Helpful in drafting responses, provides more empathy into a response without me taking time to type it all out."

"Not perfect but decreases time I spend on it and has a kind tone."

"While not perfect, I think there have been a good number of cases where I use the draft as a starting point. I expect the AI responses to get better over time."

JAMA Network | Open

Original Investigation | Health Informatics

AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication

Ming Tai-Seale, PhD, MPH; Sally L. Baxter, MD, MSc; Florin Vaida, PhD; Amanda Walker, MS; Amy M. Sitapati, MD; Chad Osborne, MD; Joseph Diaz, MD; Nimit Desai, BS; Sophie Webb, MS; Gregory Polston, MD; Teresa Helsten, MD; Erin Gross, MD; Jessica Thackaberry, MD; Ammar Mandvi, MD; Dustin Lillie, MD; Steve Li, MD; Geneen Gin, DO; Suraj Achar, MD; Heather Hofflich, DO; Christopher Sharp, MD; Marlene Millen, MD; Christopher A. Longhurst, MD, MS

Abstract

IMPORTANCE Timely tests are warranted to assess the association between generative artificial intelligence (GenAI) use and physicians' work efforts.

OBJECTIVE To investigate the association between GenAI-drafted replies for patient messages and physician time spent on answering messages and the length of replies.

Key Points

Question Would access to generative artificial intelligence-drafted replies correlate with decreased physician time on reading and replying to patient messages, alongside an increase in reply length?

JAMA Network Open, October 2024

JAMA Network | Open

Research Letter | Health Informatics

Utility of Artificial Intelligence–Generative Draft Replies to Patient Messages

Eden English, MD; Janelle Laughlin, MD; Jeffrey Sippel, MD, MPH; Matthew DeCamp, MD, PhD; Chen-Tan Lin, MD

Introduction

Messages from patients to care te
burnout.[2] Health systems are dep
drafting replies to patient messag

In contemporary team-base
iterative changes in an LLM prom

Table 1. Comparison of Survey Respondents (Registered Nurses, Medical Assistants, and Clinicians) Who Replied Agree or Strongly Agree

Question	Respondents, No. (%) (N = 69)			
	Registered nurses (n = 12)	Medical assistants (n = 14)	Clinicians (n = 43)[a]	*P* value (χ^2)
"I can reply more quickly."	11 (92)	7 (50)	20 (46)	.03 (10.9)[b]
"It's easier to express written empathy."	11 (92)	6 (43)	18 (42)	.01 (13.1)[b]
"The message tone is appropriate."	11 (92)	11 (79)	28 (65)	.21 (5.9)[b]
"PAM Chat is fun to use."	11 (92)	8 (57)	22 (51)	.08 (8.2)[b]
"Minimal risk of giving incorrect information."	8 (67)	2 (14)	12 (28)	.04 (10.1)[b]
"Significant edits are rarely needed."	NA	NA	5 (12)	<.001 (NA)[c]
"I can address more messages without forwarding."	9 (75)	3 (21)	NA	.002 (12.8)[b]
"This helps me to stay within my scope of practice."	6 (50)	4 (29)	NA	.01 (9.0)[b]
Net promoter score (likely to recommend to others)	58	−29	−43	.004 (NA)[d]

Abbreviations: NA, not applicable; PAM Chat, patient advice message chatbot.

[a] Clinicians included physicians and advance practice clinicians.

[b] χ^2 Likelihood ratio.

[c] Two-sided t test for single-group measurement.

[d] Analysis of variance.

UC San Diego Health

The New York Times

That Message From Your Doctor? It May Have Been Drafted by A.I.

Overwhelmed by queries, physicians are turning to artificial intelligence to correspond with patients. Many have no clue that the replies are software-generated.

NEJM AI

DOI: 10.1056/AIp2401167

PERSPECTIVE

A Call for Disclosure When Using AI for Patient Communications

Marlene Millen, M.D.,[1] Ming Tai-Seale, Ph.D.,[1,2] and Christopher A. Longhurst, M.D.[1]

Received: December 1, 2024; Revised: January 19, 2025; Accepted: March 17, 2025; Published: May 9, 2025

Abstract

In the spring of 2023, the UC San Diego Health system began generating draft replies to patient messages with next-generation artificial intelligence (AI) tools as an early adopter. The rollout presented our AI governance committee with novel challenges, particularly regarding transparency in the use of AI with our patients, which we chose to make explicit with an automated disclosure. Since our launch, hundreds of other institutions have adopted this same tool without clear guidelines on transparency regarding the disclosure of AI use to patients. We share the subsequent experience and call for professional organizations to help create disclosure guidelines that can be used nationally.

AB 3030: Health care services: artificial intelligence.

Session Year: 2023-2024 **House:** Assembly

Current Status: PASSED ✓ (2024-09-28: Chaptered by Secretary of State - Chapter 848, Statutes of 2024.)

Introduced → First Committee Review → First Chamber → Second Committee Review → Second Chamber → Enacted

Summary | Bill Text | Status | Votes | Supporters & Opponents | Analysis

Version: Chaptered (9/28/2024)

Bill Author

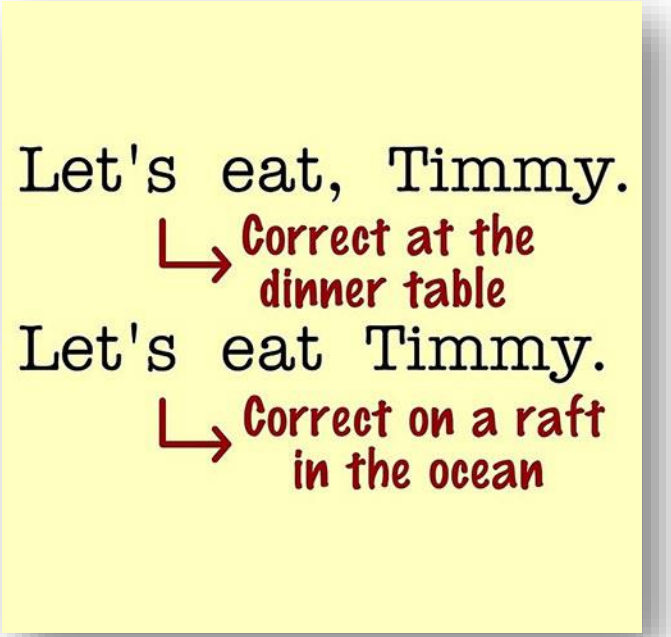
Lisa Calderon (D)

Bill Co-Author(s):

Susan Rubio (D)

Starting January 1, 2025, AB 3030 requires any health facility, clinic, or physician's office (including a group practice) to notify their patients when using GenAI to communicate patient clinical information. **The law, however, exempts from this requirement any communication that is read and reviewed by a licensed or certified human health care provider.**

April 2025



npj | digital medicine

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01586-2

Opportunities and risks of artificial intelligence in patient portal messaging in primary care

"This study assessed whether PCPs could identify and correct errors in AI-generated draft responses to patient portal messages. Twenty practicing PCPs reviewed 18 patient portal messages, 4 of which contained errors categorized as objective inaccuracies or potentially harmful omissions. **Each error was insufficiently addressed by 13–15 participants, and 35–45% of erroneous drafts were submitted entirely unedited.**

Compared with What? Measuring AI against the Health Care We Have

Isaac S. Kohane, M.D., Ph.D.

A recent faculty recruit in my department, an expert in artificial intelligence (AI), asked me whether I could recommend a primary care doctor. I assured my colleague that I would promptly provide a set of excellent options to choose from. When I reached out to several doctors who were my contemporaries in medical training, however, I learned that they had all retired. That was not completely surprising, so I tried another tack: I contacted doctors who were of more recent vintage and whom I'd encountered in a supervisory capacity. One of them cheerfully informed me that he was no longer practicing primary

ened my search further. I discovered that primary care clinics at some local hospitals, such as Massachusetts General Hospital and Brigham and Women's Hospital, were turning away new patients. I did not end up helping my new colleague and instead suggested contacting their health insurance company for recommendations.

Massachusetts has among the highest number of doctors per capita in the United States, which suggests that an attempt to make a referral for primary care would have been even more challenging in large swaths of the country. A recent report from the Association of American Medical Col-

menting the work of clinicians — including doctors, nurse practitioners, and physician assistants — with AI is being seriously considered.[2] Yet AI tools that could reduce burdens on clinicians are most likely the tip of the iceberg when it comes to AI's influence in medicine. Whether out of desperation, frustration, or curiosity, large numbers of patients are already using AI to obtain medical advice, including second opinions — sometimes with dramatic therapeutic consequences.[3] But individual anecdotes do not substitute for systematic evaluation. With any new clinical intervention, rigorous trials are the med-

JAMA Open, Oct 2024

An unexpected result was that the LLM alone performed significantly better than both groups of humans:

When A.I. worked independently to diagnose patients, it achieved 92 percent accuracy, while physicians using A.I. assistance were only 76 percent accurate — barely better than the 74 percent they achieved without A.I.

JAMA Network Open

Original Investigation | Health Informatics

Large Language Model Influence on Diagnostic Reasoning
A Randomized Clinical Trial

Ethan Goh, MBBS, MS; Robert Gallo, MD; Jason Hom, MD; Eric Strong, MD; Yingjie Weng, MHS; Hannah Kerman, MD; Joséphine A. Cool, MD; Zahir Kanjee, MD, MPH; Andrew S. Parsons, MD, MPH; Neera Ahuja, MD; Eric Horvitz, MD, PhD; Daniel Yang, MD; Arnold Milstein, MD; Andrew P. J. Olson, MD; Adam Rodman, MD, MPH; Jonathan H. Chen, MD, PhD

Abstract

IMPORTANCE Large language models (LLMs) have shown promise in their performance on both multiple-choice and open-ended medical reasoning examinations, but it remains unknown whether the use of such tools improves physician diagnostic reasoning.

OBJECTIVE To assess the effect of an LLM on physicians' diagnostic reasoning compared with conventional resources.

DESIGN, SETTING, AND PARTICIPANTS A single-blind randomized clinical trial was conducted from November 29 to December 29, 2023. Using remote video conferencing and in-person participation across multiple academic medical institutions, physicians with training in family medicine, internal medicine, or emergency medicine were recruited.

INTERVENTION Participants were randomized to either access the LLM in addition to conventional diagnostic resources or conventional resources only, stratified by career stage. Participants were allocated 60 minutes to review up to 6 clinical vignettes.

MAIN OUTCOMES AND MEASURES The primary outcome was performance on a standardized rubric of diagnostic performance based on differential diagnosis accuracy, appropriateness of supporting and opposing factors, and next diagnostic evaluation steps, validated and graded via blinded expert consensus. Secondary outcomes included time spent per case (in seconds) and final diagnosis accuracy. All analyses followed the intention-to-treat principle. A secondary exploratory analysis evaluated the standalone performance of the LLM by comparing the primary outcomes between the LLM alone group and the conventional resource group.

RESULTS Fifty physicians (26 attendings, 24 residents; median years in practice, 3 [IQR, 2-8]) participated virtually as well as at 1 in-person site. The median diagnostic reasoning score per case was 76% (IQR, 66%-87%) for the LLM group and 74% (IQR, 63%-84%) for the conventional resources-only group, with an adjusted difference of 2 percentage points (95% CI, −4 to 8 percentage points; $P = .60$). The median time spent per case for the LLM group was 519 (IQR, 371-668) seconds, compared with 565 (IQR, 456-788) seconds for the conventional resources group, with a time difference of −82 (95% CI, −195 to 31; $P = .20$) seconds. The LLM alone scored 16 percentage points (95% CI, 2-30 percentage points; $P = .03$) higher than the conventional resources group.

Key Points

Question Does the use of a large language model (LLM) improve diagnostic reasoning performance among physicians in family medicine, internal medicine, or emergency medicine compared with conventional resources?

Findings In a randomized clinical trial including 50 physicians, the use of an LLM did not significantly enhance diagnostic reasoning performance compared with the availability of only conventional resources.

Meaning In this study, the use of an LLM did not necessarily enhance diagnostic reasoning of physicians beyond conventional resources; further development is needed to effectively integrate LLMs into clinical practice.

A.I. Performance Exceeds Physician Plus A.I.?

Interpretative Task	AI Comparison	Study Design	Main Results AI vs MDs + AI	Citation
CXR	227 radiologists	CXR for 15 tasks	AI better than 2/3	Agarwal, NBER, 2024
Mammography	14 radiologists	C		n HE, Lancet Digital Health, 20
CXR	13 physicians			purkar P, NPJ Digital Medicine, 20
Diagnostic accuracy	20 internists	R		Duff D, arXiv, 2023
Diagnostic accuracy	50 physicians			h E, JAMA Network Open, 2024
Management reasoning	25 physicians	G		odeur P, arXiv, 2024

*Top-10 accuracy; ^ media controlled trial

UCSa

The New York Times

OPINION

GUEST ESSAY

The Robot Doctor Will See You Now

Feb. 2, 2025

By Pranav Rajpurkar and Eric J. Topol
Dr. Rajpurkar researches the application of A.I. in clinical settings. Dr. Topol is a cardiologist.

This research is early and may evolve. But the findings more broadly indicate that right now, simply giving physicians A.I. tools and expecting automatic improvements doesn't work. Physicians aren't completely comfortable with A.I. and still doubt its utility, even if it could demonstrably improve patient care.

But A.I. will forge ahead, and the best thing for medicine to do is to find a role for it that doctors can trust. The solution, we believe, is a deliberate division of labor. Instead of forcing both human doctors and A.I. to review every case side by side and trying to turn A.I. into a kind of shadow physician, a more effective approach is to let A.I. operate independently on suitable tasks so that physicians can focus their expertise where it matters most.

VIEWPOINT

Resolving the Productivity Paradox of Health Information Technology

A Time for Optimism

Robert M. Wachter, MD
Department of Medicine, University of California, San Francisco.

Michael D. Howell, MD, MPH
Google, LLC, Mountain View, California.

During the past decade, the US health care system has gone digital. In 2008, fewer than 1 in 10 US hospitals had an electronic health record (EHR) system; today, fewer than 1 in 10 does not. The increase in use of an EHR system in ambulatory practices has been similarly steep.

Although the evidence that health care digitization has led to improvements in quality and safety is generally positive,[1] these have been context-dependent improvements and relatively small. Moreover, rates of physician self-reported burnout are high, partly because little useful intelligence is delivered back to physicians despite all their time spent performing data entry. In addition, the presence of the computer often separates the physician from the patient, rather than enhancing communication.

Perhaps most surprising, gains in efficiency and productivity have not materialized. This represents a clear example of the productivity paradox of information technology—the observation that productivity does not immediately increase with investment in information technology.[2]

The good news from other industries is that the productivity paradox eventually resolves. In health care, the combination of widespread digitization, advances in machine learning and artificial intelligence, the opening of the health information technology business environment to both established technology companies and start-ups focused on solving specific health care problems, and growing competencies in systems thinking, work redesign, and population health have put health care on the cusp of potentially substantial improvements in quality, safety, patient and clinician experience, and efficiency. In this Viewpoint, we describe several reasons for optimism.

to new public repo
tems that began h
clinicians accoun
health care organi
ity and safety dep
provement techn
focused on higher
Even though
these efforts, its c
have been expect
care, EHR decisio
delivered as poor
high rates of false
tions that lead to l
problems. Clinical
more useful, ofte
questionable vera
data from EHRs f
has been the cap
tools and apps—
parties—back into

Why Might Now

The lesson from o
ity paradox even
decades. This was
factories at the sta
zation of manufact
of this century. Re
this resolution: im
reimagining of the
similar improveme
The first reaso
digital business e

JAMA | Special Communication | AI IN MEDICINE

Will Generative Artificial Intelligence Deliver on Its Promise in Health Care?

Robert M. Wachter, MD; Erik Brynjolfsson, PhD

IMPORTANCE Since the introduction of ChatGPT in late 2022, generative artificial intelligence (genAI) has elicited enormous enthusiasm and serious concerns.

OBSERVATIONS History has shown that general purpose technologies often fail to deliver their promised benefits for many years ("the productivity paradox of information technology"). Health care has several attributes that make the successful deployment of new technologies even more difficult than in other industries; these have challenged prior efforts to implement AI and electronic health records. However, genAI has unique properties that may shorten the usual lag between implementation and productivity and/or quality gains in health care. Moreover, the health care ecosystem has evolved to make it more receptive to genAI, and many health care organizations are poised to implement the complementary innovations in culture, leadership, workforce, and workflow often needed for digital innovations to flourish.

CONCLUSIONS AND RELEVANCE The ability of genAI to rapidly improve and the capacity of organizations to implement complementary innovations that allow IT tools to reach their potential are more advanced than in the past; thus, genAI is capable of delivering meaningful improvements in health care more rapidly than was the case with previous technologies.

JAMA. 2024;331(1):65-69. doi:10.1001/jama.2023.25054
Published online November 30, 2023.

Viewpoint page 17
CME at jamacmelookup.com

Author Affiliations: Department of Medicine, University of California, San Francisco (Wachter); Digital Economy Lab and Institute for Human-Centered AI, Stanford University, Palo Alto, California (Brynjolfsson).

Corresponding Author: Robert M. Wachter, MD, Department of Medicine, University of California, San Francisco, 505 Parnassus Ave, Room M994, Box 0120, San Francisco, CA 94143-0120 (robert.wachter@ucsf.edu).

Since they became publicly available in late 2022, generative artificial intelligence (genAI) tools such as ChatGPT have elicited enormous enthusiasm, as well as concern, in all sectors of the economy. But the potential impact of genAI in health care seems particularly noteworthy. In a field in which an estimated 30% of the $4.3 trillion spent each year in the US adds little to no value,[1] in which many tens of thousands of people die yearly from preventable mistakes,[2,3] and in which access to care is fragmented and inequities are commonplace, it is natural to be enthusiastic about the

The Productivity Paradox of Information Technology

Research on the productivity paradox has demonstrated that there are 2 fundamental reasons that technologies don't rapidly create their promised value.[6] The first is that early versions of many technologies are flawed; the tools that ultimately succeed are ones that have improved with successive iterations.

Thank you!

The purpose of knowledge is action, not knowledge.

- Aristotle

clonghurst@health.ucsd.edu

@calonghurst

Melbourne Children's
A world leader in child and adolescent health
The Royal Children's Hospital Melbourne
murdoch children's research institute
THE UNIVERSITY OF MELBOURNE
Supported by The Royal Children's Hospital Foundation

Centre for Health Analytics

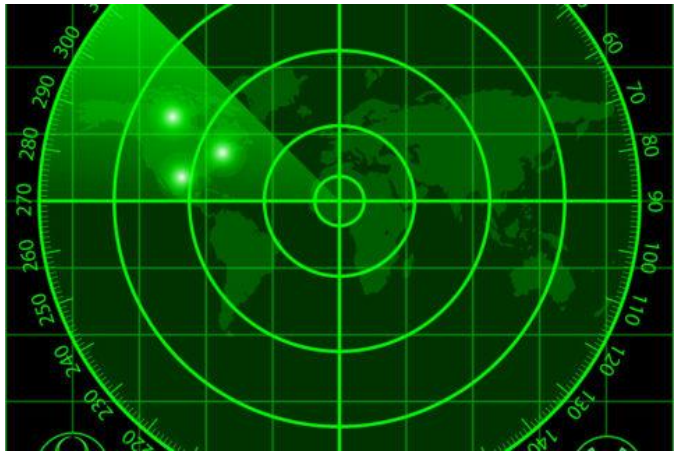
Global Vaccine Data Network®

Towards Machine Learning monitoring and response systems in vaccine safety and communication

Jim Buttery

Centre for Health Analytics
Royal Children's Hospital, Melbourne
Department of Paediatrics, University of Melbourne
Epidemiology Informatics, Murdoch Childrens Research Institute
Co-director, Global Vaccine Data Network

ML in vaccine safety and communication



Machine learning to detect Vaccine Adverse Event Mentions (VAEM)

JMIR Medical Informatics | Journal Information | Browse Journal

Published on 16.06.2022 in Vol 10, No 6 (2022): June

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/34305, first published October 16, 2021.

Vaccine Adverse Event Mining of Twitter Conversations: 2-Phase Classification Study

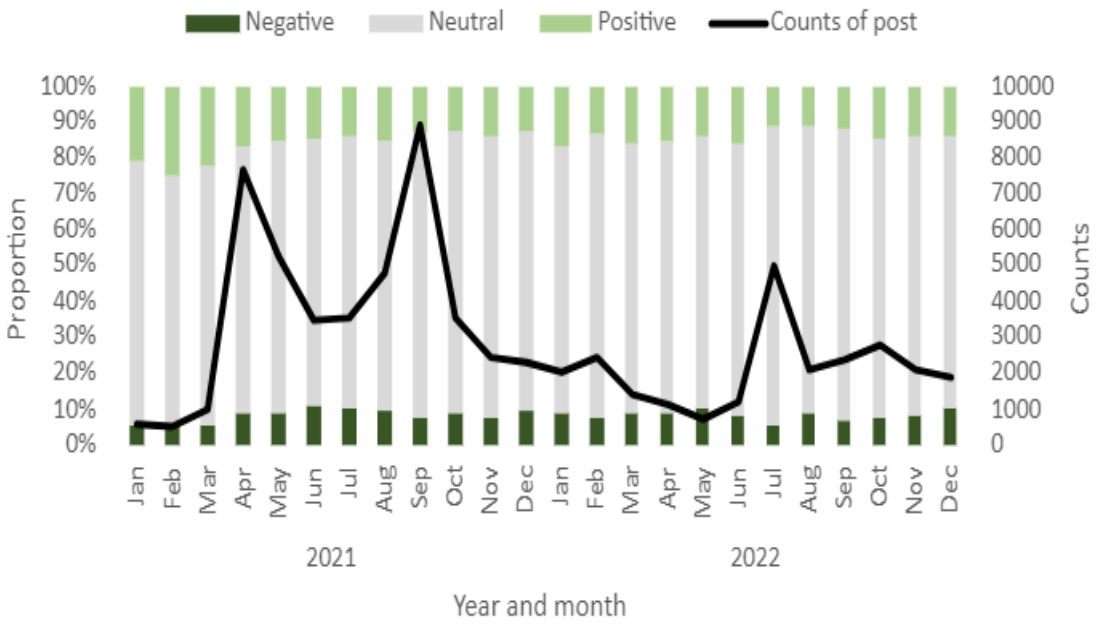
Sedigheh Khademi Habibabadi[1,2]; Pari Delir Haghighi[3]; Frada Burstein[3]; Jim Buttery[1,4]

Surveillance type	Coverage	Reactogenicity	Signal detection	Cost	Comments
Spontaneous					Disproportionality, Vigibase
Solicited/ CEM					Rapid, response fatigue
Hospital sentinel			/		Severe AESI
Syndromic*					Non-specific, rapid, high volume
Social media/ media			?		Geolocation, also sentiment
Large linked database					Rapid cycle analysis, GVDN, VAC4EU

Global Vaccine Data Network®

Jim P Buttery[a,b,c,d] and Hazel Clothier[b,c,e]
HUMAN VACCINES & IMMUNOTHERAPEUTICS
2022, VOL. 18, NO. 6, e2100173 (7 pages)

VaxPulse



·As a woman it is extremely disturbing the amount of women on that sub that are having menstrual side effects.

X USER
Feeling negative

Has there been anything published to show these vaccines have any effect on menstrual cycles?

REDDIT USER
Feeling neutral

I got my first dose 5 days ago on day 7 of my cycle and so far no menstrual issues.

X USER
Feeling positive

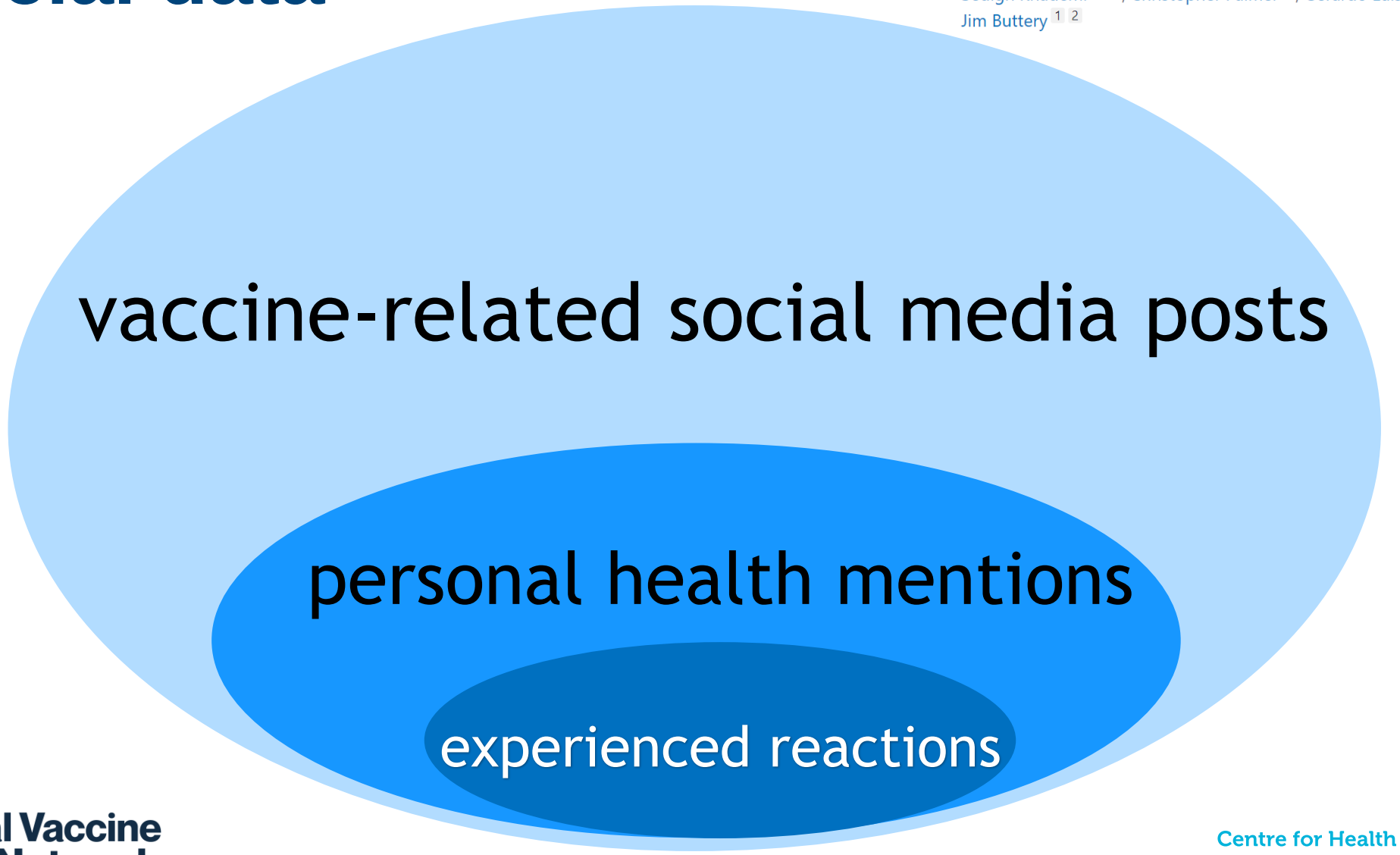
Global Vaccine Data Network®

Social data

Stud Health Technol Inform. 2024 Sep 24:318:30-35. doi: 10.3233/SHTI240887.

Exploring Large Language Models for Detecting Online Vaccine Reactions

Sedigh Khademi [1 2], Christopher Palmer [1], Gerardo Luis Dimaguila [1 2], Muhammad Javed [1], Jim Buttery [1 2]



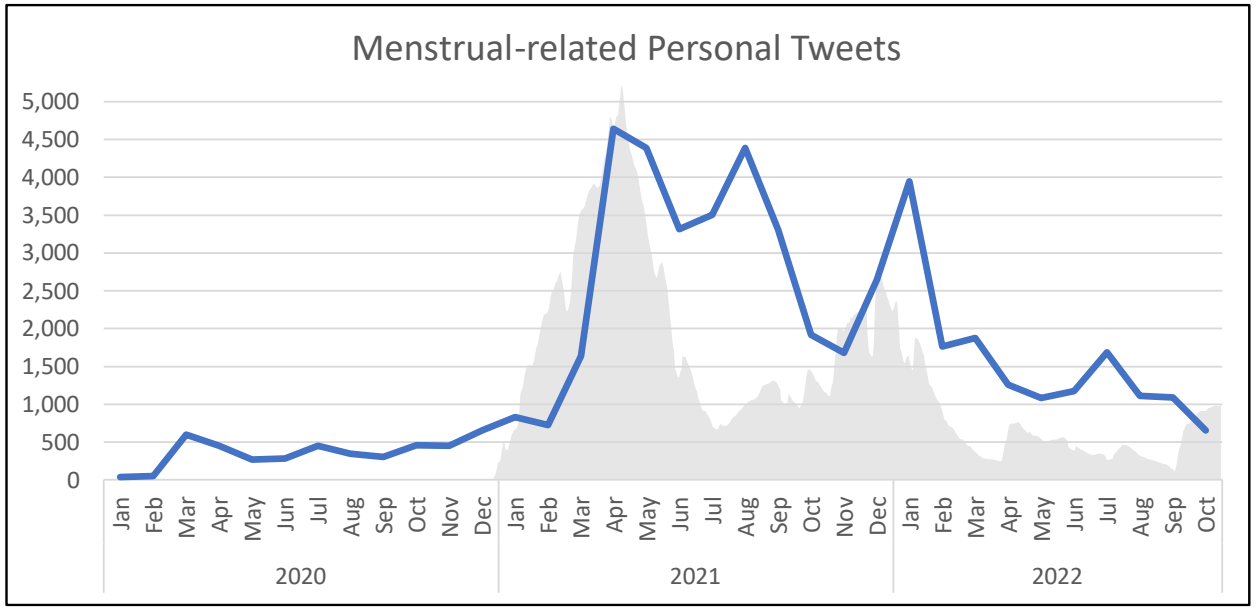
Stud Health Technol Inform. 2024 Jan 25:310:644-648. doi: 10.3233/SHTI231044.

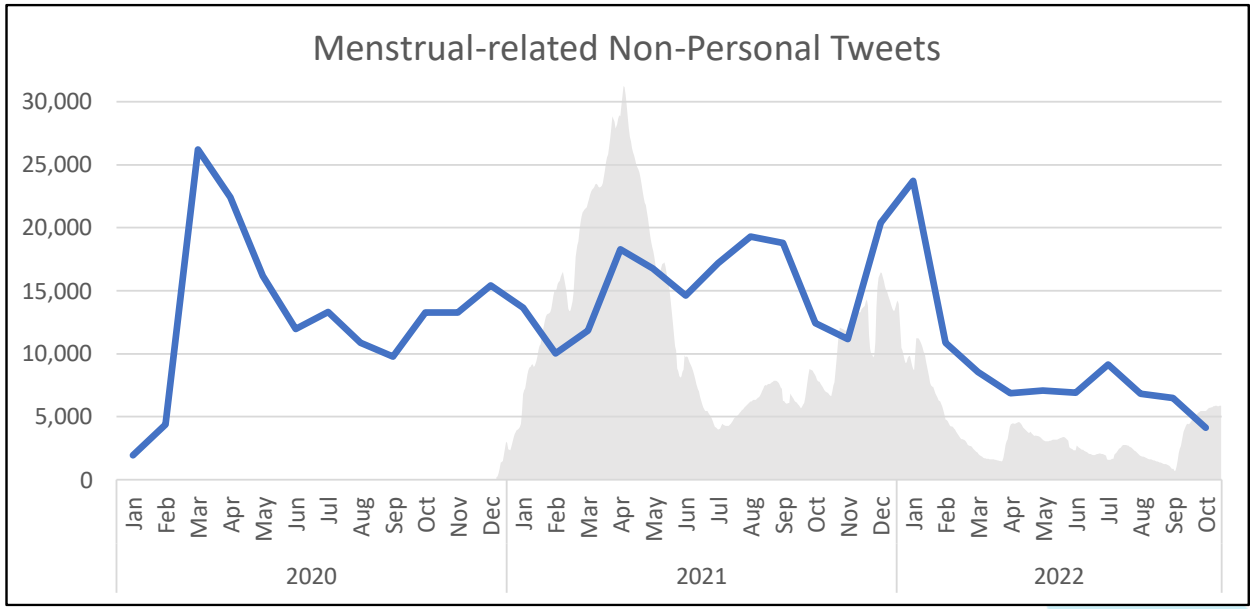
Fertility-Related Conversations in the Context of COVID-19 and Vaccinations

Christopher Palmer [1], Sedigh Khademi [1], Muhammad Javed [1], Gerardo Luis Dimaguila [1], Jim Buttery [1 2]

Tweets collected from January 2020 to October 2021

Keywords: COVID-19, COVID-19 vaccine, or vaccination keywords, combined with lists of words pertaining to fertility





Global Vaccine Data Network®

Centre for Health Analytics

SAFESIGGP

Contents lists available at ScienceDirect

Vaccine

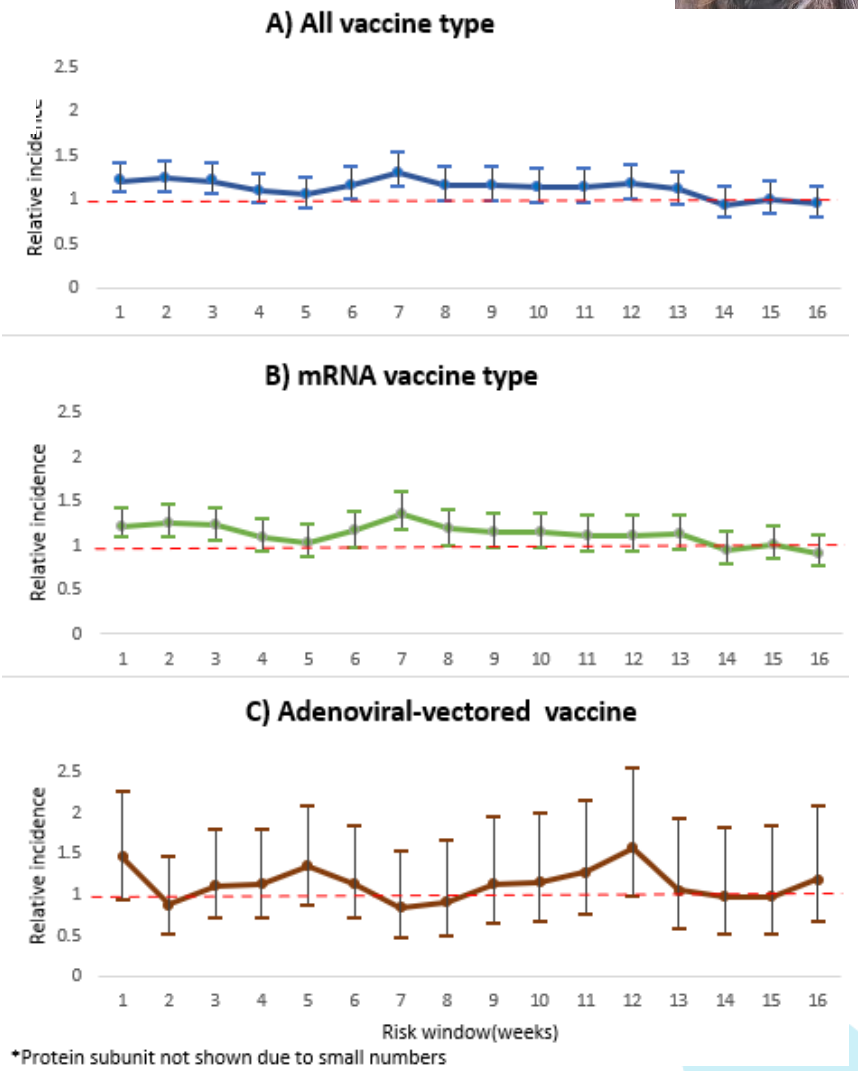
journal homepage: www.elsevier.com/locate/vaccine

Validating community concerns of menstrual changes associated with COVID-19 vaccination using a self-controlled case series analysis of real-world data

Aishwarya N. Shetty [a,b], Gonzalo Sepulveda Kattan [a,b], Muhammad Javed [a,b], Christopher Pearce [c], Jim P. Buttery [a,b,d,e,*,1], Hazel J. Clothier [a,b,d,*,1]

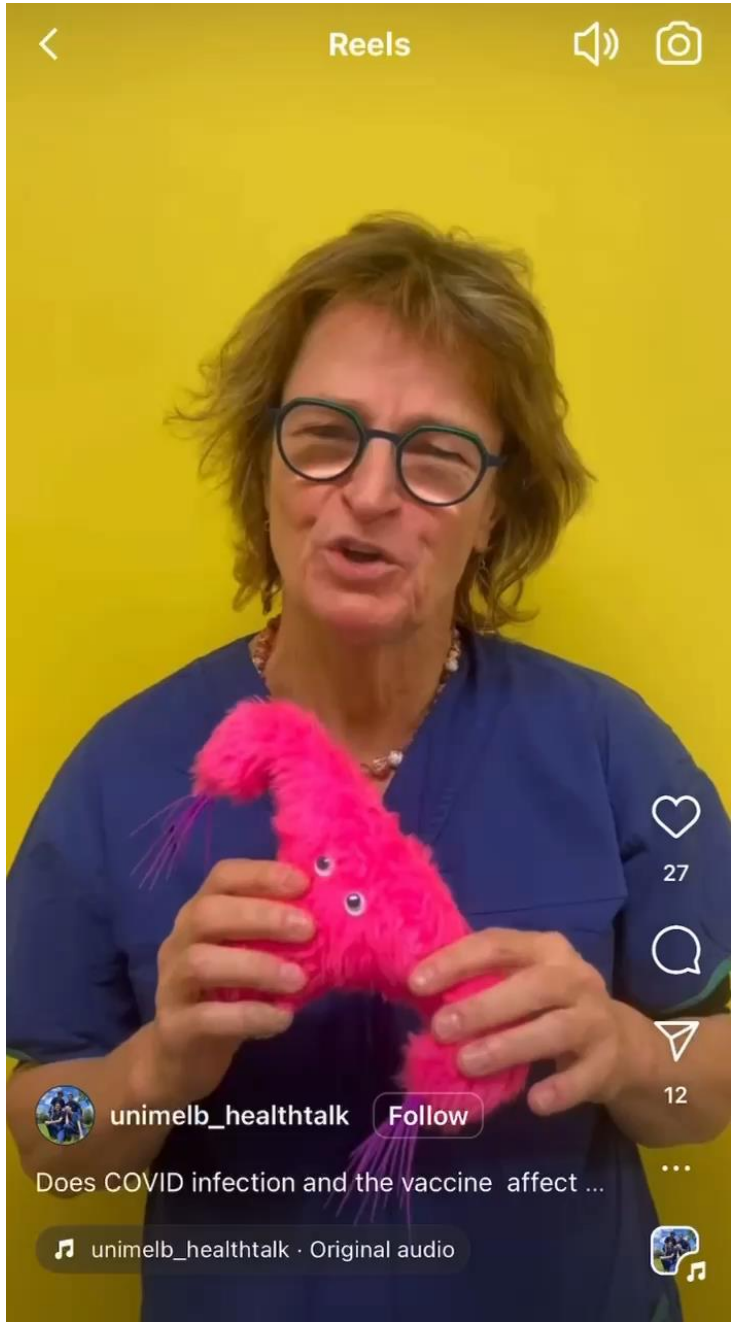
Menstrual cycle change presentations within six-weeks post COVID-19 vaccination by vaccine type

Vaccine type	Events	Relative incidence	95% Confidence interval	p-value
All vaccine type	1063	1.15	1.08, 1.22	<0.001
mRNA (Comirnaty® and Spikevax®)	953	1.14	1.07, 1.22	<0.001
mRNA (Comirnaty®)	927	1.15	1.07,1.21	<0.001
mRNA (Spikevax ®)	28	1.11	0.76, 1.63	0.57
Adenoviral-vectored (Vaxzevria®)	110	1.15	0.94, 1.40	0.15
Protein-subunit (Nuvaxovid®)	4	1.41	0.50, 3.93	0.52



*Protein subunit not shown due to small numbers

Global Vaccine Data Network®



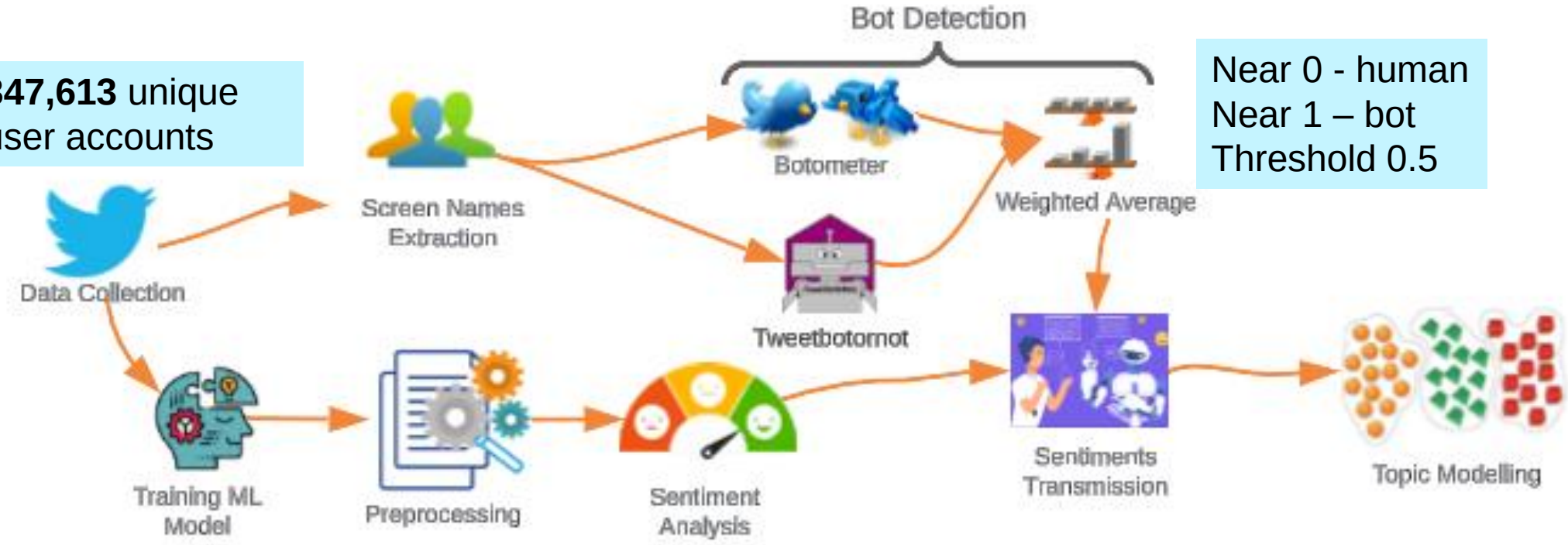
Global Vaccine Data Network®

Back to main

Social Bots Influence on COVID-19 Vaccination-Related Discussions:

Global Vaccine Data Network®

347,613 unique user accounts



Near 0 - human
Near 1 – bot
Threshold 0.5

Results

7.41% (n=25,759) of unique accounts classified as **bots**

92.59% (n=321,854) as **human** users

Social bots contributed **23.72%** (n=287,520) of **tweets** including retweets.

After **bot-to-human** transmission, we noticed significant **amplification** in tweets about discussion about **AEFI**, and some new discussions also identified.

Selected topics of discussion extracted from the humans posted tweets

Topic of discussion	Number of tweets
Blood Clots	597
Deaths in relation to COVID-19 vaccine	432
Women's hesitancies about AEFI during pregnancy	246
Heart Issues	233
Allergic Reactions	65
Side effects (in general)	295

. Selected topics of discussion extracted after the bot-to-humans transition

Topic of discussion	Number of tweets
Blood Clots	2,553
Deaths in relation to COVID-19 vaccine	1,137
Women's hesitancies about AEFI during pregnancy	960
Heart Issues	568
Allergic Reactions	141
Side effects (in general)	2,691
Hesitancies for the babies on breastfeeding	47

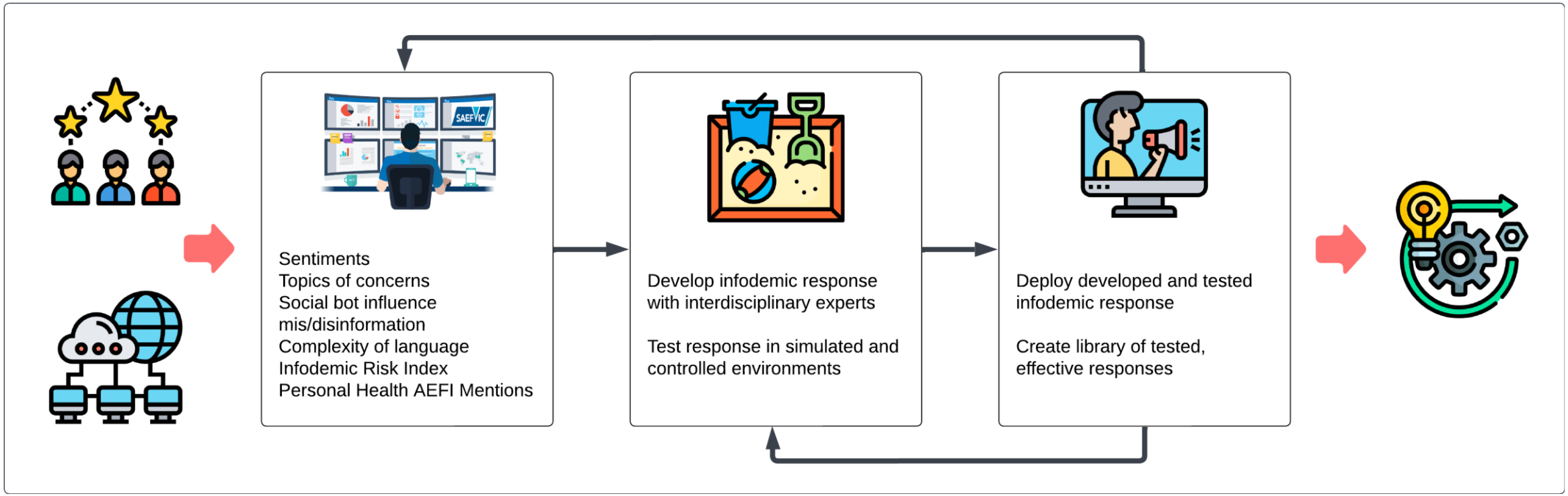
VaxPulse

Stud Health Technol Inform. 2024 Jan 25:310:1146-1150. doi: 10.3233/SHTI231144.

Interdisciplinary Learning Health System Response to Public Vaccine Concerns

Gerardo Luis Dimaguila [1 2], Muhammad Javed [1 2], Hazel J Clothier [1 2], Jo Hickman [1 2], Dan Petrovic [3], Francesca Machingaifa [4], Jessica Kaufman [5], Sedigh Khademi Habibabadi [1 2], Christopher Palmer [1 2], Jim Buttery [1 2]

Learning Health System



Global Vaccine Data Network®

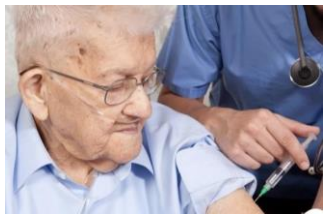
DOI: 10.3233/SHTI231144

Centre for Health Analytics

Opportunities

Cost-effective post-implementation

Equitable



Global Vaccine Data Network®

Epic

OHDSI

HL7 FHIR

Challenges

Differential SM platform use esp LMIC

Shifting access to data

Acknowledgements

Gerardo Luis Dimaguila

Gonzalo Sepulveda-Kattan

Sedigh Khademi

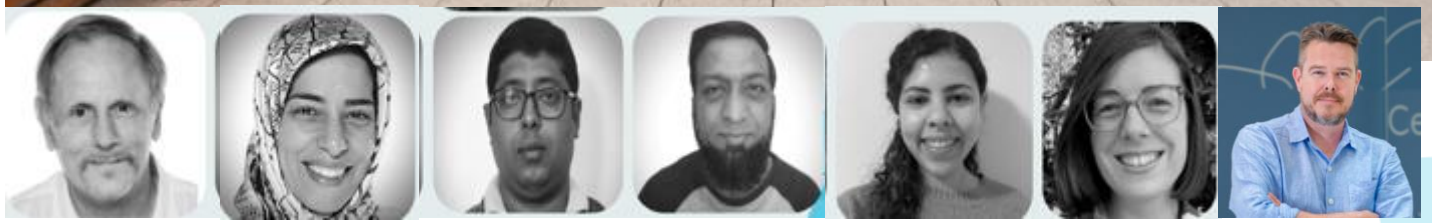
Muhammad Javed

Chris Palmer

Victoria Iliadis

Jo Hickman

Aishwarya Shetty



Thank you

Jim.buttery@mcri.edu.au

@jimbuttery..bsky.social

https://www.thisiscolossal.com/2017/06/enchanting-libraries-by-photographer-thibaud-poirier/

Citations from Chris Longhurst during the Q&A

Large Language Model Influence on Diagnostic Reasoning
A Randomized Clinical Trial

https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395

Health system-scale language models are all-purpose prediction engines

https://doi.org/10.1038/s41586-023-06160-y

Received: 14 October 2022

Accepted: 2 May 2023

Published online: 7 June 2023

Open access

Check for updates

Lavender Yao Jiang[1,2], Xujin Chris Liu[1,3], Nima Pour Nejatian[4], Mustafa Nasir-Moin[1], Duo Wang[5], Anas Abidin[4], Kevin Eaton[6], Howard Antony Riina[1], Ilya Laufer[1], Paawan Punjabi[6], Madeline Miceli[6], Nora C. Kim[1], Cordelia Orillac[1], Zane Schnurman[1], Christopher Livia[1], Hannah Weiss[1], David Kurland[1], Sean Neifert[1], Yosef Dastagirzada[1], Douglas Kondziolka[1], Alexander T. M. Cheung[1], Grace Yang[1,2], Ming Cao[1,2], Mona Flores[4], Anthony B. Costa[4], Yindalon Aphinyanaphongs[5,7], Kyunghyun Cho[2,8,9,10] & Eric Karl Oermann[1,2,11]✉

Physicians make critical time-constrained decisions every day. Clinical predictive models can help physicians and administrators make decisions by forecasting clinical and operational events. Existing structured data-based clinical predictive models have limited use in everyday practice owing to complexity in data processing, as well as model development and deployment[1–3]. Here we show that unstructured clinical notes from the electronic health record can enable the training of clinical language models, which can be used as all-purpose clinical predictive engines with low-resistance development and deployment. Our approach leverages recent advances in natural language processing[4,5] to train a large language model for medical language (NYUTron) and subsequently fine-tune it across a wide range of clinical and operational predictive tasks. We evaluated our approach within our health system for five such tasks: 30-day all-cause readmission prediction, in-hospital mortality prediction, comorbidity index prediction, length of stay prediction, and insurance denial prediction. We show that NYUTron has an area under the curve (AUC) of 78.7–94.9%, with an improvement of 5.36–14.7% in the AUC compared with traditional models. We additionally demonstrate the benefits of pretraining with clinical text, the potential for increasing generalizability to different sites through fine-tuning and the full deployment of our system in a prospective, single-arm trial. These results show the potential for using clinical language models in medicine to read alongside physicians and provide guidance at the point of care.

Physicians make difficult decisions every day requiring the integration of a tremendous amount of information. The information needed to make these medical decisions is scattered across various records, for example, a patient's medical history and laboratory and imaging reports. When physicians perform their work, however, all of this information is ultimately integrated into the notes written by physicians to document and summarize patient care.

Clinical predictive models are frequently derived from rules that have existed for decades[6–9], as well as from machine learning methods[10–12], with most relying on structured inputs pulled from the electronic health record (EHR) or direct clinician inputs. This reliance on structured inputs introduces complexity in data processing, as well as in model development and deployment, which in part is responsible for the overwhelming majority of medical predictive algorithms being trained, tested and published, yet never deployed to assess their impact on real-world clinical care. This is frequently referred to as the 'last-mile problem' (refs. 1–3).

One of the most exciting recent developments in modern artificial intelligence (AI) research is large language models (LLMs). These massive neural networks (with millions or even billions of parameters) have been shown to obtain impactful results on a wide range of problems that rely on the reading and interpretation of human language. Several styles of LLMs have been developed over the past few years, broadly ranging from encoder models (such as BERT[4]) to decoder models (such as GPT3; ref. 5). We theorized that LLMs could potentially solve the last-mile problem in medical predictive analytics by simply reading the notes written by physicians, thereby immediately accessing a comprehensive description of a patient's medical state to provide decision support at the point of care across a wide range of clinical and operational tasks.

Here we present our results from developing, evaluating, deploying and prospectively assessing NYUTron, an LLM-based system that can integrate in real time with clinical workflows centred around writing notes and placing electronic orders. Our approach relies on the fact that all clinically useful data and medical professionals' decision-making

[1]Department of Neurosurgery, NYU Langone Health, New York, NY, USA. [2]Center for Data Science, New York University, New York, NY, USA. [3]Electrical and Computer Engineering, Tandon School of Engineering, New York, NY, USA. [4]NVIDIA, Santa Clara, CA, USA. [5]Predictive Analytics Unit, NYU Langone Health, New York, NY, USA. [6]Department of Internal Medicine, NYU Langone Health, New York, NY, USA. [7]Department of Population Health, NYU Langone Health, New York, NY, USA. [8]Prescient Design, Genentech, New York, NY, USA. [9]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA. [10]Canadian Institute for Advanced Research, Toronto, Ontario, Canada. [11]Department of Radiology, NYU Langone Health, New York, NY, USA. ✉e-mail: eric.oermann@nyulangone.org

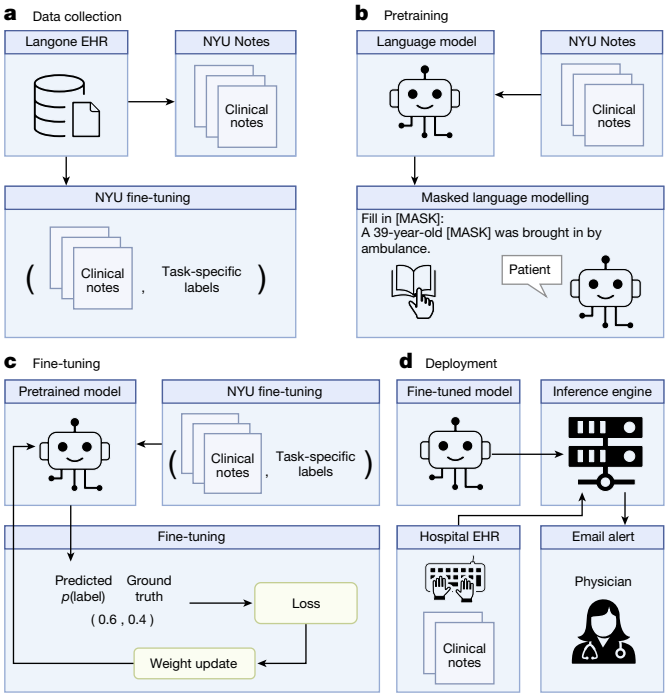


Fig. 1 | Overview of the language model-based approach for clinical prediction. a, We queried the NYU Langone EHR for two types of datasets. The pretraining dataset, NYU Notes, contains 10 years of inpatient clinical notes (387,144 patients, 4.1 billion words). There are five fine-tuning datasets. Each contains 1–10 years of inpatient clinical notes (55,791–413,845 patients, 51–87 million words) with task-specific labels (2–4 classes). **b**, We pretrained a 109 million-parameter BERT-like LLM, termed NYUTron, on the entire EHR using an MLM task to create a pretrained model for medical language contained within the EHR. **c**, We subsequently fine-tuned the pretrained model on specific tasks (for example, 30-day all-cause readmission prediction) and validated it on held-out retrospective data. **d**, Lastly, the fine-tuned model was compressed into an accelerated format and loaded into an inference engine, which interfaces with the NYU Langone EHR to read discharge notes when they are signed by treating physicians.

processes can be found as structured or unstructured text in the EHR (for example, as notes, laboratory results and reports on studies). Our approach leverages recent advances in natural language processing that suggest that sufficiently scaled, self-supervised LLMs can outperform strongly supervised approaches on non-medical predictive tasks[4,5,13]. We investigate our hypothesis in the NYU Langone Health System ('NYU Langone'), a large multi-borough hospital system with a diverse patient population in New York, with 4 urban hospitals and 350 outpatient sites. We assess NYUTron on a battery of five tasks, including three clinical and two operational tasks (30-day all-cause readmission prediction, in-hospital mortality prediction, comorbidity index prediction, length of stay (LOS) prediction and insurance denial prediction) and provide a detailed analysis of our 30-day readmission task to look at questions of data efficiency, generalizability, deployability and potential clinical impact. By rethinking all of medical predictive analytics (see Supplementary Information section 1.1 for previous works) as a natural language processing problem, we show that it is possible to use LLMs as universal prediction engines for a wide range of medical predictive tasks.

Language model-based clinical prediction

Our language model-based approach has four steps: data collection, pretraining, fine-tuning and deployment. In the first step (Fig. 1a), we collected a vast set of unlabelled clinical notes and five task-specific labelled clinical notes from the NYU Langone EHR. Unlike other studies, our datasets come from the entire hospital system with a diverse patient population from different clinical departments. Our large unlabelled dataset, 'NYU Notes', comprises 7.25 million clinical notes (for example, radiographic reads, history and physicals) from 387,144 patients across four hospitals, resulting in a 4.1 billion-word corpus curated from January 2011 to May 2020. Each one of our labelled fine-tuning sets contains 1–10 years of inpatient clinical notes (55,791–413,845 patients, 51–87 million words) with task-specific labels (2–4 classes). See Extended Data Table 1 for dataset statistics.

In the second and third steps (Fig. 1b,c), we pretrained and fine-tuned an LLM for each downstream task using a bidirectional encoder model known as BERT (Bidirectional Encoder Representation with Transformer) and a masked language modelling (MLM) objective on the NYU Notes dataset[11] until the validation loss plateaued. The MLM objective randomly masks words or subwords in clinical notes and trains the language model to fill in the masked word correctly. Next, using the fine-tuning dataset, we fine-tuned the pretrained model (termed 'NYUTron') to predict the task label using the relationships learned in pretraining with clinical notes.

In the fourth step (Fig. 1d), we deployed our best model to a high-performance inference engine, NYUTriton, that interfaces with the NYU Langone EHR. Deployment enabled real-time LLM-guided inference at the point of care. In a single-arm, non-interventional, prospective trial, we validated NYUTron's performance on 30-day readmission prediction in a real-world environment and assessed its potential clinical impacts.

Overall performance on five tasks

To assess the breadth of NYUTron's applicability, we evaluated NYUTron's performance on five tasks retrospectively. We trained with the full dataset and evaluated performance with two test sets: (1) a random test set (clinical notes sampled from the same time as the training data) and (2) a temporal test set (clinical notes sampled from the future of the training data). The temporal test set more closely resembles the deployment scenario, in which inference data come from the future of the training data. Our battery of tasks consisted of three clinical tasks and two operational tasks, as shown in Fig. 2a. We compared NYUTron against structured baselines, which forward structured features used by traditional clinical predictive models into an extreme gradient-boosted tree[14] model.

NYUTron is capable of being extended to multiple clinical and operational tasks. Figure 2b and Fig. 2c show that, on prediction tasks (in-hospital mortality, readmission, LOS and insurance denial), NYUTron had an area under the curve (AUC) of 78.7–94.9%, with an improvement of 5.36–14.7% in AUC from traditional clinical predictive models. On the comorbidity index imputation task, NYUTron had a median AUC of 89.4% ± 0.275%. We first present our results across four of the tasks and conclude with a focused look at readmission prediction that addresses questions of data efficiency, model generalizability and deployment in a real-world environment.

NYUTron is capable of predicting risk of in-hospital mortality on admission and imputing a comorbidity index. The task of in-hospital mortality prediction was to estimate (at admission) the likelihood of a patient's death during the present inpatient encounter. Figure 2b shows that, for in-hospital mortality prediction, NYUTron had a median AUC of 94.9% ± 0.168%, with a 7.43% improvement from its structured baseline based on Simplified Acute Physiology Score (SAPS2)[15] and Acute Physiology and Chronic Health Evaluation (APACHE2)[16] features such as age and mean heart rate. The task of comorbidity index imputation was to predict (at admission) the Charlson comorbidity index (CCI)[17] with no available structured features for chronic diseases. We framed this as a data imputation problem, as 22% of our dataset lacked CCI scores and this was a known area for documentation improvement

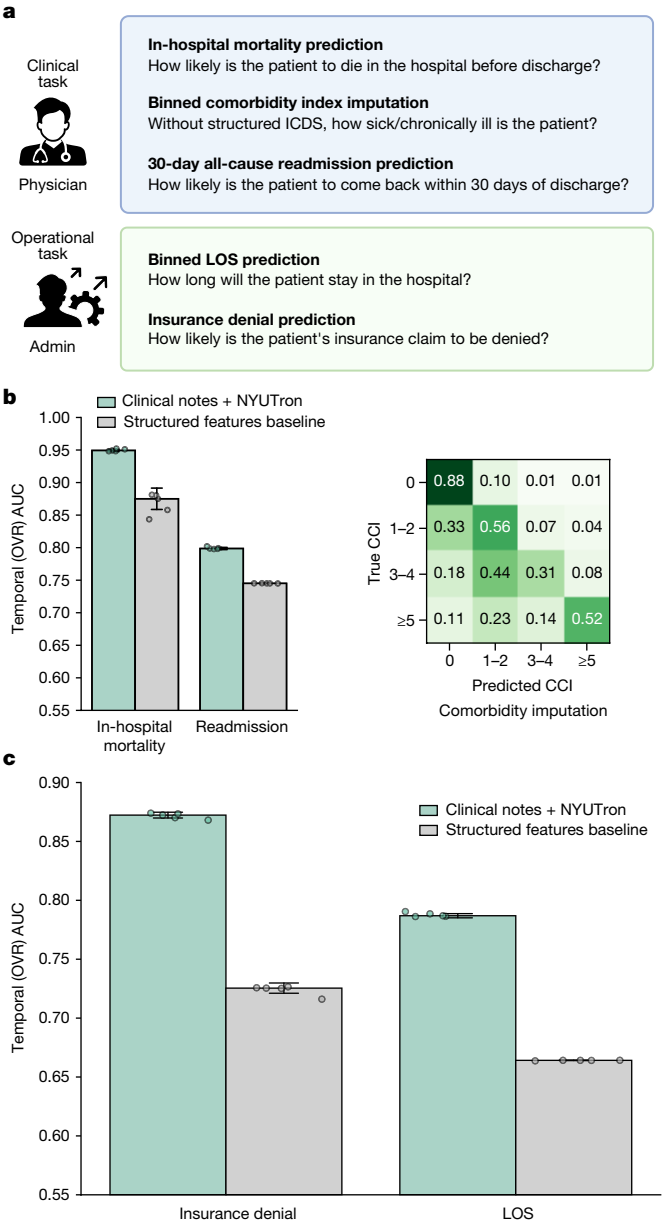


Fig. 2 | Overall temporal test performance across five tasks. a, The five tasks include three clinical tasks and two operational tasks. **b**, On readmission prediction, NYUTron had a median AUC of 79.9% ± 0.168% with a 5.36% improvement. On in-hospital mortality prediction, NYUTron had a median AUC of 94.9% ± 0.168% with a 7.43% improvement. On comorbidity index imputation, NYUTron had an OVR median AUC of 89.4% ± 0.275%. A confusion matrix is shown on the right. **c**, On binned LOS prediction, NYUTron had a median AUC of 78.7% ± 0.179% with a 12.3% improvement from the structured baseline. On insurance denial prediction, NYUTron had a median AUC of 87.2% ± 0.246% with a 14.7% improvement. For **b**,**c**, the height of the error bar is the median AUC and the half-width of the error bar is 1 s.d. The grey points are individual data points from $n = 5$ experiments using distinct random seeds.

(see Supplementary Information section 2.3 for more context). We discretized the index into four bins according to the original paper's grades of severity (0, none; 1–2, mild; 3–4, moderate; ≥5, severe). Figure 2b shows that, on comorbidity imputation, NYUTron had a median AUC of 89.4% ± 0.275% and 88% precision when identifying patients whose CCI score was 0.

NYUTron can also be used for operational endpoints and to predict in-patient LOS and insurance claim denial on admission. The task of LOS prediction was to predict (at admission) the likely range of days a patient would stay in the hospital. We discretized LOS into four bins (0–25% quantile, 25–50% quantile, 50–75% quantile, >75% quantile). Figure 2c shows that, for LOS prediction, NYUTron had a median one-versus-rest (OVR) AUC of 78.7% ± 0.179%, with a 12.3% improvement from the structured baseline, which used an available subset of 'Lisbon Portugal' features[18]. The task of insurance claim denial prediction was to predict (at admission) whether the insurance claims submitted for an encounter would be accepted or initially denied. Figure 2c shows that, for insurance denial prediction, NYUTron had a median AUC of 87.2% ± 0.246%, with a 14.7% improvement from the structured baseline, which used an available subset of 'claim form' features[19] such as age and insurance provider. NYUTron is also capable of predicting different types of denials from both admission notes and discharge notes with similar performance (Supplementary Information section 2.2).

Detailed analysis on readmission

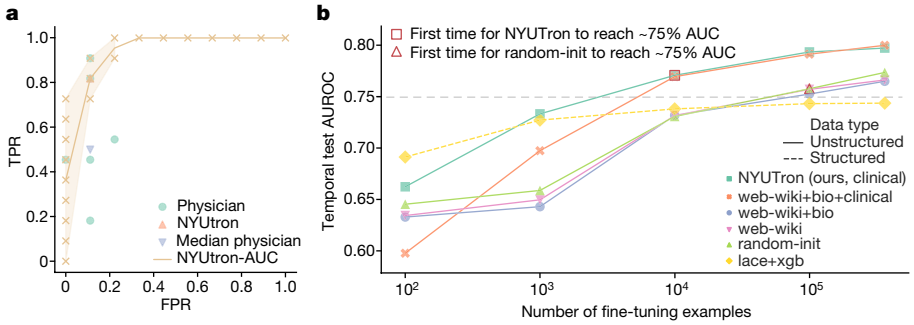
To better understand NYUTron's performance, we carried out a detailed analysis of 30-day all-cause readmission prediction. The task of readmission prediction is to predict (at discharge) the likelihood of a patient coming back to the hospital within 30 days and is a well-studied problem in the medical informatics literature (see Supplementary Information section 2.1 for more details on the readmission prediction task). Figure 2b shows that, for 30-day all-cause readmission prediction, NYUTron had a median AUC of 79.87% ± 0.168%, with a 5.36% improvement from its structured baseline, which used LACE[20] features (a mnemonic for LOS, acuity of admission, Charlson comorbidity index and number of emergency department visits in the past 6 months). We performed five additional evaluations in both retrospective and prospective settings: (1) a human comparison with six attending physicians for prediction of readmission for 20 patient cases sampled from a random split, (2) a study of NYUTron's scaling properties with respect to data in which NYUTron and other models were compared using a different number of fine-tuned data points, (3) an assessment of NYUTron's cross-site generalizability using pretraining, fine-tuning and test data from different locations, (4) a prospective, single-arm, non-interventional study to evaluate NYUTron's deployability and (5) a qualitative evaluation by a physician panel of NYUTron's prospective performance to assess clinical impacts.

Retrospective study of readmission

On small samples, NYUTron was competitive with a small group of physicians at predicting 30-day readmission. We tested a group of six physicians at different levels of seniority against NYUTron in a head-to-head comparison to establish a baseline difficulty for predicting 30-day all-cause readmission at the time of discharge. Discharge summaries ($n = 20$, including 11 positive cases and 9 negative cases) were sampled from a random split and uploaded to an online evaluation platform. Median physician performance was worse than that of NYUTron (Fig. 3a). For physicians and NYUTron, the median false positive rate (FPR) was 11.11%, whereas the median true positive rate (TPR) was 50% for physicians compared with 81.82% for NYUTron. Physicians had a median F1 score of 62.8% and substantial variance of 22.2% compared with NYUTron, which had a median F1 score of 77.8%.

The random split does not resemble the deployment scenario, in which the test data come from the future of the training data. We therefore created a temporal split to simulate deployment and observed a meaningful difference in test statistics compared with the random split (the random test AUC was 84.13%, whereas the temporal test AUC was 80.2%), confirming the importance of this second testing phase (further comparison in Extended Data Fig. 1).

NYUTron is competitive with traditional models and other LLMs. We evaluated the effectiveness of NYUTron by comparing its test performance on the temporal split against that of a traditional model



c Test Manhattan

Pretrain	Fine-tune: Manhattan	Fine-tune: Brooklyn
Manhattan	81.73% ± 0.26%	80.70% ± 0.32%
Brooklyn	80.91% ± 0.34%	78.6% ± 0.60%
All sites	81.61% ± 0.32%	80.99% ± 0.17%

d Test Brooklyn

Pretrain	Fine-tune: Manhattan	Fine-tune: Brooklyn
Manhattan	78.64% ± 0.27%	80.55% ± 0.29%
Brooklyn	78.74% ± 0.56%	79.65% ± 0.69%
All sites	79.14% ± 0.92%	81.24% ± 0.55%

Fig. 3 | Retrospective study of NYUTron's readmission prediction. a, On 20 cases sampled from a random split, we compared NYUTron's TPR and FPR with those for six physicians. NYUTron (orange triangles) had a higher TPR and the same FPR when compared with the median physician performance (green circles). The error band for AUC ranges from the minimum to maximum, and the orange crosses indicate TPR and FPR using all possible thresholds. We chose NYUTron's threshold on the basis of validation data. **b**, Comparison of the temporal test AUCs of different pretrained LLMs with an increasing number of fine-tuning examples. For simplicity, we omit the variance and only plot the median performance of five trials. Differences in median performance with 100 and 1,000 examples are less notable because AUCs with sparse fine-tuning examples have high variance (at 100 examples, we had 4.26% to 9.56% variance; at 1,000 examples, we had 0.44% to 9.46% variance). AUC variance decreases with more fine-tuning examples. The horizontal dashed line at 0.75 corresponds to the threshold for performance. See alternative presentations in Extended Data Fig. 7. **c**,**d**, Temporal test performance of NYUTron using pretraining, fine-tuning and test data from different sites. For both the Manhattan and Brooklyn tests, the column corresponding to local fine-tuning shows better performance than that with external fine-tuning. Each entry in **c**,**d** is presented as the mean ± 1 s.d. for $n = 5$ experiments using distinct random seeds.

and four different types of LLMs. NYUTron had the highest AUC when fine-tuned with the full dataset (Fig. 3b), with a median AUC of 79.87% ± 0.17%, which was similar to the clinical+web-wiki+bio AUC of 80.14% ± 0.26%. Compared with LLMs pretrained with non-clinical text (web-wiki+bio and web-wiki), NYUTron's median AUC was 2.37% to 3.23% higher. Compared with the traditional model that uses structured features (lace+xgb), NYUTron had a 5.36% higher AUC. Compared with a model using traditional natural language processing (NLP) embedding (tf-idf+xgb), NYUTron had a 12.8% higher median AUC (Extended Data Fig. 2a).

An LLM trained on unstructured clinical notes better scales with data than traditional structured models. Compared with lace+xgb, NYUTron benefits from an increasing amount of labelled examples and achieved a better AUC when fine-tuned with the full dataset. Figure 3b shows that lace+xgb (dashed yellow line) and NYUTron (solid green line) had similar AUCs at 100 and 1,000 examples. However, NYUTron's AUC consistently improved with more examples whereas lace+xgb's AUC started to plateau (from 100 to 1,000 examples, NYUTron's AUC increased by 7.27% whereas that of lace+xgb increased by 3.98%; from 10,000 to 392,336 examples, NYUTron's AUC increased by 2.15% whereas that of lace+xgb increased by 0.63%). With the full fine-tuning dataset, NYUTron had a 7.04% higher AUC than lace+xgb.

Pretraining on a large amount of unlabelled clinical notes contributes to performance. Compared with the randomly initialized LLM (random-init), NYUTron learns to generalize better from fewer examples. Figure 3b shows that, whereas NYUTron needed 10,000 examples to achieve an AUC of around 75%, random-init needed 100,000 examples. We also observed a similar trend in another clinical prediction task: NYUTron performed better than the random-init model (36.83% higher F1 score) and the non-clinically pretrained models (2.06% to 3.73% higher F1 score) on the clinical named entity recognition (NER) task from the 2012 i2b2 challenge (Extended Data Fig. 2b).

It is beneficial to match the domain of the pretraining corpus and the domain of the fine-tuning corpus. Figure 3b shows three pieces of evidence: LLMs pretrained on non-clinical text (web-wiki and web-wiki+bio) had similar performance as random-init. A separate LLM, web-wiki+bio+clinical, had similar performance as NYUTron. Third, compared with LLMs pretrained on non-clinical text (web-wiki and web-wiki+bio), clinically pretrained LLMs (NYUTron and web-wiki+bio+clinical) learned to generalize better from fewer examples. See Extended Data Fig. 3 for comparison of the pretraining corpus.

Having a close domain match during pretraining is particularly beneficial in the low-data setting during fine-tuning. We compared two language models that were pretrained on clinical text from different hospital systems, NYUTron (NYU Langone Health) and web-wiki+bio+clinical (University of Florida). Figure 3b shows that, at 1,000 examples, NYUTron (the in-domain model) had a higher AUC for NYU Langone readmission prediction than web-wiki+bio+clinical (the out-of-domain model). Notably, NYUTron's advantage disappeared as the number of fine-tuning examples increased, suggesting that sufficient in-domain fine-tuning can adapt models that were pretrained out of domain.

Clinical language models show generalizability to different sites through local fine-tuning. To investigate the robustness of NYUTron across clinical environments, we chose two hospitals that are geographically separated within the NYU Langone Health System. For brevity, we refer to Tisch Hospital in Manhattan as 'Manhattan', NYU Langone Hospital–Brooklyn as 'Brooklyn' and all four hospitals within the NYU Langone Health System (Manhattan, Brooklyn, NYU Langone Orthopedic Hospital and NYU Langone Hospital–Long Island) as 'all sites'. We considered three LLMs pretrained on different sites: the first one was pretrained in Manhattan, the second one was pretrained in Brooklyn and the third one was pretrained on all sites. For each of the pretrained LLMs, we fine-tuned the LLM with a readmission dataset from either Manhattan or Brooklyn. Finally, we asked the fine-tuned LLM to predict readmission on the basis of discharge notes from either Manhattan or Brooklyn. Figure 3c,d shows that the LLM pretrained on all sites had the best performance on both 'test Manhattan' and 'test Brooklyn'. For all the LLMs, fine-tuning with the local dataset ('fine-tune Manhattan/Brooklyn') led to a higher test AUC at the test site ('test

Manhattan/Brooklyn') compared with fine-tuning at another site ('fine-tune Brooklyn/Manhattan'). Therefore, pretraining with data from all sites and local fine-tuning is the best way to optimize performance. We performed additional analyses that showed that NYUTron is able to generalize to a different health system through local fine-tuning (Supplementary Information section 4.1 and Extended Data Fig. 4) and compared the robustness of NYUTron and lace+xgb with respect to training sites (Supplementary Information section 4.2). We also found that NYUTron is sensitive to notes from different clinical departments and patients with different demographics and that its performance fluctuates over months (Extended Data Figs. 5 and 6). The causes of the discrepancies can be very complex (discussed in Supplementary Information section 4.3) and will be studied in future work.

Prospective study of readmission

To assess NYUTron's performance outside the development environment, we selected a model on the basis of the retrospective trial results and ran a prospective trial from January to April 2022. During this time period, we deployed NYUTron in an accelerated format and loaded it into an inference engine, which interfaces with the EHR, to read discharge notes as they were signed by treating physicians. In this period, there were 29,286 discharged encounters, with 3,271 patients (11.17%) returning within 30 days. NYUTron predicted 2,692 of the 3,271 readmissions (82.30% recall) with 20.58% precision. Figure 4a shows that NYUTron had an AUC of 78.70%.

To gauge the potential clinical impact, a group of six physicians performed a qualitative evaluation of 100 randomly sampled readmitted cases that were captured by NYUTron following the trial's conclusion. Physician review suggested that some true positive predictions by NYUTron are clinically meaningful, preventable readmissions. Overall, readmitted patients who were predicted to be readmitted were 6.02 times more likely to die in hospital and stay 2.93 days longer ($P < 10^{-4}$). As shown in Fig. 4b, 61% of the predicted case were unplanned, and the mean predicted probabilities for these unplanned readmissions were lower than those for planned readmissions (31.9% ± 31.1% versus 82.1% ± 27.3%; $P < 10^{-4}$). Among the unplanned readmissions, 19.67% of patients experienced an adverse event or death on readmission, with 50% of these events considered preventable by the physician panel. From a financial standpoint, 81.9% of the unplanned readmissions would be penalized according to Centers for Medicare and Medicaid Services (CMS) guidelines. Among the penalizable cases, 54% were considered preventable. Notably, 3 of the 27 preventable readmissions had *Clostridioides difficile* enterocolitis, a contagious, healthcare-associated bacterial infection that causes 1 in 11 people over age 65 to die within 1 month[21].

Discussion

We present our work in developing, training, validating and deploying NYUTron, a health system-scale LLM designed and validated for clinical use. We demonstrate NYUTron's performance on three clinical tasks (in-patient mortality prediction, comorbidity index prediction and readmission prediction) and two operational tasks (insurance claim denial prediction and inpatient LOS prediction). We also performed a detailed analysis of readmission prediction owing to its clinical and operational importance and its well-documented history in the medical informatics literature. We view the flexibility of our approach in using an encoder architecture (BERT), which relies on only unstructured text inputs to generate a single prediction, as being a virtue, and we anticipate many future tasks built on this fundamental paradigm to assist with multiple aspects of patient care and automating hospital operations.

An ethical consideration in deployment is that physicians and administrators could over-rely on NYUTron's predictions owing to its seamless integration with existing medical workflows, thereby

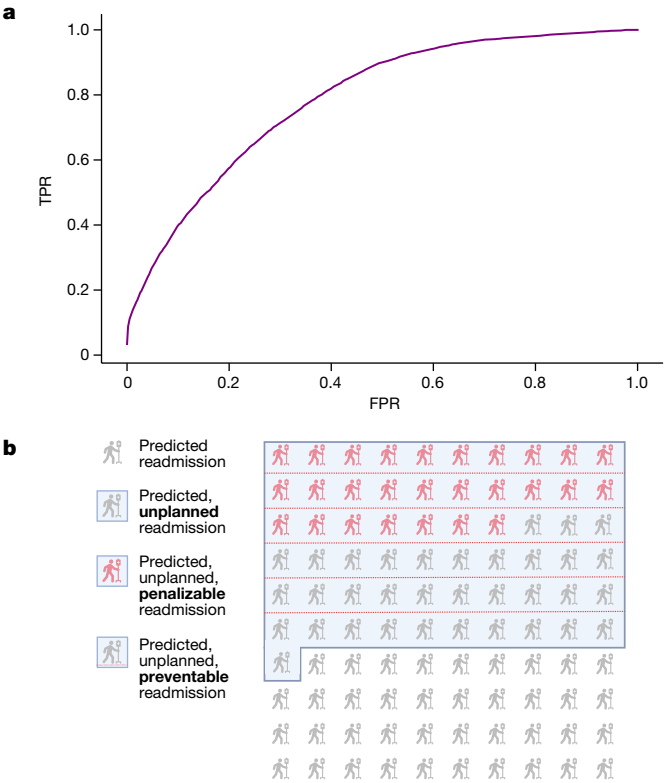


Fig. 4 | Prospective study of NYUTron's predictive performance. a, NYUTron had an AUC of 78.70% in a prospective, single-arm, non-interventional trial with recall of 82.3% and precision of 20.6%. **b**, A panel of six physicians reviewed NYUTron's results for potential clinical impact. Of 100 readmissions that were successfully identified by NYUTron, 61% were unplanned readmissions, 50% would have resulted in a penalty under CMS guidelines and 27% were preventable at the time of discharge according to the consensus opinion of the multi-specialty panel of physicians who reviewed cases from the prospective trial. See Supplementary Information section 2.1 for a discussion of the readmission label and the practical significance of the observed performance.

leading to undesirable outcomes. Further research is needed to optimize human–AI interactions, as well as development of standardized assessments for sources of bias or other unexpected failure points. Ongoing work from our group around measuring the similarity between language models' sensitivity patterns and those of physicians through token-level perturbations of the clinical notes[22] is one among many such efforts.

Large, generative LLMs also present a unique opportunity for integration into medical workflows; however, they are highly dependent on user inputs and prompting[23] and are not as easily adapted for automation of basic clinical and operational tasks. The seamless integration into existing medical informatics workflows is a virtue of our approach, and we hope that this work presents itself as a flexible solution to the last-mile problem—any structured data algorithm can be reconceptualized and rapidly prototyped within this framework. As part of monitoring the impact of such a system on physician behaviour and on patients, there should be a level of continuous supervision to capture human–machine interactions, as well as mitigate the risk of model drift over time. We discuss our implementation of such a system in Supplementary Information section 5.

Our approach of using a smaller (<1 billion parameters) encoder language model trained on highly tailored data represents a marked departure from the current trend in language model research that focuses on massive (>1 billion parameters), generative models pretrained on large, non-specific datasets. Nonetheless, even relatively small LLMs,

such as the ones used in this study, require a substantial amount of compute time for pretraining. Our pretraining used 24 NVIDIA A100 GPUs with 40 GB of VRAM for 3 weeks, and our fine-tuning used 8 A100 GPUs for 6 hours per run. This amount of computation is not commonly accessible to research groups, although we note that it is less than that in similar LLM projects routinely pursued by industry research groups and that our results indicate that massive pretraining may not be necessary to obtain highly performant models. Our results show that high-quality datasets for fine-tuning are more valuable than pretraining, and, on the basis of our experimental results, we recommend that users locally fine-tune an externally pretrained language model when computational ability is limited. Regarding the choice for the externally pretrained model, we further recommend using a model pretrained with a large amount of in-domain clinical text, although we note that large, out-of-domain models can be highly performant, particularly when combined with in-domain fine-tuning. Work with larger decoder-based architectures has also demonstrated a benefit with fine-tuning on medical data or prompt tuning with chain of thought, instructions and related techniques[24,25], which further emphasizes the necessity of accounting for the domain shift from general to medical text for LLM work in the medical sciences. Although we have not compared these approaches directly (which would require more medical text or fusion with general-domain text for training a compute-optimal model[26]), we believe that this could be an interesting future direction for research and that, in the end, approaches combining these different approaches to language modelling may prove to be complementary depending on the use case.

The ultimate validation of our approach must come from randomized controlled trials of interventions tied to individual task predictions to assess their clinical impact and from user feedback as we continue to integrate NYUTron into health systems. As we plan this within our own health system, we recommend the consideration of different levels of intervention depending on the predicted risk of patients for each task. For instance, for a patient at low risk for 30-day readmission, follow-up calls could be scheduled; for a high-risk patient, care should be taken to limit premature discharge. All interventions should be decided on with physician supervision, although many of the operational uses can probably be fully automated.

It is a long-standing dream for physicians to have AI assistants observing care along with them and chiming in with predictions and advice. To take a step towards this futuristic vision, we trained an LLM, NYUTron, on the entire EHR of a large healthcare system to read physician notes and make several of these predictions across a wide range of clinical and operational tasks. We deployed NYUTron in a live healthcare environment and demonstrate its efficacy at predicting 30-day readmission while being integrated seamlessly into clinical workflows. We believe that this work opens the door to translating the progress in modern natural language processing and deep learning to improving the quality and affordability of healthcare, and we are excited to see what comes next.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06160-y.

1. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intel.* **3**, 199–217 (2021).
2. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
3. Gaube, S. et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* **4**, 31 (2021).
4. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. in *Proc. 2019 NAACL: Human Language Technologies* (eds Burstein, J., Doran, C. & Solorio, T.) 4171–4186 (Association for Computational Linguistics, 2019).
5. Brown, T. B. et al. Language models are few-shot learners. in *Proc. NeurIPS* (eds Wallach, H. et al.) 1877–1901 (Neural Information Processing Systems, 2020).
6. Gage, B. F. et al. Selecting patients with atrial fibrillation for anticoagulation: stroke risk stratification in patients taking aspirin. *Circulation* **110**, 2287–2292 (2004).
7. Child, C. G. & Turcotte, J. G. Surgery and portal hypertension. *Major Prob. Clin. Surg.* **1**, 1–85 (1964).
8. Pugh, R. N. H., Murray-Lyon, I. M., Dawson, J. L., Pietroni, M. C. & Williams, R. Transection of the oesophagus for bleeding oesophageal varices. *Br. J. Surg.* **60**, 646–649 (2005).
9. Wells, P. et al. Accuracy of clinical assessment of deep-vein thrombosis. *Lancet* **345**, 1326–1330 (1995).
10. Tomašev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
11. Wu, N. et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE TMI* **39**, 1184–1194 (2020).
12. Liang, H. et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat. Med.* **25**, 433–438 (2019).
13. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
14. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. in *Proc. 2016 SIGKDD* 785–794 (Association for Computing Machinery, 2016).
15. Le Gall, J.-R. A. New simplified acute physiology score (SAPS II) based on a European/North American multicenter study. *J. Am. Med. Assoc.* **270**, 2957–2963 (1993).
16. Knaus, W. A., Draper, E. A., Wagner, D. P. & Zimmerman, J. E. APACHE II: a severity of disease classification system. *Crit. Care Med.* **13**, 818–829 (1985).
17. Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chron. Dis.* **40**, 373–383 (1987).
18. Caetano, N., Laureano, R. M. S. & Cortez, P. A data-driven approach to predict hospital length of stay—a Portuguese case study. in *Proc. 2014 ICEIS* (eds Hammoudi, S., Maciaszek, L. & Cordeiro, J.) 407–414 (SCITEPRESS Digital Library, 2014).
19. Johnson, M., Albizri, A. & Harfouche, A. Responsible artificial intelligence in healthcare: predicting and preventing insurance claim denials for economic and social wellbeing. *Inf. Syst. Front.* https://doi.org/10.1007/s10796-021-10137-5 (2021).
20. van Walraven, C., Wong, J. & Forster, A. J. LACE+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. *Open Med.* **6**, 80–90 (2012).
21. Center for Disease Control. What is *C. diff*? https://www.cdc.gov/cdiff/what-is.html (2022).
22. Yang, G. et al. Language model classifier aligns better with physician word sensitivity than XGBoost on readmission prediction. Preprint at https://doi.org/10.48550/arXiv.2211.07047 (2022).
23. Perez, E., Kiela, D. & Cho, K. True few-shot learning with language models. in *Proc. NeurIPS* (eds Ranzato, M. et al.) 11054–11070 (Neural Information Processing Systems, 2021).
24. Singhal, K. et al. Large language models encode clinical knowledge. Preprint at https://doi.org/10.48550/arXiv.2212.13138 (2022).
25. Bolton, E. et al. PubMedGPT 2.7B. Technical report. *Stanford University Center for Research on Foundation Models* https://crfm.stanford.edu/2022/12/15/pubmedgpt.html (2022).
26. Hoffmann, J. et al. An empirical analysis of compute-optimal large language model training. in *Proc. NeurIPS* (eds Koyejo, S. et al.) 30016–30030 (Neural Information Processing Systems, 2022).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023, corrected publication 2023

Methods

Pretraining datasets

NYU Notes. We created this dataset of unlabelled clinical notes directly from the NYU Langone EHR. The dataset contains 387,144 patients, 7,247,694 notes and 4,112,249,482 words in total. We built NYU Notes as follows: we wrote structured query language (SQL) scripts to query the NYU Langone EHR. We first prototyped the queries with an interactive web-based editor (Cloudera Hue) and then download the query results as comma-separated files (CSVs) to NYU Langone's high-performance computing cluster. We included notes signed by medical professionals (physicians, residents, physician assistants, nurse practitioners and fellows) at Tisch Hospital, NYU Langone Hospital–Brooklyn, NYU Langone Hospital–Long Island and NYU Langone Orthopedic Hospital from 2011 to 2020 (inclusive). We excluded any notes that were derived from billing, labelled as invalid or empty. We split the notes into three sets, training, validation and test sets, with a ratio of 949:50:1. Lastly, we masked tokens with 15% probability to create masked text and labels.

NYU Notes–Manhattan. We created this dataset of unlabelled clinical notes as the subset of NYU Notes that were written in Tisch Hospital in Manhattan. The dataset contains 256,217 patients, 4,342,602 notes and 2,381,466,993 words in total.

NYU Notes–Brooklyn. We created this dataset of unlabelled clinical notes as the subset of NYU Notes that were written in NYU Langone Health–Brooklyn. The dataset contains 104,521 patients, 1,337,352 notes and 1,102,078,012 words in total.

Fine-tuning datasets

NYU Readmission. We created this dataset of labelled discharge notes (with binary labels for readmission) from the NYU Langone EHR. Most of the notes from this dataset are a subset of NYU Notes, with additional discharge notes from 2021 for the temporal test. The dataset contains 413,845 patients, 506,740 notes and 487,395,462 words in total. We built this dataset as follows: for each encounter that ended between January 2011 and November 2021, we included its discharge note with a binary label for 30-day all-cause readmission. We assigned the 'readmitted' label if the patient had an admission note within 30 days of being discharged. To focus on modelling acute care readmission, we excluded discharge notes from the rehabilitation, dialysis and palliative care departments because these were not acute care admissions. We split the dataset into four sets: training, validation, test and temporal test sets. The first three sets were notes from January 2011 to May 2021, with a ratio of 8:1:1. The temporal test set included notes from June to December 2021. See Extended Data Fig. 8a for a visualization of the four-way split.

NYU Readmission–Manhattan. We created this dataset of unlabelled clinical notes as the subset of notes in the NYU Readmission dataset that were written in Tisch Hospital in Manhattan. The dataset contains 240,824 patients, 296,519 notes and 253,622,053 words.

NYU Readmission–Brooklyn. We created this dataset of unlabelled clinical notes as the subset of clinical notes from the NYU Readmission dataset that were written in NYU Langone Health–Brooklyn. The dataset contains 94,653 patients, 113,275 notes and 142,767,957 words.

NYU Mortality. We created this dataset of history and physical (H&P) notes with binary labels for in-hospital mortality from the NYU Langone EHR. Most of the notes from this dataset are a subset of NYU Notes, with additional H&P notes from 2021 for the temporal test. The dataset contains 371,922 patients, 469,162 notes and 484,467,141 words in total. We built this dataset as follows: for each encounter that ended between January 2011 and November 2021, we included its H&P note with a binary label for in-hospital mortality. We assigned the positive label if the patient's discharge disposition was 'expired'. We split the dataset into four sets: training, validation, test and temporal test sets. The first three sets were notes from January 2011 to May 2021, with a ratio of 8:1:1, and the temporal test set included notes from June to December 2021.

NYU Binned Comorbidity. We created this dataset of H&P notes with five class labels for hospital LOS from the NYU Langone EHR. Most of the notes from this dataset were a subset of NYU Notes, with additional H&P notes from 2021 for the temporal test. The dataset contains 327,039 patients, 403,579 notes and 422,485,417 words in total. The dataset contains fewer labelled encounters than the NYU Mortality and NYU Binned LOS datasets because 22% of the encounters had no International Classification of Diseases (ICD) codes to calculate the CCI score. This missingness motivated our task of predicting binned CCI score with a lack of structured ICD codes. We built this dataset as follows: for each encounter that ended between January 2011 and November 2021, we included its H&P note with a five-class label for binned CCI score. To generate the labels, we first calculated the comorbidity index using the ICD codes and the scoring function in ref. 27. We then discretized the scores into five classes: we assigned label 0 for a comorbidity index below the 50% quantile (0 days), label 1 for a comorbidity index between the 50% and 75% quantile (1–2 days), label 2 for a comorbidity index between the 75% and 90% quantile (3–4 days), label 3 for a comorbidity index between the 90% and 99% quantile (4–7 days) and label 4 for a comorbidity index above the 99% quantile (>7 days). We split the dataset into four sets: training, validation, test and temporal test sets. The first three sets were notes from January 2011 to May 2021, with a ratio of 8:1:1, and the temporal test set included notes from June to December 2021.

NYU Binned LOS. We created this dataset of H&P notes with quantile labels for hospital LOS from the NYU Langone EHR. Most of the notes from this dataset were a subset of NYU Notes, with additional H&P notes from 2021 for the temporal test. The dataset contains 371,922 patients, 469,162 notes and 484,467,141 words in total. We built this dataset as follows: for each encounter that ended between January 2011 and November 2021, we included its H&P note with a binary label and a quantile label for LOS. For the quantile label, we assigned label 0 for an LOS below the 25% quantile (0–2 days), label 1 for an LOS between the 25% and 50% quantile (3 days), label 2 for an LOS between the 50% and 75% quantile (4–5 days) and label 3 for an LOS above the 75% quantile (>5 days). We split the dataset into four sets: training, validation, test and temporal test sets. The first three sets were notes from January 2011 to May 2021, with a ratio of 8:1:1, and the temporal test set included notes from June to December 2021.

NYU Insurance Denial. We created this dataset of H&P notes with binary labels for whether the patient's insurance claim was initially rejected or directly approved. The dataset contains 54,563 patients, 55,791 notes and 51,270,256 words in total. We built this dataset as follows: for each encounter that occurred between May 1, 2021, and April 30, 2022, we included its H&P note with a binary label for insurance denial. We assigned a positive label if the patient's insurance claim status was 'final, adverse determination' (claim was rejected by insurance and was again rejected following appeal) or 'final, favorable determination' (claim was rejected by insurance and approved following appeal). We split the dataset into four sets: training, validation, test and temporal test sets. The first three sets were notes from May 1, 2021, to February 30, 2022, with a ratio of 18:1:1. The temporal test set included notes from March 1 to April 30, 2022.

NYU Insurance Denial–Discharge Notes. We created this dataset of discharge notes with binary labels for whether the patient's insurance

claim was initially rejected or directly approved. The dataset contains 54,563 patients, 55,791 notes and 49,405,133 words in total. We built this dataset as follows: for each encounter that occurred between May 1, 2021, and April 30, 2022, we included its discharge note with a binary label for insurance denial. The label assignment and four-way split were the same as in the NYU Insurance Denial dataset.

NYU Insurance Eventual Denial, H&P. This dataset contained the same notes as the NYU Insurance Denial dataset, but the labels were different. The binary label indicated whether the patient's insurance claim was eventually rejected (even after appeal) or was eventually approved (direct approval or approval after appeal).

NYU Insurance Eventual Denial–Discharge. This dataset contained the same notes as the NYU Insurance Denial–Discharge Notes dataset, but the labels were different. The binary label indicated whether the patient's insurance claim was eventually rejected (even after appeal) or was eventually approved (direct approval or approval after appeal).

i2b2-2012 NER. This is an open dataset released by the Harvard Medical School as part of an annual clinical NLP challenge[28]. This dataset is a well-known benchmark in the clinical NLP community. The task is to identify and classify clinical concepts (for example, treatments), clinical departments (for example, surgery), occurrences of events (for example, admission) and evidentials (for example, the patient complained) from de-identified clinical notes from Beth Israel Medical Center in Boston. The dataset contains no more than 310 patients, 310 notes and 636,000 words. We downloaded the dataset as a compressed tar.gz file from the n2c2 data portal after our use application was approved.

MIMIC-III Readmission. This is an open dataset for an intensive care unit (ICU) EHR released by MIT and Boston Beth Israel Medical Center[29]. We collected a set of 52,726 discharge notes and created a 30-day all-cause readmission label by checking whether there was any subsequent encounter within 30 days. The readmission rate was 6%. We split the data into training, validation and test sets in a 8:1:1 ratio.

Deployment dataset

NYU Readmission–Deployment. This dataset consists of discharge notes with binary labels for readmission from our deployment engine and the NYU Langone EHR. From January to April 2022, every time a discharge note was signed by a physician, the note was sent to our custom inference engine for NYUTron's prediction. The paired discharge note and prediction were recorded in a database. The database contained 27,376 patients, 29,287 notes and 34,669,963 words by the end of the study period.

Structured datasets

NYU Readmission–LACE. We created this dataset of structured LACE[30] features with binary labels for readmission for comparison against the unstructured models. The dataset contains structured features for all encounters in the NYU readmission dataset. LACE is a traditional clinical prediction rule for readmission with four features: LOS, acuity of readmission, Charlson comorbidity index, and number of recent emergency department visits in the past 6 months. We built the dataset as follows: for every encounter in the NYU Readmission dataset, we collected data on the four LACE features from the NYU Langone EHR. LOS was the difference (in days) between the discharge date and the admission date. Acuity of readmission was a binary feature indicating whether the patient was admitted to the emergency department. The comorbidity index was calculated with the ICD-9 or ICD-10 codes for chronic diseases, on the basis of the mapping algorithm in ref. 31 and the scoring function in ref. 27. The number of emergency department visits was calculated from the patient's encounter history up to 6 months before the admission date.

NYU Readmission–LACE, Manhattan. We created this dataset of structured LACE features from the subset of notes from the NYU Readmission–LACE dataset that were written in Tisch Hospital in Manhattan.

NYU Readmission–LACE, Brooklyn. We created this dataset of structured LACE features from the subset of notes from the NYU Readmission–LACE dataset that were written in NYU Langone Health–Brooklyn.

NYU Mortality–SAPS2 + APACHE2. We created this dataset of structured SAPS2 + APACHE2 features with binary labels for in-hospital mortality to compare against the unstructured data. The dataset contains a subset of structured SAPS2 + APACHE2 features for all encounters in the NYU Mortality dataset. SAPS2 + APACHE2 features are a subset of the features used in the SAPS2 model[15] and the APACHE2 model[16] for ICU mortality prediction. We selected the subset of features that were available in the NYU Langone EHR. We included the following 12 features: age (numerical), mean heart rate (numerical), systolic blood pressure (numerical), atrial temperature (numerical), blood urea nitrogen concentration (numerical), sodium concentration (numerical), potassium concentration (numerical), bilirubin concentration (numerical), white blood cell count (numerical), pH (numerical), creatine concentration (numerical) and haematocrit (numerical). We additionally included department specialty (categorical). We excluded the following features owing to their unavailability: PaO_2/FiO_2 (ratio of arterial oxygen partial pressure to fractional inspired oxygen), whether the patient was on mechanical ventilation or continuous positive airway pressure (CPAP), bicarbonate concentration, urine output, Glasgow Coma Scale score, presence of metastatic cancer or haematological malignancy or AIDS, and whether the admission was scheduled.

NYU Binned LOS–Lisbon Portugal. We created this dataset of structured 'Lisbon Portugal' features with binary labels for in-hospital mortality to compare against the unstructured data model. The dataset contains a subset of the features used in the Lisbon Portugal dataset[18] (which is widely used in the LOS prediction literature) for all encounters in the NYU Binned LOS dataset. We selected a subset of 12 features that were available in the NYU Langone EHR: gender (categorical), age as measured by the difference in years between the birth date and the admission date (numerical), highest level of education (categorical), country (categorical), postal code as address (categorical), marital status (categorical), admission type (categorical), admission service type (categorical), provider ID (categorical), department specialty (categorical), procedure name (categorical) and number of previous admissions (numerical). We left out diagnosis because it is not always available at the time of writing H&P notes. We excluded the following three features owing to difficulty in finding them in the NYU Langone EHR: homogeneous group diagnosis code, great diagnostic category and treatment.

NYU Insurance Denial–Claim Forms. We created this structured dataset based on the NYU Insurance Denial dataset for comparison against the unstructured data model. The dataset contains structured features for all encounters in the NYU Insurance Denial dataset and has the same splits as the NYU Insurance Denial dataset. Selection of structured features was based on the features in ref. 19, which built a model that predicts insurance claim denial from demographic and care-related features found in the claim form. We found eight available features in the NYU Langone EHR: patient name (categorical), age (numerical), gender (categorical), postal code as a generalization of address (categorical), insurance brand (categorical), first insurance plan name (categorical), provider ID (categorical) and provider type (categorical). We additionally added four features based on the clinician's inputs: second insurance plan code (categorical), a binary flag for surgical cases (categorical), a binary flag for emergency department cases (categorical) and a binary flag for Medicare fee-for-service users

(categorical). We left out six features in ref. 19 owing to difficulty in searching for them: the patient's relationship to the insured person, network type, whether the claim was a resubmission, diagnosis pointer, charge of service and prior authorization number.

Preprocessing

Pretraining datasets (NYU Notes, NYU Notes–Manhattan, NYU Notes–Brooklyn). Using these datasets, we trained an uncased BERT wordpiece tokenizer with a vocabulary size of 50,000 tokens, a maximum sequence length of 512 tokens and special tokens [SEP], [PAD], [UNK], [MASK] and [CLS]. Because most of the clinical notes had more than 512 tokens, we split each long note into non-overlapping chunks that were under the maximum sequence length. Specifically, we split each note into sentences using natural language toolkit (nltk)[32] and tokenized each sentence. For sentences that were longer than 512 tokens, we truncated them. Next, for all tokenized sentences in the same note, we concatenated them into groups such that each group had exactly the maximum sequence length. We discarded any remaining group (with a length strictly less than the maximum) of a long note.

Fine-tuning datasets (NYU Readmission, NYU Readmission–Manhattan, NYU Readmission–Brooklyn, NYU Mortality, NYU Binned LOS, NYU Insurance Denial, NYU Binned Comorbidity). Using the tokenizer trained with NYU Notes, we first tokenized the discharge note. We truncated notes that exceeded the maximum sequence length of 512 tokens. We leave it for the future to design a language model that efficiently reads longer clinical notes (see Extended Data Fig. 8b for the impact of note length on language model performance).

i2b2-2012 NER. We first decompressed the tar.gz files into folders of xml files. We then converted the xml files to brat format. Next, we converted the brat files to bio files. Finally, we wrote a custom HuggingFace[33] data loader to convert the folder of bio files into a HuggingFace dataset. Our code for preprocessing is available at GitHub.

Deployment datasets. We first cleaned the notes by stripping out html artifacts. We then tokenized the discharge note using NYUTron's tokenizer. We truncated notes that exceeded the maximum sequence length of 512 tokens.

Structured dataset (NYU Readmission–LACE, NYU Mortality–SAPS2+APACHE2, NYU Binned LOS–Lisbon Portugal, NYU Insurance Denial–Claim Forms). When there was a missing numerical feature (for example, the average heart rate was NaN), we filled in the feature as the average feature across the training set. For missing categorical features (for example, the admitting department was 'unspecified'), we left them as category 'none'.

Pretraining

We pretrained a 109 million-parameter BERT model using preprocessed NYU Notes and the MLM objective for 3 weeks (96 epochs) on 24 NVIDIA A100 GPUs distributed over three compute nodes until the validation loss started to plateau. The model has 12 hidden layers with dimension 768, with 12 attention heads per layer. We used a per-device training batch size of 64 and saved every 2,000 steps. We used the Zero Redundancy AdamW optimizer (an improvement over the Adam optimizer) with a constant learning rate of 5×10^{-5}, FP16 mixed precision and stage 2 parallelization[34–36].

Fine-tuning

NYUTron + discharge notes for readmission prediction. We replaced the trained MLM classifier with a randomly initialized linear classifier after the last hidden layer of the pretrained BERT model. We fine-tuned the model end to end using the training set of the NYU Readmission dataset for ten epochs, evaluating the validation AUC every half epoch and stopping early with a patience of five. We used the following hyperparameters from manual tuning based on the validation AUC: a learning rate of 2×10^{-5}, a weight decay of 0.01 and a per-device batch size of 4. We optimized the cross-entropy loss using the AdamW optimizer. While varying the size of the dataset ($N \in \{10^2, 10^3, 10^4, 10^5, 3.92336 \times 10^5\}$), we fine-tuned the pretrained model using subsamples of the NYU Readmission dataset and evaluated their AUC on the temporal test set. For each size of subsample, we ran five experiments with distinct random seeds (0, 13, 24, 36, 42). For comparison, we looked at the median AUC and the standard deviation of the five experiments.

NYUTron + H&P notes for in-hospital mortality prediction. We replaced the trained MLM classifier with a randomly initialized linear classifier after the last hidden layer of the pretrained BERT model. We fine-tuned the model end to end using the training set of the NYU Mortality dataset for ten epochs, evaluating the validation AUC every half epoch and stopping early with a patience of 5. We used the following hyperparameters from manual tuning based on the validation AUC: a learning rate of 2×10^{-5}, a weight decay of 0.01 and a per-device batch size of 4. We optimized the cross-entropy loss using the AdamW optimizer. Using the full dataset, we fine-tuned the pretrained model using subsamples of the NYU Mortality dataset and evaluated their AUC on the temporal test set. For each size of subsample, we ran five experiments with distinct random seeds (0, 13, 24, 36, 42). For comparison, we looked at the median AUC and the standard deviation of the five experiments.

NYUTron + H&P notes for binned comorbidity prediction. We replaced the trained MLM classifier with a randomly initialized linear classifier after the last hidden layer of the pretrained BERT model. We fine-tuned the model end to end using the training set of the NYU Binned Comorbidity dataset for ten epochs, evaluating the validation OVR AUC every half epoch and stopping early with a patience of 5. We used the following hyperparameters from manual tuning based on the validation OVR AUC: a learning rate of 2×10^{-5}, a weight decay of 0.01 and a per-device batch size of 4. We optimized the cross-entropy loss using the AdamW optimizer. Using the full dataset, we fine-tuned the pretrained model with subsamples of the NYU Binned Comorbidity dataset and evaluated their OVR AUC on the temporal test set. For each size of subsample, we ran five experiments with distinct random seeds (0, 13, 24, 36, 42). For comparison, we looked at the median OVR AUC and the standard deviation of the five experiments.

NYUTron + H&P notes for binned LOS prediction. We replaced the trained MLM classifier with a randomly initialized linear classifier after the last hidden layer of the pretrained BERT model. We fine-tuned the model end to end using the training set of the NYU Binned LOS dataset for ten epochs, evaluating the validation AUC every half epoch and stopping early with a patience of 5. We used the following hyperparameters from manual tuning based on the validation OVR AUC: a learning rate of 2×10^{-5}, a weight decay of 0.01 and a per-device batch size of 4. We optimized the cross-entropy loss using the AdamW optimizer. Using the full dataset, we fine-tuned the pretrained model with subsamples of the NYU Binned LOS dataset and evaluated their AUC on the temporal test set. For each size of subsample, we ran five experiments with distinct random seeds (0, 13, 24, 36, 42). For inference, we combined the last two classes, label 3 (90–99% quantile) and label 4 (>99% quantile) because label 4 was very sparse. For comparison, we looked at the median OVR AUC and the standard deviation of the five experiments.

NYUTron + H&P notes for insurance denial prediction. We replaced the trained MLM classifier with a randomly initialized linear classifier after the last hidden layer of the pretrained BERT model. We fine-tuned the model end to end using the training set of the NYU Insurance

Denial dataset for ten epochs, evaluating the validation AUC every half epoch and stopping early with a patience of 5. We used the following hyperparameters from manual tuning based on the validation AUC: a learning rate of 2×10^{-5}, a weight decay of 0.01 and a per-device batch size of 4. We optimized the cross-entropy loss using the AdamW optimizer. Using the full dataset, we fine-tuned the pretrained model using subsamples of the NYU Insurance Denial dataset and evaluated their AUC on the temporal test set. For each size of subsample, we ran five experiments with distinct random seeds (0, 13, 24, 36, 42). For comparison, we looked at the median AUC and the standard deviation of the five experiments.

NYUTron + clinical notes for NER. We performed the fine-tuning experiments as follows. For each LLM in Extended Data Table 2, we initialized a HuggingFace token classification model with the LLM as the pretrained checkpoint. We fine-tuned the model using i2b2-2012 NER for ten epochs using the AdamW optimizer with a learning rate of 2×10^{-5}, a weight decay of 0.01 and a batch size of 4, evaluating every 50 steps and stopping early on the basis of area under the receiver operating characteristic (AUROC) with a patience of 1. This took 20 to 40 min on one node of four NVIDIA 17-GB V100 GPUs. We performed fine-tuning five times with random seeds 0, 13, 24, 36 and 42 and recorded the average and standard deviation of the micro-averaged F1 score (excluding the label for non-entity, 'O').

NYUTron + MIMIC-III readmission. We performed the fine-tuning experiments as follows: For both NYUTron and BioClinicalBert, we initialized a HuggingFace token classification model with the LLM as the pretrained checkpoint. We fine-tuned the model using MIMIC-III Readmission for ten epoch using the AdamW optimizer with a learning rate of 2×10^{-5}, a weight decay of 0.01 and a batch size of 16, evaluating every half epoch. We performed fine-tuning five times with random seeds 0, 13, 24, 36 and 42.

Deployment

The fine-tuned model was converted to a high-performance format (Onnx or TensorRT) and loaded into our deployment platform, an NVIDIA Triton inference engine that interfaces with the NYU Langone EHR through the HLA7 Fast Health Interoperability Resources (FHIR)[37] interface. For our consideration of performance, security, reliability and interpretability, see Supplementary Information section 5.

Our deployment platform consisted of a modified version of NVIDIA's Triton Inference Server that we named NYUTriton (pronounced 'nutrition' because it is good for the health system). NVIDIA Triton supports GPU-, x86- and ARM CPU-based inferencing and several key features, including dynamic batching, concurrent execution, a highly flexible model specification interface, and the ability to support a wide range of deep learning frameworks and accelerated model formats for maximum throughput. We modified NVIDIA Triton to interface seamlessly with HuggingFace-formatted language models so as to provide a uniform and highly flexible crossover point between our development and production pipelines. Trained models were saved in a standard HuggingFace-style format and converted into Onnx and then TensorRT to obtain sub-millisecond-scale inference results. NYUTriton is hosted on a dedicated inference server that consists of an AMD Threadripper 3960X (24 cores, 3.8 GHz), two RTX 3090 GPUs and 128 GB of DDR5 system memory purchased from Lambda Labs.

Following the signing of discharge summaries in Epic, the HL7 FHIR interface connects with NYUTriton and sends a JavaScript Object Notation (JSON) payload consisting of the discharge summary and metadata specifying the underlying readmission model and sender. NYUTriton preprocesses the text, runs an inference job with the accelerated NYUTron readmission model and returns the model's inference result to a secondary orchestration server, which writes the result to a database and generates an email to the signing physician.

Structured baselines

The structured baselines were (1) SAPS2/APACHE2 features + XGBoost for in-hospital mortality prediction, (2) LACE features + XGBoost for readmission prediction, (3) Lisbon Portugal features + XGBoost for binned LOS prediction and (4) claim form features + XGBoost for insurance denial prediction.

For all structured baselines, we used the xgboost library to train an extreme gradient-boosted tree classifier with a binary logistic loss (multiclass softmax loss for more than two classes). We used scikit-learn's randomized search to search hyperparameters among minimum_child_weight from {1, 5, 10}, gamma from {0.5, 1, 1.5, 2, 5}, subsample from {0.6, 0.8, 1}, col_sample_bytree from {0.6, 0.8, 1.0}, max_depth from {3, 4, 5}, learning_rates from {0.001, 0.01, 0.1, 0.5} and n_estimators from {10, 100, 1000} for 100 iterations based on AUROC score (ovr-auroc score for multiple classes) from threefold cross-validation[38]. We ran each experiment five times with distinct random seeds (0, 13, 24, 36, 42). For mortality, binned comorbidity, binned LOS and insurance denial, we ran the experiment with the full dataset. For readmission, we trained the model using subsamples ($N \in \{10^2, 10^3, 10^4, 10^5, 3.92336 \times 10^5\}$) of the NYU Readmission–LACE dataset.

Metrics

We evaluated the five tasks (in-hospital mortality prediction, binned comorbidity index prediction, 30-day all-cause readmission prediction, binned LOS prediction and insurance denial prediction) with AUC for binary classes and OVR AUROC for multiple classes. AUROC is the area under the two-dimensional curve consisting of tuples of the form (TPR, FPR) resulting from different decision thresholds.

We additionally evaluated readmission prediction with the following metrics: TPR, FPR, precision, recall and F1 score, all of which have a range of [0, 1]. We evaluated NER using a micro-averaged NER F1 score. The NER F1 score is similar to the normal F1 score except that the non-entity label 'O' is excluded for calculation.

Baseline algorithms for retrospective study

We compared NYUTron against physicians. We worked with six physicians with different levels of seniority: three attending physicians and three residents. The physicians were asked to review discharge summaries and predict whether the described patient would come back to the hospital within 30 days.

We compared NYUTron against four other LLMs and two machine learning models. 'random-init' is a BERT-base uncased model with randomly initialized parameters. 'web-wiki' is a BERT-base uncased model that is pretrained using web text (from the BookCorpus dataset[39]) and Wikipedia articles (from the English Wikipedia dataset[40]). 'web-wiki+bio' is a BERT model pretrained using web text, Wikipedia articles, PubMed abstracts[41] and PubMed Central (PMC) full articles[42]. 'web-wiki+bio+clinical', or gatortron-og[43], is a Megatron-BERT[44] model pretrained using web text, Wikipedia articles, PubMed abstracts, PMC full articles, MIMIC-III notes and de-identified clinical notes from University of Florida Health. 'lace+xgb' reads structured LACE features (from a traditional clinical prediction rule) with an extreme gradient-boosted tree model[14]. 'tf-idf+xgb' reads corpus-level bag-of-words features with an extreme gradient-boosted tree model. For detailed statistics and examples of the pretraining corpora, see Extended Data Table 2 and Extended Data Fig. 3.

Comparison with physicians

We randomly sampled 20 discharge notes from the random test set and asked six doctors with different seniority to predict whether the patient would come back within 30 days. The six physicians included three attending neurosurgeons, two neurosurgery residents and one ICU resident.

We used REDCap to perform the survey and gave physicians unlimited time. The survey was structured as follows: for each case, we

asked "Will this person be admitted within 30 days?", followed by the discharge summary. The physician could choose to answer "yes" or "no". If the patient came back within 30 days, we had three follow-up questions to assess the characteristics of the subsequent readmission. First, we asked "Is this readmission related to the prior discharge?", followed by the H&P note of the subsequent readmission. The physician could answer "yes", "no", "partial" or "does not meet Medicare criteria for 30-day readmission". The second follow-up question was "Is this readmission preventable?", to which the physician could answer "yes", "no" or "partial". The third follow-up question, "Any comments?", had a free-text response where the physician could explain why the readmission was partially related to the prior discharge or why the readmission was partially preventable.

To collect NYUTron's predictions, we used the text classification pipeline from HuggingFace to perform inference on the 20 discharge notes. For each discharge note, the pipeline output a predicted probability for readmission. We converted this predicted probability to a binary label with a threshold of 0.07 (a predicted probability no less than 0.07 was converted to a positive label). We chose 0.07 as the decision boundary because it was the minimum threshold that gave us above 80% validation recall among the thresholds $\{0.01 \times n : n \in \{1, ..., 90\}$ (the 80% criterion was chosen on the basis of clinical applicability). See Extended Data Fig. 8c for NYUTron's calibration curve.

Comparison with other language models

Discharge notes + other LLMs for readmission prediction. The dataset, hyperparameters, and evaluation and software libraries for fine-tuning other LLMs were the same as when fine-tuning NYUTron. The pretrained LLMs were constructed as follows: random-init is a BERT-base uncased model with reset parameters. web-wiki is a BERT-base uncased model. web-wiki+bio is a dmis-lab/biobert-base cased v1.2 model. web-wiki+bio+clinical was Gatortron-og downloaded from NVIDIA NGC and converted to a HuggingFace checkpoint using convert megatron bert checkpoint.

Clinical notes + other LLMs for NER. The dataset, hyperparameters, and evaluation and software libraries for fine-tuning other LLMs were the same as for fine-tuning NYUTron. The pretrained LLMs were the same as the baseline LLMs for predicting readmission from discharge notes.

Comparison with machine learning models

LACE features + XGBoost for readmission prediction. Using the NYU Readmission–LACE dataset, we used the xgboost library to train an extreme gradient-boosted tree classifier with binary logistic loss with hyperparameter search. We used scikit-learn's randomized search to search among minimum_child_weight from {1, 5, 10}, gamma from {0.5, 1, 1.5, 2, 5}, subsample from {0.6, 0.8, 1}, col_sample_bytree from {0.6, 0.8, 1.0}, max_depth from {3, 4, 5}, learning_rates from {0.001, 0.01, 0.1, 0.5} and n_estimators from {10, 100, 1000} for 100 iterations on the basis of AUROC score on the validation set[37]. We trained the model using subsamples ($N \in \{10^2, 10^3, 10^4, 10^5, 3.92336 \times 10^5\}$) of the NYU Readmission–LACE dataset and evaluated their AUROC on the temporal test set. For each size of subsample, we ran five experiments with distinct random seeds (0, 13, 24, 36, 42). For comparison, we looked at the median AUROC and the standard deviation of the five experiments.

XGBoost + TF-IDF for readmission prediction. We transformed the text from the NYU Readmission dataset into tf-idf (term frequency–inverse document frequency) embeddings and used an xgboost classifier with binary logistic loss to predict readmission. We used raytune[45] to search hyperparameters, including max_tf-idf features from {512, 5000}, max_depth from a quantized random integer of 3 to 16 with an interval of 4, learning_rate from a log uniform distribution from 10^{-2} to 10^{-1}, gamma from a quantized uniform distribution from 0 to 12 with an interval of 4, minimum_child_weight from a quantized uniform distribution from 0 to 8 with an interval of 4, reg lambda from a quantized uniform distribution from 0 to 10 with an interval of 2, colsample_bytree from a uniform distribution from 0.7 to 1, scale pos weight from a quantized uniform distribution from 0 to 50 with an interval of 10 and n_estimator from a quantized integer distribution from 50 to 300 with an interval of 50. We trained the model using subsamples ($N \in \{10^2, 10^3, 10^4, 10^5, 3.92336 \times 10^5\}$) of the NYU Readmission dataset and evaluated their AUROC on the temporal test set. For each size of subsample, we ran five experiments with distinct random seeds (0, 13, 24, 36, 42). For comparison, we looked at the median AUROC and the standard deviation of the five experiments.

Comparison of multi-site pretraining and fine-tuning

We compared NYUTron with its four variants (pretrained and fine-tuned using data from different sites): (1) NYU Notes–Manhattan + NYU Readmission–Manhattan, (2) NYU Notes–Manhattan + NYU Readmission–Brooklyn, (3) NYU Notes–Brooklyn + NYU Readmission–Brooklyn and (4) NYU Notes–Brooklyn + NYU Readmission–Manhattan. The hyperparameters and evaluation and software libraries for fine-tuning NYUTron variants were the same as for fine-tuning NYUTron.

Analysis of prospective performance

On the basis of the temporal test performance in the retrospective study, we selected a fine-tuned model with a decision threshold of 0.07 for use in the prospective trial.

Comparison of mortality rate and LOS. To assess the condition of the readmitted patients who were correctly predicted (n = 3,298), we compared their in-hospital mortality rate and length of hospitalization with that of patients who were admitted in the same period. We collected data on patients who were admitted from February to May 2022 (n = 30,548) and compared their in-hospital mortality rate and LOS with that of the readmitted patients caught by NYUTron from January to April 2022. We used two-sided Welch's t tests (with the null hypothesis that the two groups had the same average) to assess the statistical significance of our comparison[46].

Assessing NYUTron's clinical impacts with physician review. We performed a post hoc analysis of readmitted patients in the prospective cohort to better understand model performance in a real-world environment and in anticipation of creating targeted interventions based on model outputs. One hundred readmitted patients were sampled from the five largest departments at NYU Langone by patient volume: internal medicine, pediatrics, general surgery, obstetrics and gynaecology, and haematology and oncology. Each department contributed 20 cases, with 10 cases having the highest predicted probabilities in that department and 10 cases having the lowest predicted probabilities. All cases had their encounter IDs logged for their index discharge and readmission on a secure online platform. A standardized questionnaire was constructed for manual review asking whether the readmission was planned, whether the readmission met CMS criteria for a penalized 30-day readmission, whether the readmission was preventable, whether an adverse event occurred on readmission, whether any adverse events were preventable and whether the reviewing physicians had any comments on the case. A team of ten physicians from internal medicine and neurosurgery were randomly assigned cases to be reviewed in pairs, with any disagreement between the reviewers adjudicated by a third physician reviewer. To determine whether a readmission was preventable, the reviewer looked at the discharge note of the inference encounter and the H&P note of the readmission encounter.

Ethical approval

Our research was approved by the NYU Langone institutional review board as 's21-01189 NYUtron', and the methods were carried out in accordance with the institutional review board's relevant guidelines and regulations.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The clinical data used for the pretraining, fine-tuning, validation and test sets were collected from the NYU Langone Health System EHR maintained by the NYULH Datacore team. Text data were stripped of rich-text features and directly included in the dataset 'as is' and were augmented with structured features where noted. These data consist of the production medical records of NYU Langone and cannot be made publicly available. Researchers may obtain a limited de-identified dataset (or a test subset) from NYU Langone Health System by reasonable request and subject to local and national ethical approvals. We also used publicly available i2b2-2012 (https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/) and MIMIC-III (https://physionet.org/content/mimiciii/1.4/) datasets.

Code availability

We used sql and Python 3.8.13 to collect data from the NYU Langone EHR. We used REDCap 12.4.31 to collect physician responses. This work used several open-source libraries, including HuggingFace Transformers 4.19.2, Datasets 2.2.2, Evaluate 0.1.1, wandb 0.12.17, matplotlib 3.5.2, seaborn 0.12.2, pandas 1.4.2, ray 2.0.0, sklearn 1.1.1, deepspeed 0.8.0+384f17b, NVIDIA Apex, XGBoost 1.6.1 and nltk 3.6.3. Our experimental framework involved the use of these libraries and, in some cases, modification of them. We will release code to replicate the pretraining, fine-tuning and testing of the models described in this paper at the time of publication (code for experiments available at https://github.com/nyuolab/NYUTron, preprocessing code for i2b2-2012 available at https://github.com/nyuolab/i2b2_2012_preprocessing). We include detailed methods and implementation steps in the Methods and Supplementary Information to allow for independent replication.

27. Charlson, M. Charlson comorbidity index (CCI). *MD+CALC* https://www.mdcalc.com/calc/3917/charlson-comorbidity-index-cci (2022).
28. Sun, W., Rumshisky, A., & Uzuner, O. Annotating temporal information in clinical narratives. *J. Biomed. Inform.* **46**, 5–12 (2013).
29. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
30. van Walraven, C. et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Can. Med. Assoc. J.* **182**, 551–557 (2010).
31. Sundararajan, V. et al. New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality. *J. Clin. Epidemiol.* **57**, 1288–1294 (2004).
32. Bird, S. & Loper, E. NLTK: The Natural Language Toolkit. in *Proc. 2004 ACL Interactive Poster and Demonstration Sessions* 214–217 (Association for Computational Linguistics, 2004).
33. Wolf, T. et al. Transformers: state-of-the-art natural language processing. in *Proc. 2020 EMNLP* (eds Webber, B., Cohn, T., He, Y. & Liu, Y.) 38–45 (Association for Computational Linguistics, 2020).
34. Rajbhandari, S., Rasley, J., Ruwase, O. & He, Y. ZeRO: memory optimizations. Toward training trillion parameter models. in *Proc. Int. Conf. High Performance Computing, Networking, Storage and Analysis* 1–16 (IEEE Press, 2020).
35. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *ICLR* https://openreview.net/forum?id=Bkg6RiCqY7 (2019).
36. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *ICLR* https://arxiv.org/abs/1412.6980 (2017).
37. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R. & Stiawan, D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. *JMIR Med. Inform.* **9**, 21929 (2021).
38. Pedregosa, F. et al. Scikit-Learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
39. Zhu, Y. et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books. in *Proc. 2015 ICCV* (ed. O'Conner, L.) 19–27 (IEEE, 2015).
40. Wikimedia Foundation. Wikimedia downloads. https://dumps.wikimedia.org/ (2021).
41. NCBI Literature Resources. Download PubMed data. https://pubmed.ncbi.nlm.nih.gov/download/ (2022).
42. National Library of Medicine. PubMed Central: PMC article datasets. https://www.ncbi.nlm.nih.gov/pmc/tools/textmining/ (2022).
43. Yang, X. et al. A large language model for electronic health records. *NPJ Digit. Med.* **5**, 194 (2022).
44. Shoeybi, M. et al. Megatron-LM: training multi-billion parameter language models using model parallelism. Preprint at https://doi.org/10.48550/arXiv.1909.08053 (2020).
45. Liaw, R. et al. Tune: a research platform for distributed model selection and training. Preprint at https://doi.org/10.48550/arXiv.1807.05118 (2018).
46. Welch, B. L. The generalization of Student's problem when several different population variances are involved. *Biometrika* **34**, 28–35 (1947).

Acknowledgements E.K.O. is supported by the National Cancer Institute's Early Surgeon Scientist Program (3P30CA016087-41S1) and the W.M. Keck Foundation. We would like to acknowledge J. Golfinos, whose vision and support made this project possible. We also would like to acknowledge our collaborators M. Costantino and K. Yie from the NYU Langone High-Performance Computing (HPC) team; without their tireless assistance in building and maintaining our GPU cluster, none of this research would have been possible. We would also like to thank D. Bar-Sagi and N. Mherabi, whose support for this research has made everything possible. We would like to thank B. Guzman from the NYU Langone Predictive Analytics Unit and V.J. Major from the NYU Grossman School of Medicine for their help with learning the SQL data structures used as part of this work. We would like to thank Y.(R.) Pang for reviewing and editing the initial manuscript. We would like to thank X. Yang from University of Florida for helping us with preprocessing and evaluating the i2b2 dataset. We thank S. Ciprut for helping with the REDCap survey and research administration for our team. We thank C. Fernandez-Granda, J. Kempe, V. Dhar, N. Wu, M. Barot, A. Chen, K. Link and F. Kwon for their valuable discussions.

Author contributions E.K.O. conceptualized and supervised the project. L.Y.J. collected data (except the NYU Insurance Denial and MIMIC-III Readmission datasets) and performed experiments. L.Y.J. and X.C.L. prepared the figures. X.C.L., N.P.N., M.N.-M. and K.C. debugged and tested the model and the pretraining and fine-tuning software. E.K.O. designed the NYUTriton deployment platform, and E.K.O., A.A. and D.W. built the system and integrated it with the EHR. K.E., E.K.O., D.W. and Y.A. collected and processed the NYU Insurance Denial dataset. H.A.R., I.L., P.P., K.E., M.M., N.C.K., C.O., Z.S., C.L., H.W., D.K., S.N., Y.D., D.K. and A.T.M.C. participated in the human experiments, review of cases, and providing user feedback and testing. G.Y. and M.C. provided the scripts for tf-idf+xgb and built the MIMIC-III Readmission dataset. M.F., A.B.C., Y.A. and K.C. provided guidance and feedback throughout the project. L.Y.J., K.C. and E.K.O. wrote the initial draft. L.Y.J., E.K.O., K.C., M.N.-M., G.Y. and M.C. formatted the final submission. All authors edited and revised the manuscript.

Competing interests E.K.O. reports consulting with Sofinnova and Google, income from Merck & Co. and Mirati Therapeutics, and equity in Artisight. N.P.N., M.F. and A.B.C. are employed by NVIDIA. D.K. reports consulting with Elekta. K.C. is employed by Prescient Design, a Genentech accelerator, a subsidiary of Roche. There are no other potential conflicts of interest. The work presented herein was performed exclusively within the NYU Langone Health System.

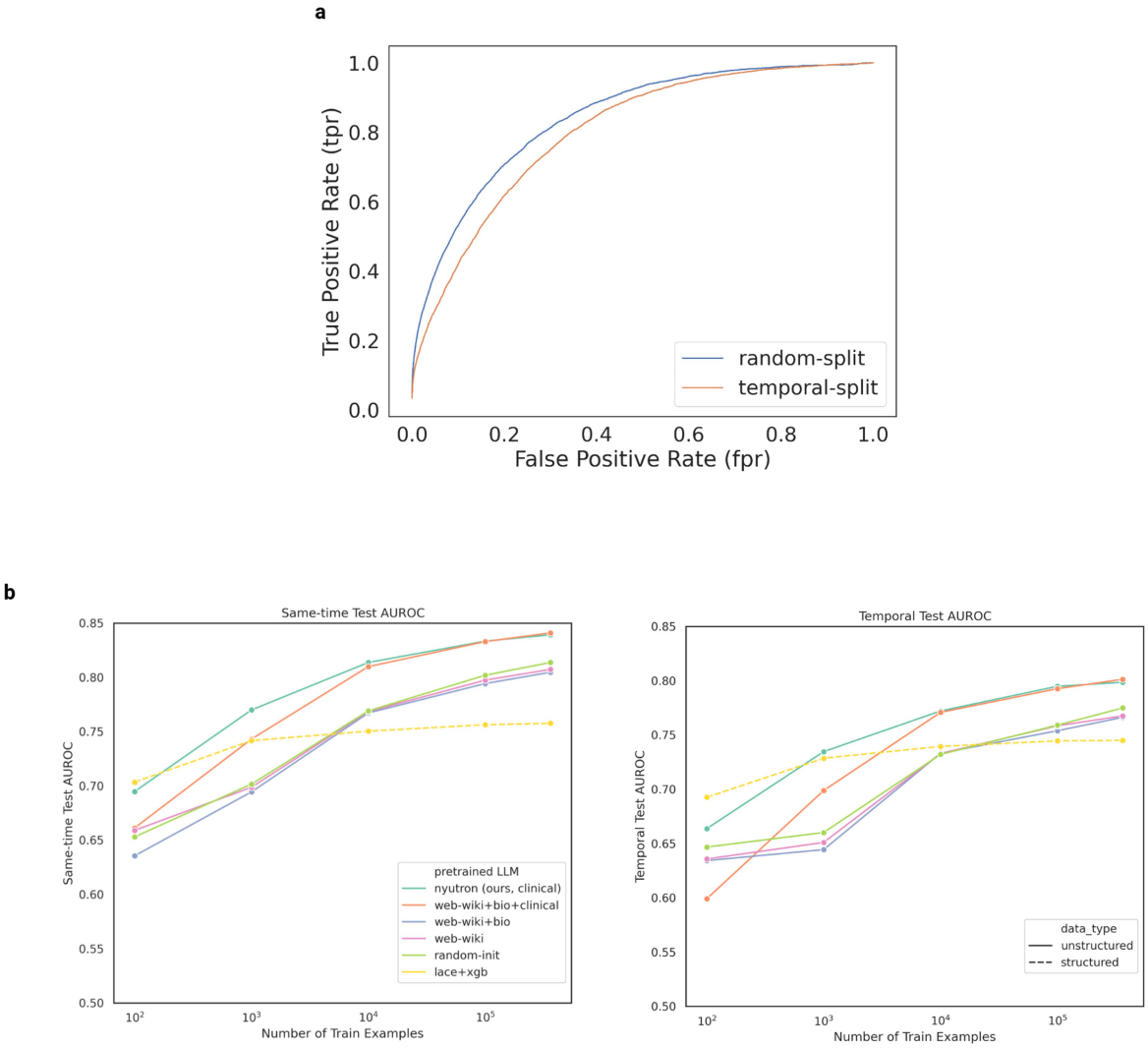
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06160-y.

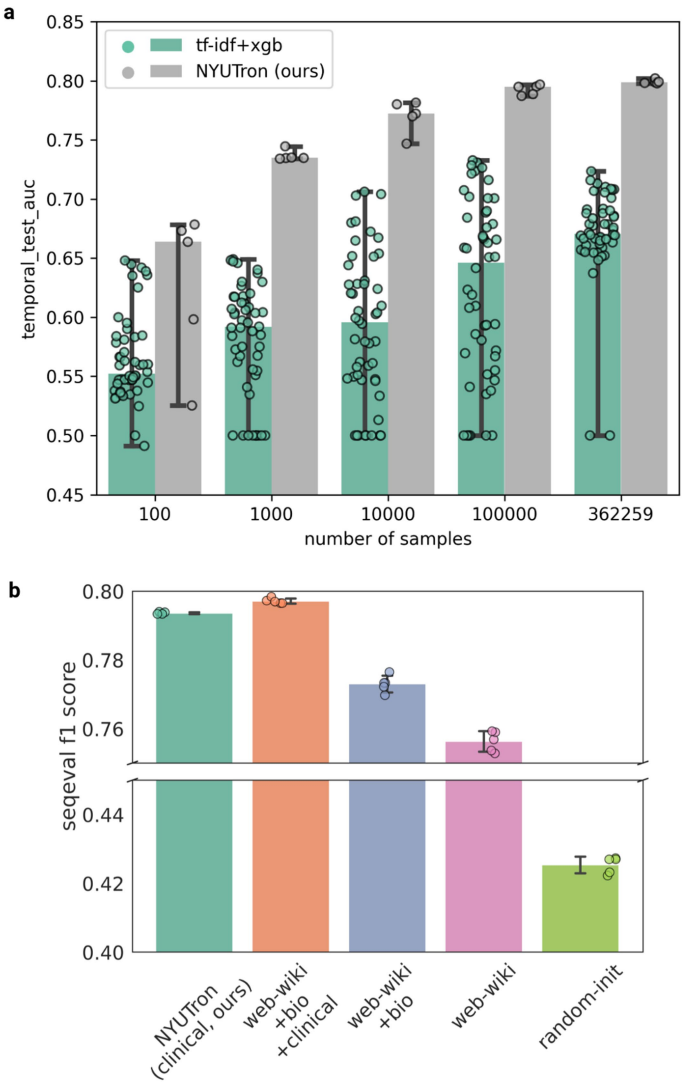
Correspondence and requests for materials should be addressed to Eric Karl Oermann.

Peer review information *Nature* thanks Ziad Obermeyer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

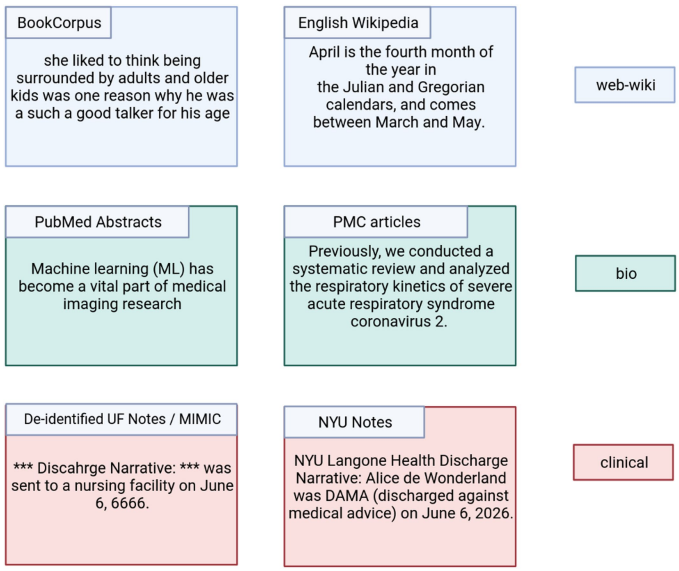
Reprints and permissions information is available at http://www.nature.com/reprints.



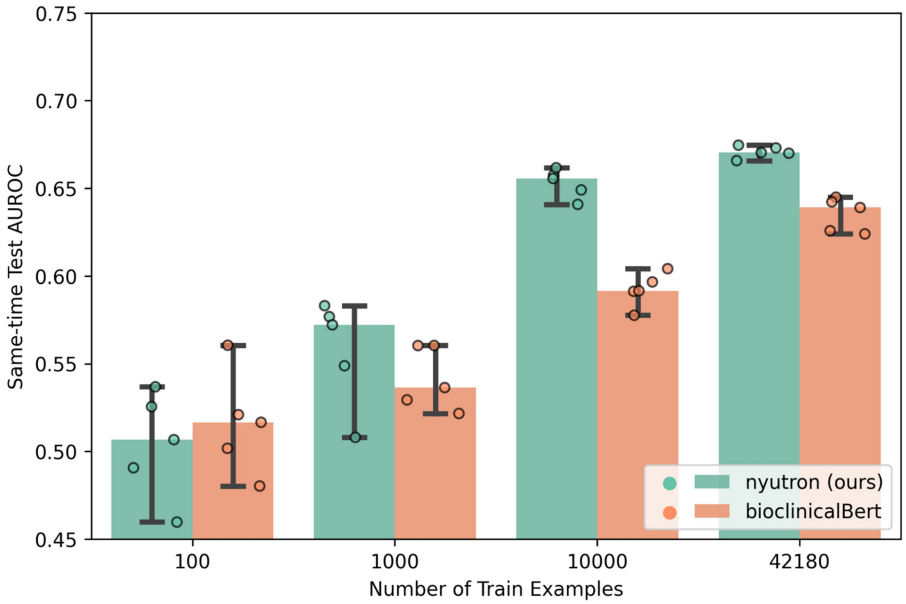
Extended Data Fig. 1 | Difference between random test and temporal test. **a**, AUC curve for the random test shows better performance than the temporal test. The random-test AUC is 84.13%, compared to the temporal-test AUC of 80.2%. The difference highlights the importance of creating a test set to reflect the problem setup. In the case of readmission prediction, the deployment set always comes from the future of the training set. Thus we use the temporal test AUC for model selection. **b**, Comparison of random-test AUC and temporal-test AUC as the number of training examples increases shows that temporal-testing is important to estimate deployment performance. Here we show that sampling a temporally split out dataset seems "harder" than a randomly sampled test dataset because all tested LLMs and lace+xgb perform worse on the temporal test (notes from the future) than the random test (notes from the same time as the training data). The colored lines on the left (random test AUCs) are generally higher than the colored lines on the right (temporal test AUCs). We conclude that this is an important distinction that temporally sampled held-out test sets give a more realistic estimate of model performance. Interestingly, the language models seem to be more sensitive to this phenomenon than the lace+xgb model.



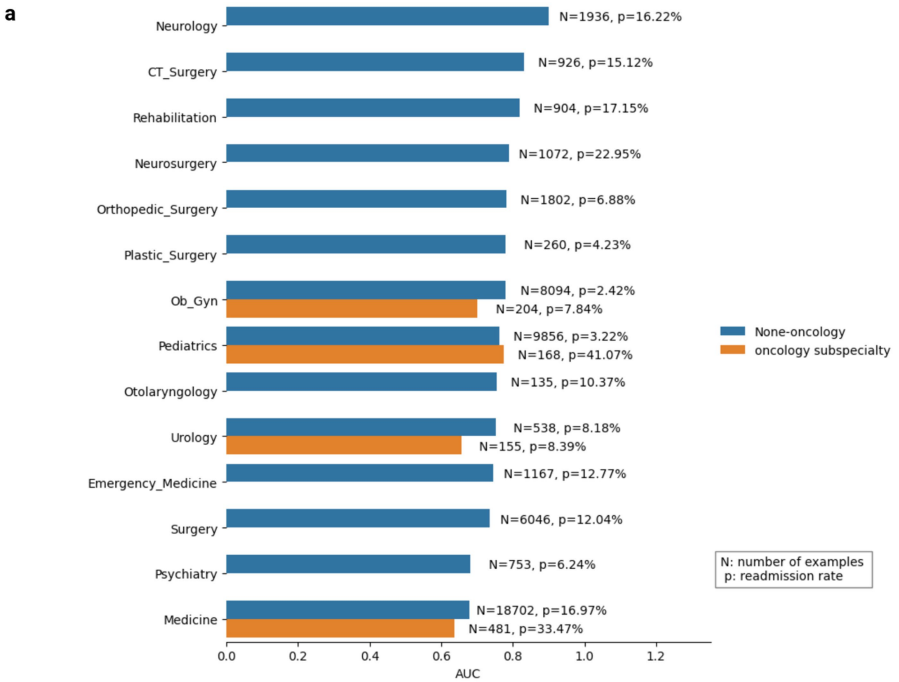
Extended Data Fig. 2 | Benchmarking NYUTron against a traditional NLP model and other language models on a different clinical prediction task (clinical concept extraction). We observe a similar trend as readmission prediction: (a) shows that NYUTron has better performance than tf-idf under different data availability settings and (b) shows that clinically pretrained language models have better performance than non-clinically pretrained language models. This corroborates our findings that health-system scale language models are general purpose clinical prediction engines and that a domain match between pretraining and finetuning corpus contributes to task performance. **a**, Comparison of temporal test AUCs between NYUTron and a traditional NLP model (tf-idf+xgb). NYUTron has a higher median AUC than tf-idf+xgb for all tested number of finetuning examples. The black vertical line indicates standard deviation over 5 trials of different random seeds (0, 13, 24, 36, 42). **b**, Comparison of LLMs' finetuning performances on the NER task. On the i2b2-2012 clinical concept extraction task, the LLMs that are pretrained with clinical corpora (NYUTron, web-wiki+bio+clinical) have a higher average f1 score than LLMs that are not pretrained with clinical corpora (web-wiki+bio, web-wiki, random-init). Specifically, NYUTron and web-wiki+bio+clinical perform better than the randomly initialized model (36.64% higher median seqeval f1 score) and non-clinically pretrained models (2.01%–3.48% higher median seqeval f1 score). Note that the height of each bar is the average f1 score and the half length of each black vertical line indicates the standard deviation over 5 trials of different random seeds (0, 13, 24, 36, 42).

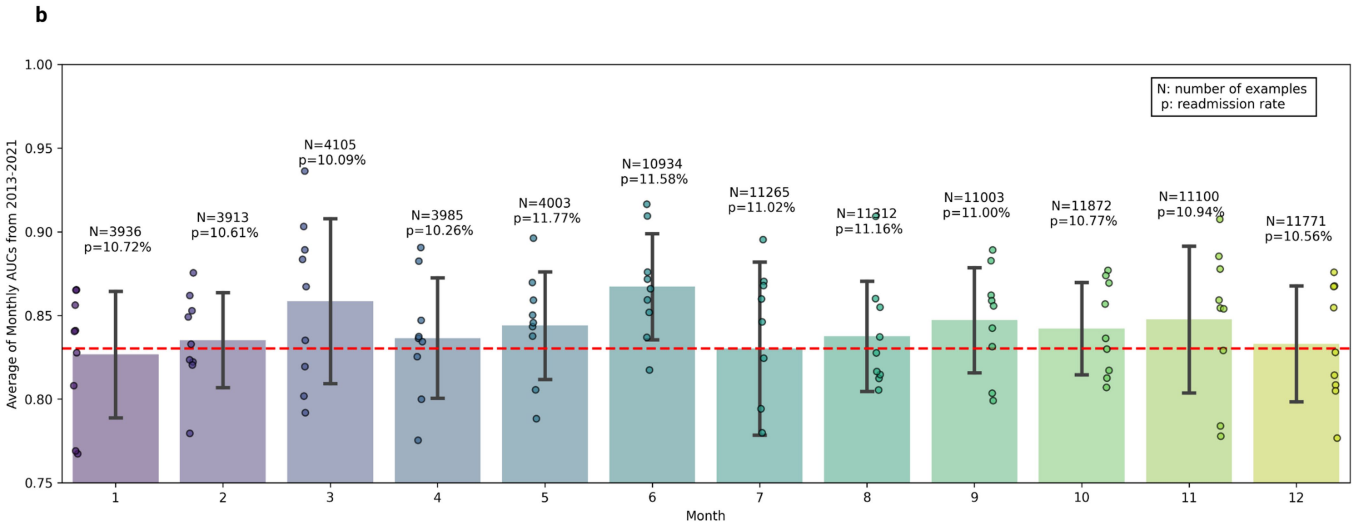


Extended Data Fig. 3 | Examples of pretraining corpora. We include here some examples from the utilized pretraining corpora to help contextualize our work. Examples from three types of pretrain corpus: (1) web-wiki (online books from bookcorpus and encyclopedia articles from English Wikipedia), (2) bio (abstracts of academic papers from Pubmed Abstracts and full articles from Pubmed Central), and (3) clinical (NYU Notes, NYU Readmission from Langone EHR and clinical notes from University of Florida Health).

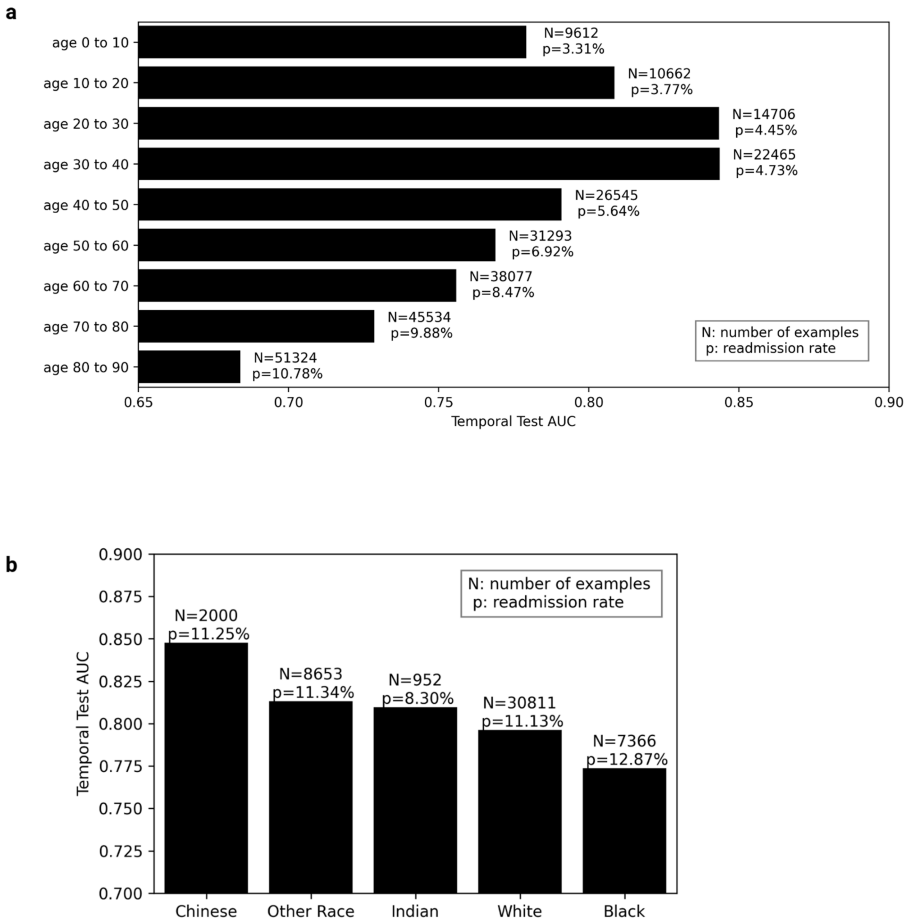


Extended Data Fig. 4 | Comparison of NYUTron's and BioClinicalBERT's performance on MIMIC-III Readmission. To test how much finetuning NYUTron needs to generalize to another health system, we finetune NYUTron and BioClinicalBERT (which has the same number of parameters and architecture as NYUTron, but pretrained on MIMIC notes, bookcorpus, pubmed and wikipedia articles) using different subsamples of MIMIC-III readmission dataset. The dataset contains 52,726 de-identified ICU discharge notes from Boston Beth Israel Hospital with 8:1:1 train-val-test split. At 100 samples, the AUC is similar. At 1000 samples, NYUTron has a 3.58% higher median AUC than BioClinicalBERT (57.22% v.s. 53.64%). At 10,000 samples, NYUTron has a 6.42% higher median AUC than BioClinicalBERT (65.56% v.s. 59.14%). Using the full dataset (42,180 samples), NYUTron has a 3.8% higher median AUC than BioClinicalBERT (67.04% v.s. 63.24%). Given that NYUTron was pretrained on identified all-department notes from NYU Langone and finetuned on de-identified ICU-specific notes from Beth-Israel, this result shows that NYUTron is able to generalize to a very different health environment through local finetuning. The height of the bar indicates the median performance of 5 experiments using distinct random seeds (0, 13, 24, 36, 42) and the error bar indicates the min-max range.

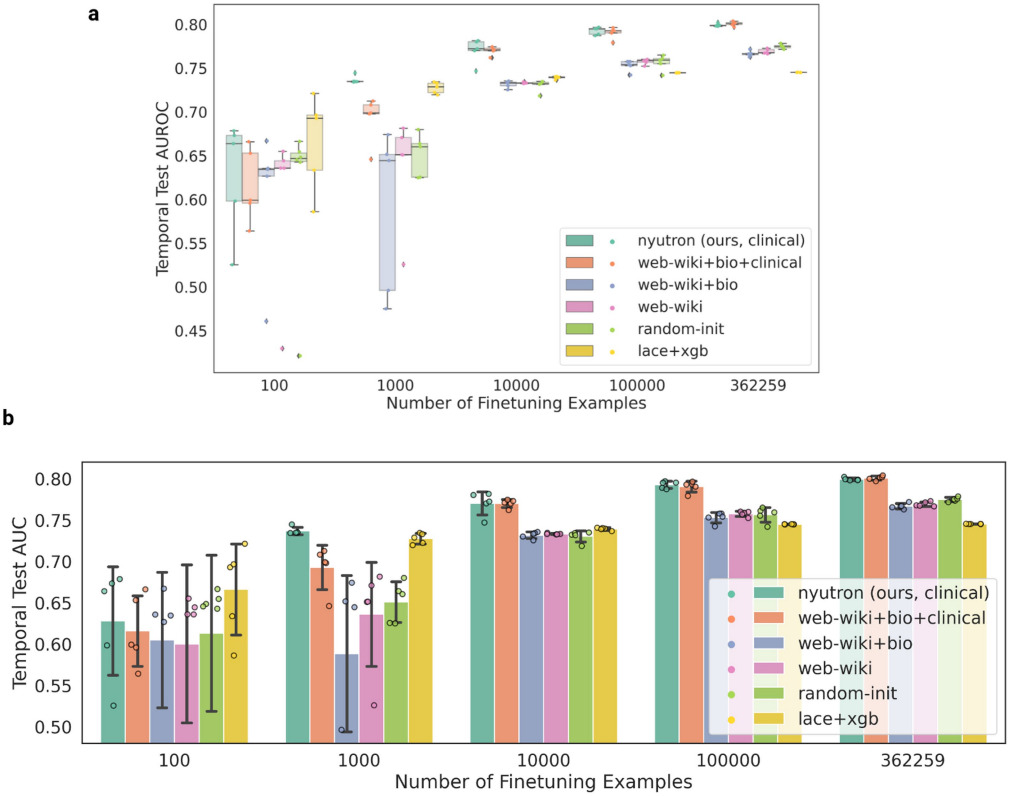




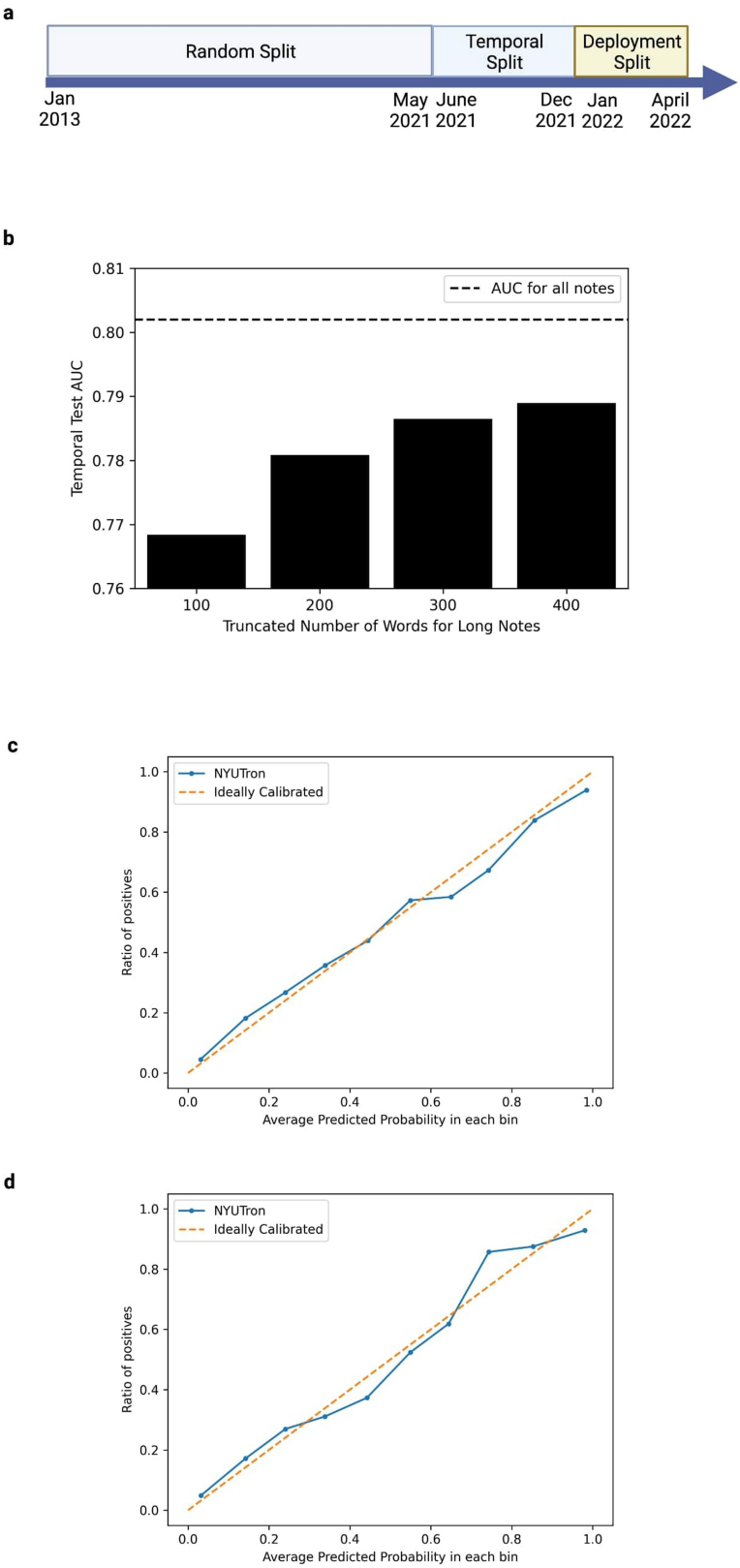
Extended Data Fig. 5 | Bias analysis stratifying NYUTron's performance by clinical departments and months. a, A stratified analysis of NYUTron's temporal test performance by clinical department and oncological subspecialty. NYUTron performs best in the Neurology Department (AUC 90.12%), and performs worst in the Internal Medicine Department (AUC 67.95% for non-oncology specialty and AUC 63.77% for oncology specialty), with a difference of about 20% AUC. This significant variance across clinical departments suggests that a more fine-grained analysis may lead to performance benefits. We annotate the number of examples (N) and the readmission rate (p) for each department. **b**, NYUTron's performance displays minor fluctuations over months. We plot the average monthly test AUC of NYUTron from January 2013 to December 2021 to look for underlying monthly trends or cycles and to test the hypothesis that performance would be worst in July when new physicians start their training with a different writing style than physicians already in practice (dashed red line indicating the monthly AUC of July). The height of the bar indicates average monthly performance across the 9 years and the vertical bar indicates the standard deviation. We annotate the number of examples (N) and the readmission rate (p) for each month. We note that July has the second lowest monthly AUC and the highest variance. We speculate (and need more years of data to verify) that clinical notes written by new physicians are associated with the temporal shift across the months and the drop in performance in July. Average AUCs from the quarters January to March, April to June, and July to September are increasing, which may coincide with residents' rotation schedule across different clinical departments. We leave further investigation of this cyclical performance to future work.



Extended Data Fig. 6 | Bias analysis stratifying NYUTron's performance by age groups and major racial groups. As part of an analysis of model performance by two possible sources of bias, age and race, we perform stratified analyses of NYUTron's performance. We annotate the number of examples (N) and the readmission rate (p) for each evaluation. **a**, We stratify the temporal test based on nine bins of ages (0 to 90 years with bins of 10 year intervals). NYUTron performs best for patients who are 10 to 40 years old, and has declining performance by decile over the age of 40 years with the worst performance in the 80–90 years of age group. We observe that this isn't an effect of sample size, the single largest sample is age 80–90, but likely reflects complexity and comorbidity burdens being disproportionately higher with advanced age. **b**, To test for potential dependencies and bias by race, we first identify the five most frequent races in the dataset (White, Other Race, Black, Chinese, Indian), then stratify the evaluation results by race. NYUTron performs best on Chinese patients and worst on Black patients with a mild variation in AUC across both groups.



Extended Data Fig. 7 | Detailed statistics of the comparison between language models and lace+xgb. a, A box plot with individual data points. For each model, 5 experiments were run using random seeds 0, 13, 24, 36, 42. The centerline of the box plot indicates the median. The upper line of the box indicates the first quantile. The lower line of the plot indicates the last quantile. The whisker extends to 1.5 times the interquartile length and the diamonds indicate outliers. **b**, A bar plot that shows the mean and standard deviation. The height of the bar indicates the mean across 5 experiments and the length of the black vertical line indicates the standard deviation.



Extended Data Fig. 8 | Additional information about readmission prediction. a, Visualization of readmission data split timelines. We visualize the random split, temporal split, and deployment split on a timeline to indicate this decision for model evaluation. The random split starts from January 2013 and ends in May 2021 (inclusive), which is further split into a 80% train set, 10% validation set and a 10% test set. The temporal split (temporal test) starts from June 2021 and ends in December 2021, a time period from which no training samples were sampled from. The deployment data is necessarily sampled from the future as it is accrued prospectively as part of our single arm, non-interventional clinical trial. **b**, NYUTron's performance increases with more complete input notes. To attempt to estimate performance as a function of sequence length we sampled a subset of "long notes" from the temporal test set. Each note in this subset has no less than 400 words, or approximately 512 tokens. We truncated these long notes to 100, 200, 300 and 400 words while keeping their readmission labels fixed in order to demonstrate the incremental gain in performance as we capture proportionally more information from each of these "long notes". The dashed line is the AUC of all notes. This figure shows that processing more words from the possible input leads to a better evaluation performance and confirms that there is a clear potential for improving performance by increasing maximum sequence length. **c**,**d** NYUTron's calibration curve for temporal test (c, number of evaluation examples is N = 53,916) and prospective deployment (d, number of evaluation examples is N = 29,286). As a reference, the orange line is the calibration curve of an ideally calibrated classifier. The blue line is NYUTron's calibration curve. Currently we do not perform any additional calibration and choose the decision threshold based on the precision and recall on the temporal validation set. The predicted probability is normalized by the largest predicted probability. Overall the model is well calibrated to the 30-day readmission task.

Extended Data Table 1 | Detailed statistics of datasets

Model	Model Size	Clinical Text	Biomedical Text	General Text
NYUTron, clinical (ours)	109m	4.1B real	0	0
web-wiki+bio+clinical	345m	82B real	6B	2.5B
web-wiki+bio	109m	0	18B	3.3B
web-wiki	109m	0	0	3.3B
random-init	109m	0	0	0

We built a comprehensive pretraining dataset (NYU Notes) with two site-specific variants (NYU Notes - Manhattan/Brooklyn) as discussed in the Methods section. For readmission prediction, we also built a finetuning dataset (NYU Readmission) with two site-specific variants (NYU Readmission Manhattan/Brooklyn), one structured-data variant (NYU Readmission - LACE), and a deployment test set (NYU Readmission - Deployment) that was sampled in real-time as part of our prospective trial. To test the breadth of NYUTron's applicability, we added 4 tasks (NYU Mortality, NYU Binned LOS, NYU Comorbidity, NYU Insurance denial) with their respective structured-data variant (NYU Mortality - SAPS2+APACHE2, NYU Binned LOS - Lisbon Portugal, NYU Insurance Denial - Claim forms). NYU Comorbidity has no structured-data variant because the task is to impute comorbidity index with the lack of structured icd codes. Finally, we have a Named Entity Recognition (NER) dataset for testing how well NYUTron generalizes to different clinical predictive tasks using non-NYU data.

Extended Data Table 2 | Sizes and pretrain corpora for LLMs

Dataset	# Notes	# Patients	# Words
NYU Notes	7,247,694	387,144	4,112,249,482
NYU Notes - Manhattan	4,342,602	256,217	2,381,466,993
NYU Notes - Brooklyn	1,337,352	104,521	1,102,078,012
NYU Readmission	506,740	413,845	487,395,462
NYU Readmission - Manhattan	296,519	240,824	253,622,053
NYU Readmission - Brooklyn	113,275	94,653	142,767,957
NYU Readmission - LACE	0	413,845	0 (structured data)
NYU Readmission - Deployment	29,287	27,376	34,669,963
NYU Mortality	469,162	371,922	484,467,141
NYU Mortality - SAPS2+APACHE2	0	371,922	0 (structured data)
NYU Binned LOS	469,162	371,922	484,467,141
NYU Binned LOS - Lisbon Portugal	0	371,922	0 (structured data)
NYU Comorbidity	403,579	327,039	422,485,417
NYU Insurance Denial	55,791	54,563	51,270,256
NYU Insurance Denial - Claim forms	0	54,563	0 (structured data)
i2b2-2012-NER	310	≤ 310	636K

We test 6 types of LLMs with different model sizes and different pretraining corpora. We list out the various corporate here as well as model parameter counts to facilitate ease of comparison. We also note that one key distinction between web-wiki+bio+clinical and NYUTron, clinical is that the former was stripped of identifying information while the latter was not.

nature portfolio

Corresponding author(s): Eric Oermann

Last updated by author(s): Mar 28, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used sql and Python 3.8.13 to collect data from NYU Langone EHR. We downloaded MIMIC-III dataset (https://physionet.org/content/mimiciii/1.4/) and i2b2-2012 dataset (https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/). We used REDCap 12.4.31 to collect physician responses.
Data analysis	This work uses several open-source libraries including HuggingFace Transformers 4.19.2, Datasets, 2.2.2, Evaluate 0.1.1, wandb 0.12.17, matplotlib 3.5.2, seaborn 0.12.2, pandas 1.4.2, ray 2.0.0, sklearn 1.1.1, deepspeed 0.8.0+384f17b, NVidia Apex, XGBoost 1.6.1 and spaCy 3.5.0. Our experimental framework involves the utilization of these libraries and in some cases modification of them. We will release code to replicate the pretraining, finetuning and testing of the models described in this paper at the time of publication. We included detailed methods and implementation steps in the Methods and Supplementary Information to allow for independent replication. Code for experiment: https://github.com/nyuolab/NYUTron Code for preprocessing i2b2-2012: https://github.com/nyuolab/i2b2_2012_preprocessing

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We provide links for the following public data:
MIMIC-III dataset (https://physionet.org/content/mimiciii/1.4/).
i2b2-2012 NER dataset (https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/).

The clinical data used for the pretraining, finetuning, validation, and test sets were collected from the NYU Langone Health System EHR maintained by the NYULH Datacore team. Text data was stripped of rich text features and directly included in the dataset "as-is", and was augmented with structured features where noted. It consists of the production medical records of NYU Langone and cannot be made publicly available.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	We did not perform sex and gender analysis
Population characteristics	We performed analysis of model performance on test sets that are stratified by race and age. Our pretraining population has a mean age of 50.66 with standard deviation of 28.52. Our pretraining population has a self-reporting male-female ratio of 3:4. Our finetuning population has a mean age of 49.17 and a standard deviation of 28.76. Our finetuning population has a self-reporting male-female ratio of 3:4.
Recruitment	We recruit anyone who is admitted to the NYU Langone Health Hospital from 2011 to 2022 April.
Ethics oversight	NYU Langone Health Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For pretraining, we used 7,247,694 notes with 387,144 patients. The notes were from 2011 to 2020. For finetuning, we used 506,740 notes with 413,845 patients. The notes were from 2011 to 2021. Each clinical note is considered a note sample and each patient id is considered a patient sample. The sample size is determined by counting the number of notes or patients in our dataset. Our goal was to build a dataset at the scale of our health system, and the sample sizes were determined by the size of the NYU Langone EHR. We performed additional scaling experiments in the manuscript to investigate effects of smaller datasets.
Data exclusions	For both pretraining and finetuning, we excluded notes that are not signed by medical professionals (physicians, residents, physician assistants, nurse practitioners, fellows). This is because other types of notes (e.g., from pastor and social workers) do not record the clinical decision making process. For finetuning, we excluded discharge notes from the rehabilitation, dialysis, and palliative care departments because these are not acute care admissions.
Replication	For retrospective study, we ran experiments using 5 different random seeds and achieve similar results. The prospective study was run in a live clinical environment, inference results were run once and then served to clinicians via e-mail.
Randomization	We split the pretraining data in 3 splits with ratio 8:1:1. We split the finetuning data into 4 sets: training, validation, test, and temporal test set. The first 3 sets are notes from January 2011 to May 2021, with a ratio of 8:1:1. The temporal test set are notes from June to December of 2021.
Blinding	For all experiments, at test time, investigators were blinded to group allocation at time of analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Did not register prospectively.
Study protocol	NYU Langone Health Medical Center Information Technology (https://nyulangone.org)
Data collection	Data was collected from NYU Langone Health System prospectively for all patients encountered in the health system for the duration of the trial.
Outcomes	Our primary outcome was successful identification of all-cause 30-day readmissions at discharge during the index admission.