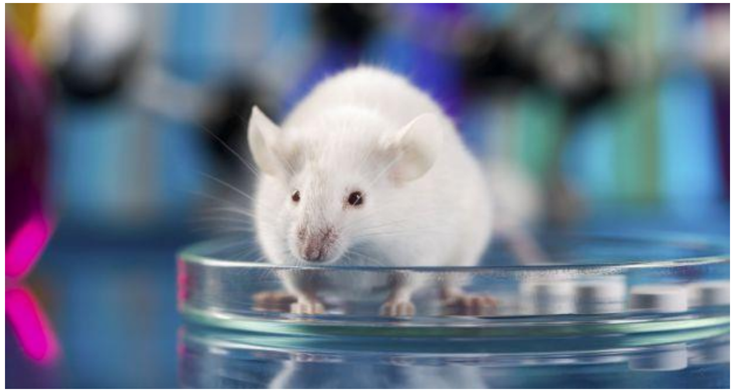
Big data in medicine

Strengths and pitfalls

Journal Club
February 2020
Marc Emmenegger

Research with human data



- For basic research questions that are purely curiosity-driven, many models available to life science researcher.
- For research on human disease, human biospecimen and associated data may be preferential in many cases.

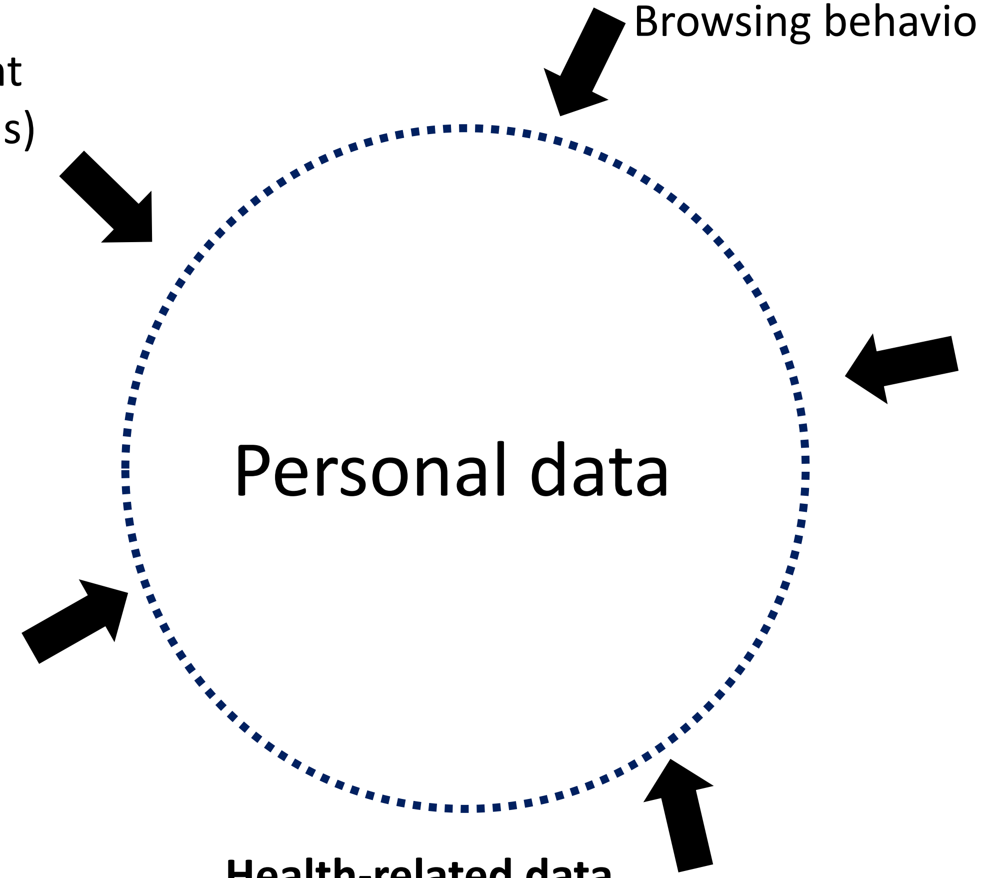
Personal data is (considered) sensitive

Details of bank account
(access, financial status)

Browsing behaviour

Political views

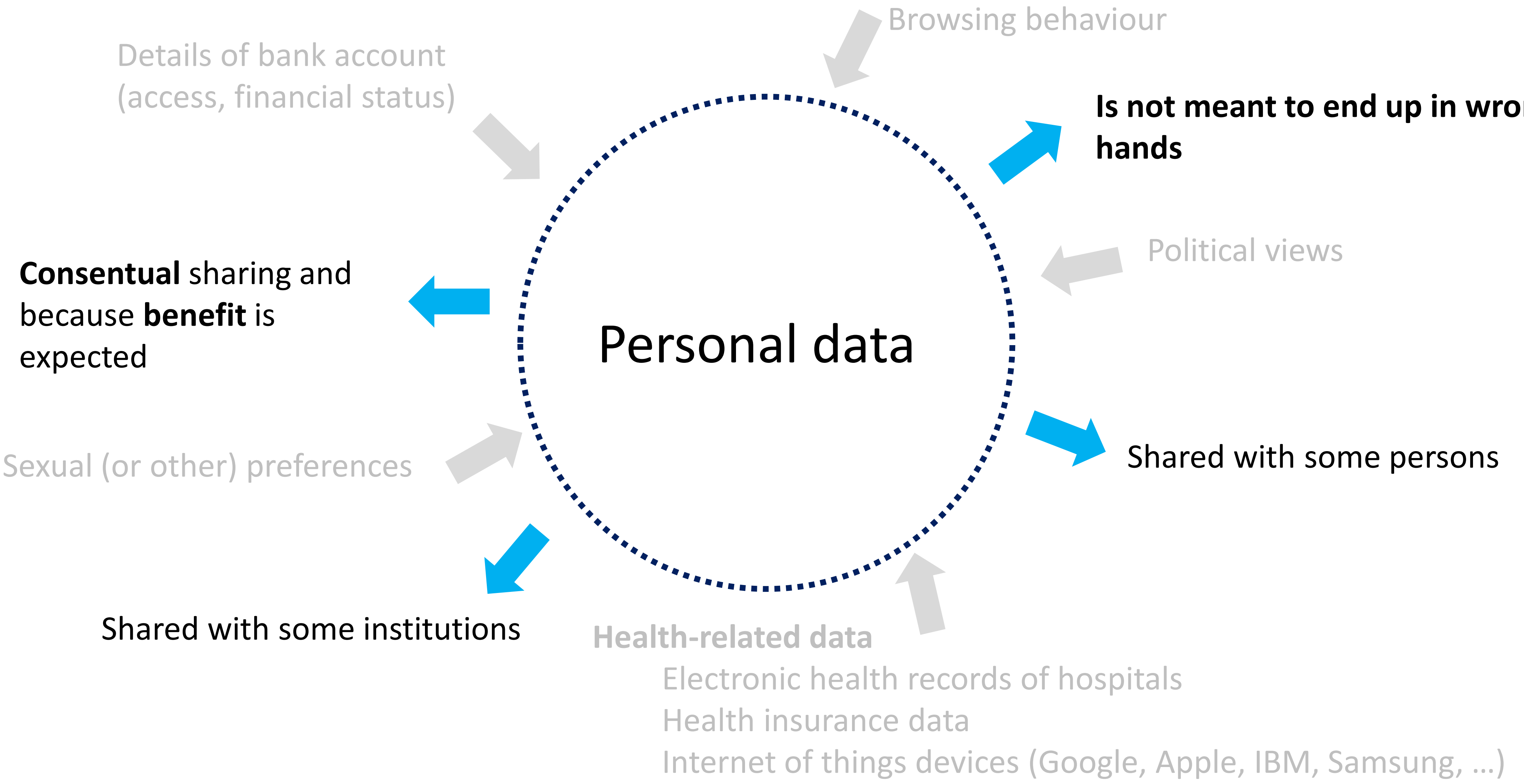
Sexual (or other) preferences



Health-related data

Electronic health records of hospitals
Health insurance data
Internet of things devices (Google, Apple, IBM, Samsung, …)

Personal data is (considered) sensitive



Personal data is (considered) sensitive

Browsing behaviour

Details of bank account
(access, financial status)

Data breach

Political views

Consentual sharing and because **benefit** is expected

Personal data

Sexual (or other) preferences

Shared with some persons

Shared with some institutions

Health-related data

Electronic health records of hospitals
Health insurance data
Internet of things devices (Google, Apple, IBM, Samsung, …)

NEWS EXPLAINER · 29 MARCH 2018

The scant science behind Cambridge Analytica's controversial marketing techniques

Nature peers into the evidence for 'psychographic targeting'.

Elizabeth Gibney

Reported data breaches

- Cambridge Analytica, involved in Trump’s 2016 election campaign.
- Received data of millions of Facebook users, without explicit user consent.
- Data hoard was used to target voters with messages personalised to their personality traits, method called **psychographic marketing**.
- Usage of personal data for non-intended purpose, with potential malicious intent.

NEWS · 19 NOVEMBER 2019

Google health-data scandal spooks researchers

Scientists fear the controversy over the Nightingale project will undermine trust in research.

Heidi Ledford

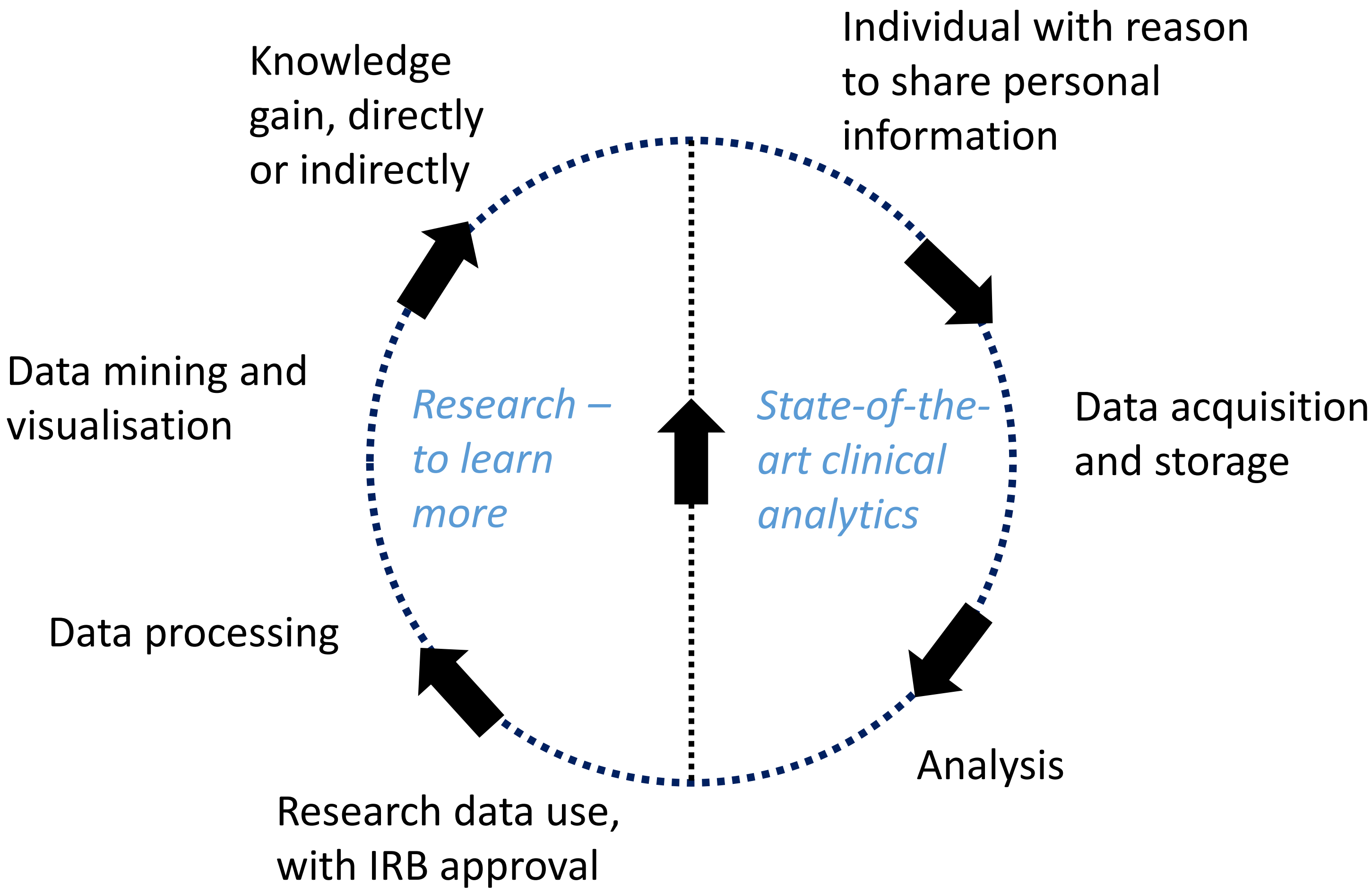
Reported data breaches

Google's secret cache of medical data includes names and full details of millions - whistleblower

The Guardian

- Nightingale: Access to Google to health-care information, including names and other identifiable data, of tens of millions of people without their knowledge.
- People treated at health network Ascension in St. Louis.
- According to Google, data is used to improve health care.
- But no patient consent.
- Reminiscent of DeepMind affair where Google gained access to health data of patients from London, without their knowledge.
- Generally, there is a lack of regulations regarding the corporate use of personal data as in most countries, companies – opposed to academic institutions – do not require an ethical approval for the commercial use of data.
- Yet, obvious lack of data privacy jeopardises public trust in data-sharing practices, also those that are used for the benefit.

Presentation outline



Presentation outline

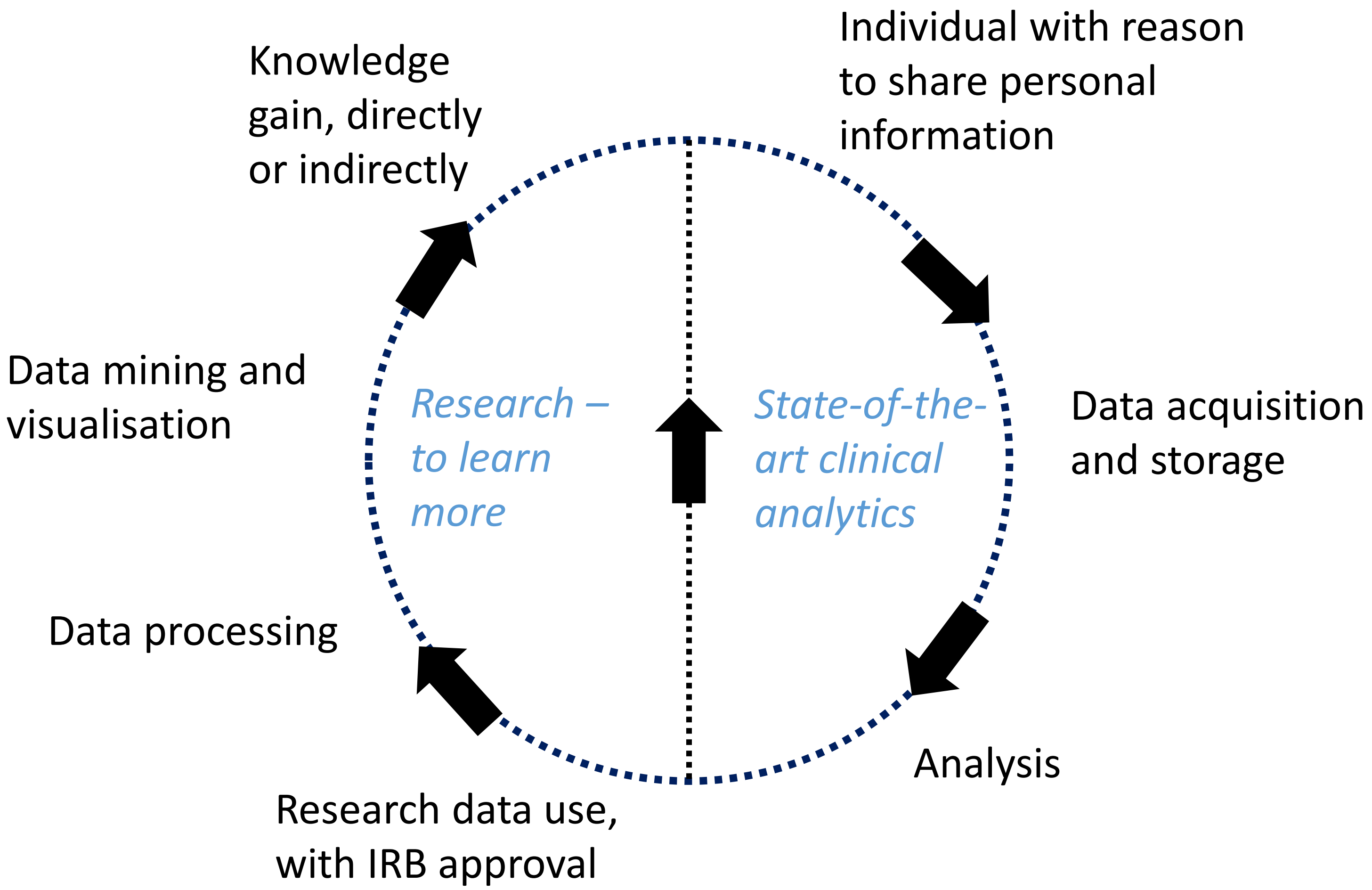
Knowledge gain, directly or indirectly → Individual with reason to share personal information

You benefit directly from research on your data

Somebody else benefits from research on your data

You benefit from research on data of many people

Setting up of big medical data study



Privacy, data governance

Proper analytics

Evaluating the impact on society

nature medicine
FOCUS | REVIEW ARTICLE
https://doi.org/10.1038/s41591-018-0272-7

Privacy in the age of medical big data

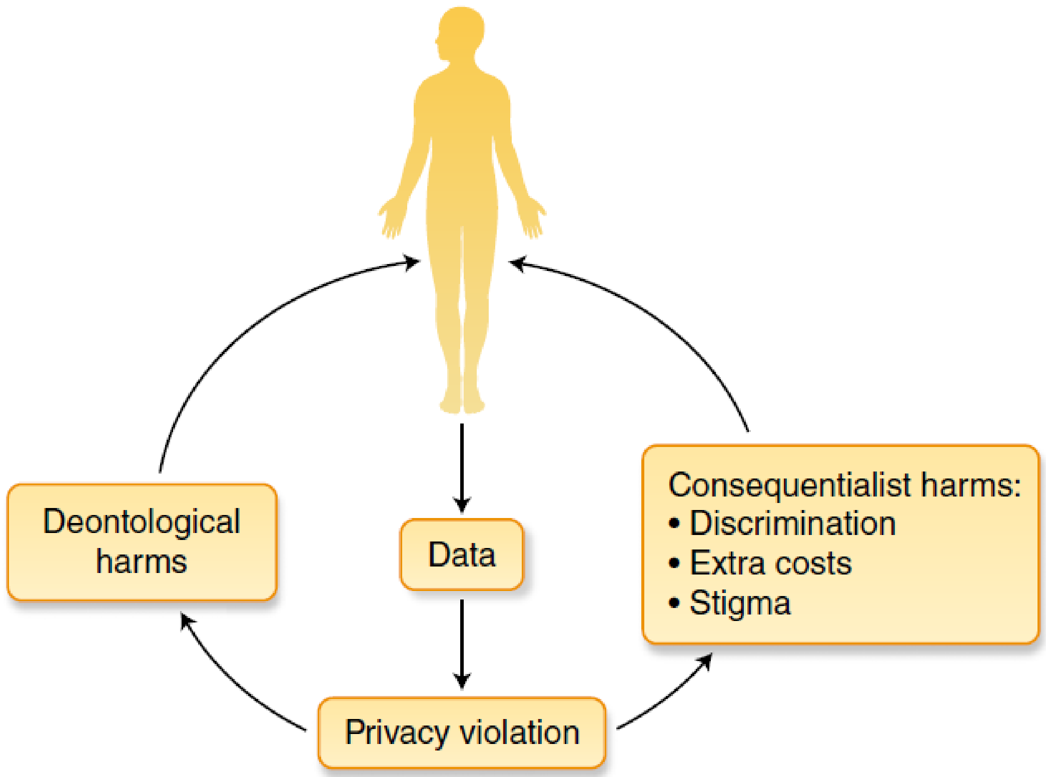
W. Nicholson Price II[1,2,3] and I. Glenn Cohen[2,3,4*]

Privacy in the age of medical big data – intro

- With big data comes big risks and challenges, among them significant questions about patient privacy.
- Advocates of big data promise increased accountability, quality, efficiency, innovation.
- Too little privacy raises concerns, too much privacy in this area can pose problems, too.
- Concept of privacy difficult to define.
- Prominent view links privacy to context: Depends on the actors involved, the process by which information is accessed, frequency of accesses, and purpose of access.
- Violation: Wrong actor gets access, purpose inappropriate.

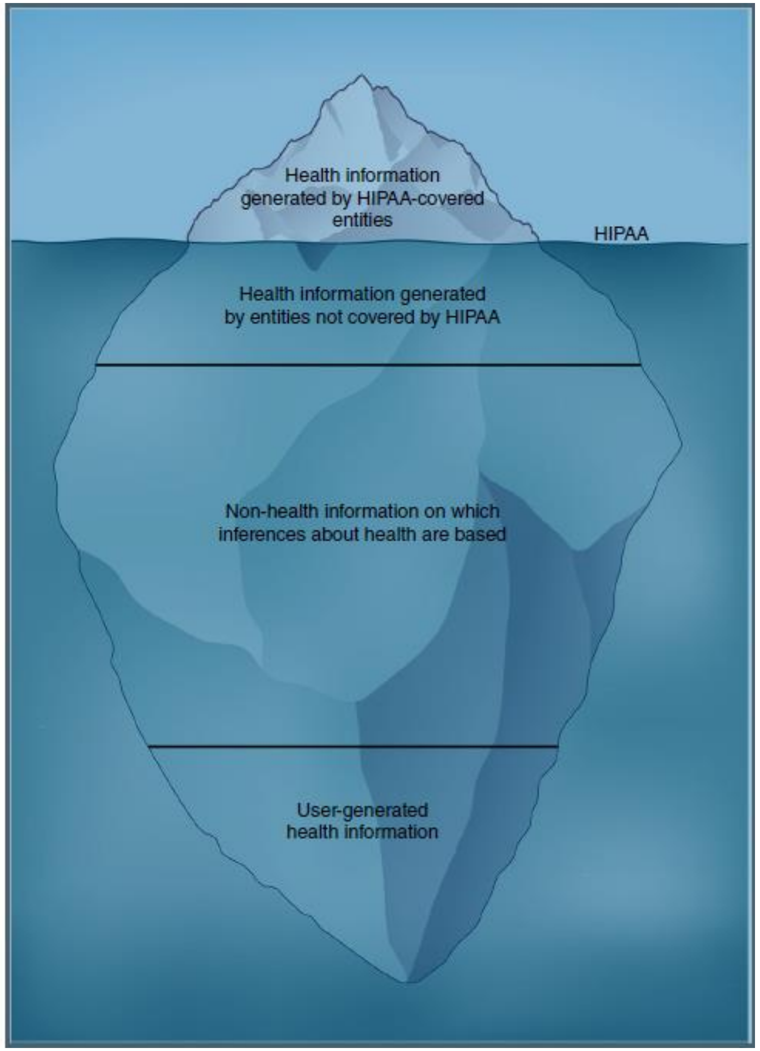
Consequentialist and deontolgogical concerns regarding data privacy

- Consequentalist concerns:
  - Result from negative consquence that affect the person whose privacy has been violated.
  - Tangible consequences.
  - Rise in health insurance premium because of poor genetic prognosis.
  - Discrimination because of HIV status.
  - Psychological distress because private information could be abused – even before actual misuse occurs.
- Deontological concerns:
  - Do not depend on potentially experiencing negative consquences.
  - Concerns from privacy violations manifests even if no one uses a person’s information against this person.
  - One may be wronged by privacy breach even if one is not harmed.



Gathering data

- Gathering of medical data raises many legal and ethical privacy questions.
- Health data come from many different sources: electronic health records, insurance claims, Internet of things devices, social media posts, …
- EU General Data Protection Regulation (unlike the US privacy law) set out single broadly defined regime for health (as well as other) data no matter what format, how it is collected, or who the custodian is.
- In the US, the Health Insurance Portability and Accountability Act (HIPAA) covers only a small fraction of health-related data – Google, Apple, IBM all operate outside the HIPAA regime.



Data anonymisation

- Many assume that ‘anonymized’ data cannot be used to reidentify the subject of the data.
- Unfortunately, as data sets proliferate, the ability to combine multiple datasets may defeat the de-identification strategy.
- State of Massachusetts purchased health insurance for state employees and subsequently released records summarizing every state employee’s hospital visits at no cost to any researcher who requested the data.
- Then-Governor William Weld assured the public that the data had been scrubbed to defeat reidentification by removing information such as names, addresses, and social security numbers.
- Unfortunately, many patient attributes were not scrubbed.
- Sweeney, then a graduate student, knew Weld resided in the city of Cambridge, and so she purchased this city’s complete voter rolls, which contained the name, address, ZIP code, birth date, and sex of every voter in the city.
- She paired that data with the state health insurance data to demonstrate that one could reidentify Weld’s prescriptions, medications diagnosis, and medical history.

Equitable data collection

- Concern is not that too much data may be taken from patients but that data collection is not occurring equitably.
- Data collection may be justified in the sense of a bargain between data sources (say patients) and data users (e.g. scientists).
- Patients are willing to, in part, compromise their privacy since added value is generated that befits their health.
- When balance is off, bargain may break down.
- Existing bias can reappear in data mining – ***example will be shown later***.
- Datasets should be inclusive.

Role of patient in data collection and access

- To what extent should an individual’s data be available for use in predictive analytics without her/his consent?
- Should health data be considered a public good?
- If the patients benefit extremely well from an analysis even if their data is used without being asked before, is it wrong?

How to address concerns?

- Perhaps data sharing should be limited to the minimal amount necessary in all contexts, data should be retained only for limited time, or data should be intentionally obfuscated if consequential harms are difficult to limit.
- Nevertheless, limits on data access can bring their own harms.
- The basic harm of privacy overprotection is the brakes it puts on data-driven innovation.
- Data holders should be stewards of data, not privacy-agnostic intermediaries.
- Privacy also interacts problematically with secrecy.
- There are many potential innovations that can arise from data, and some of these may be very lucrative, such as an algorithm that accurately selects cancer drugs.
- Innovators have incentives to keep data secret to maintain a competitive advantage in development and deployment of such valuable innovations but we might prefer as a society to have access to the data on which such innovations are based: others can use those data to create better predictors from the same data, to aggregate data to find more subtle patterns, or to validate and verify that the original innovator’s research was accurate.

What is done in Switzerland?

Legislation

Switzerland:

- Federal Act on Data Protection (SR 235.1)
- Federal Data Protection Ordinance (SR 235.11)
- Human Research Act (SR 810.30)
- Human Research Ordinance (SR 810.301)
- Swiss Penal Code (SR 311.0)

European Union:

- General Data Protection Regulation (Regulation (EU) 2016/679)

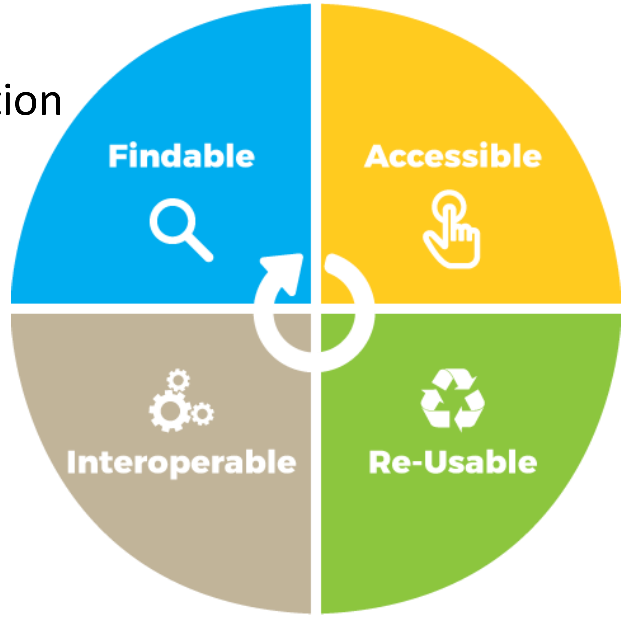
Data collection and sharing

- Legitimacy
- Proportionality
- Appropriation/purpose limitation
- Recognisability
- Responsibility
- Information security
- Control

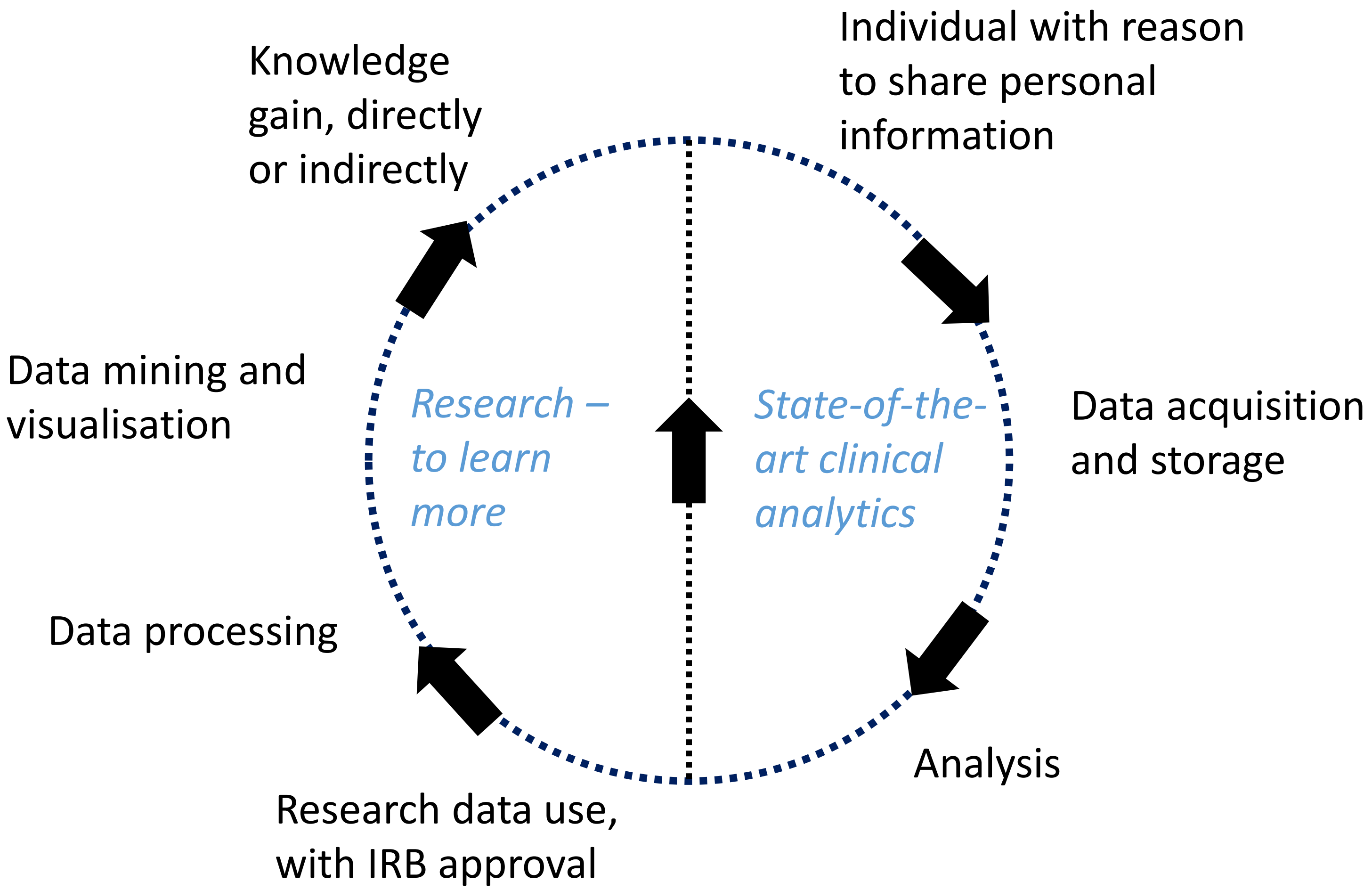
SPHN Swiss Personalized Health Network

Data stewardship and curation

The FAIR data principles



Setting up of big medical data study



Privacy, data governance

Proper analytics

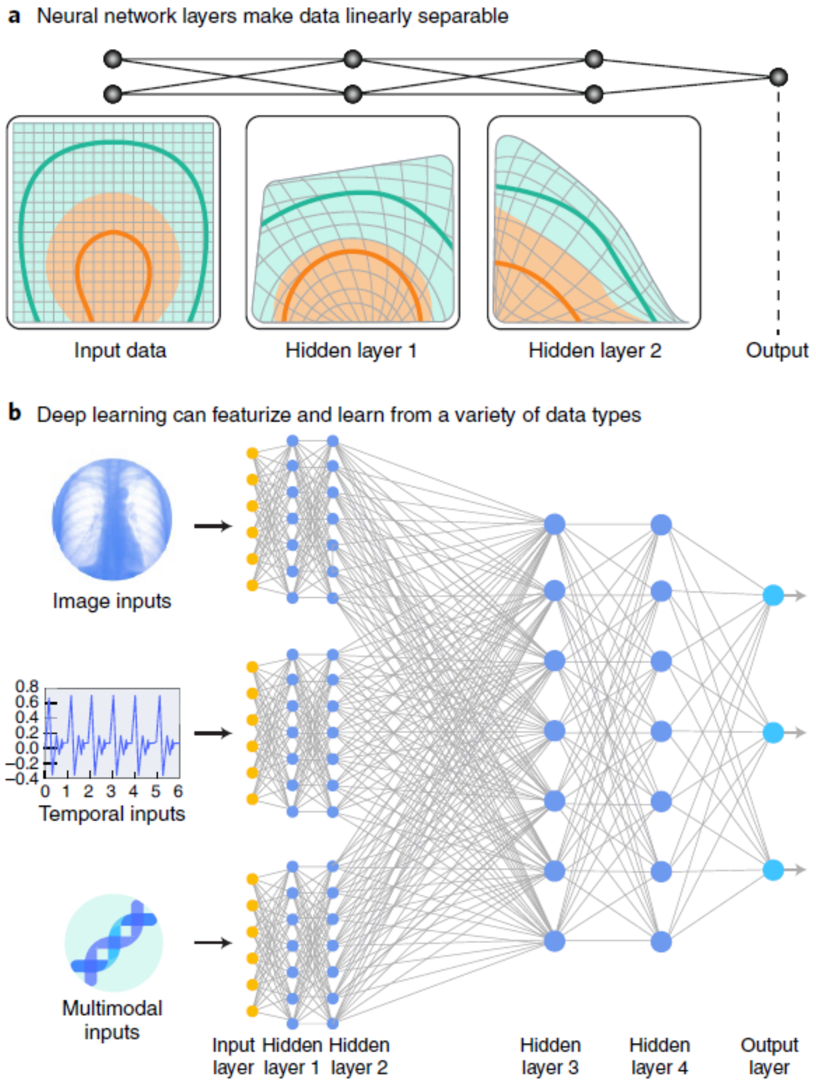
Evaluating the impact on society

PERSPECTIVE | FOCUS
https://doi.org/10.1038/s41591-018-0316-z
nature medicine

A guide to deep learning in healthcare

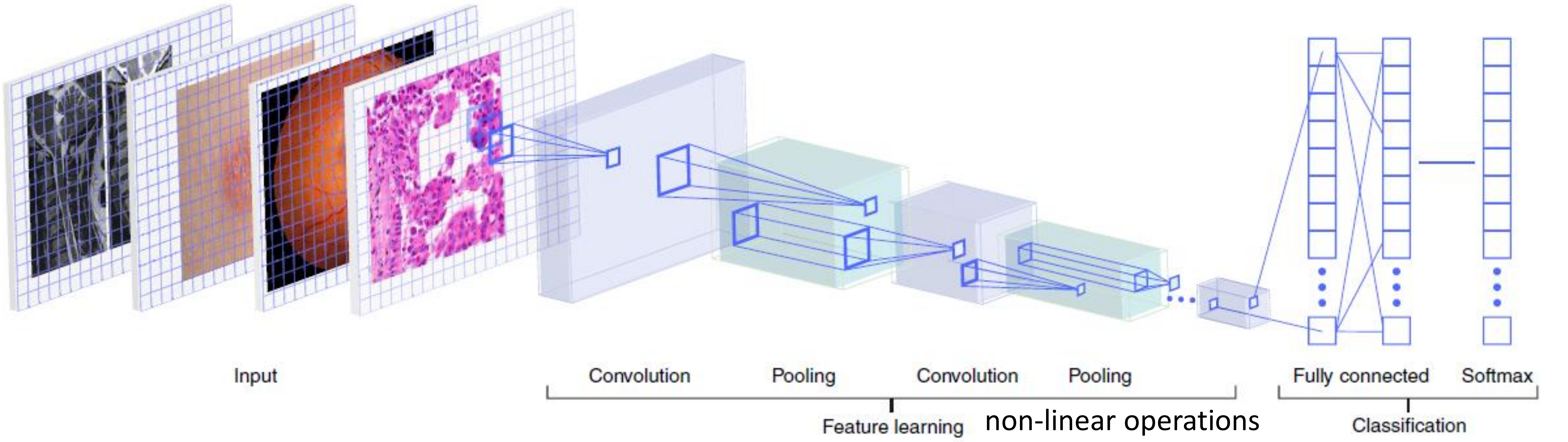
Andre Esteva[1,3]*, Alexandre Robicquet[1,3], Bharath Ramsundar[1], Volodymyr Kuleshov[1], Mark DePristo[2], Katherine Chou[2], Claire Cui[2], Greg Corrado[2], Sebastian Thrun[1] and Jeff Dean[2]

Deep learning



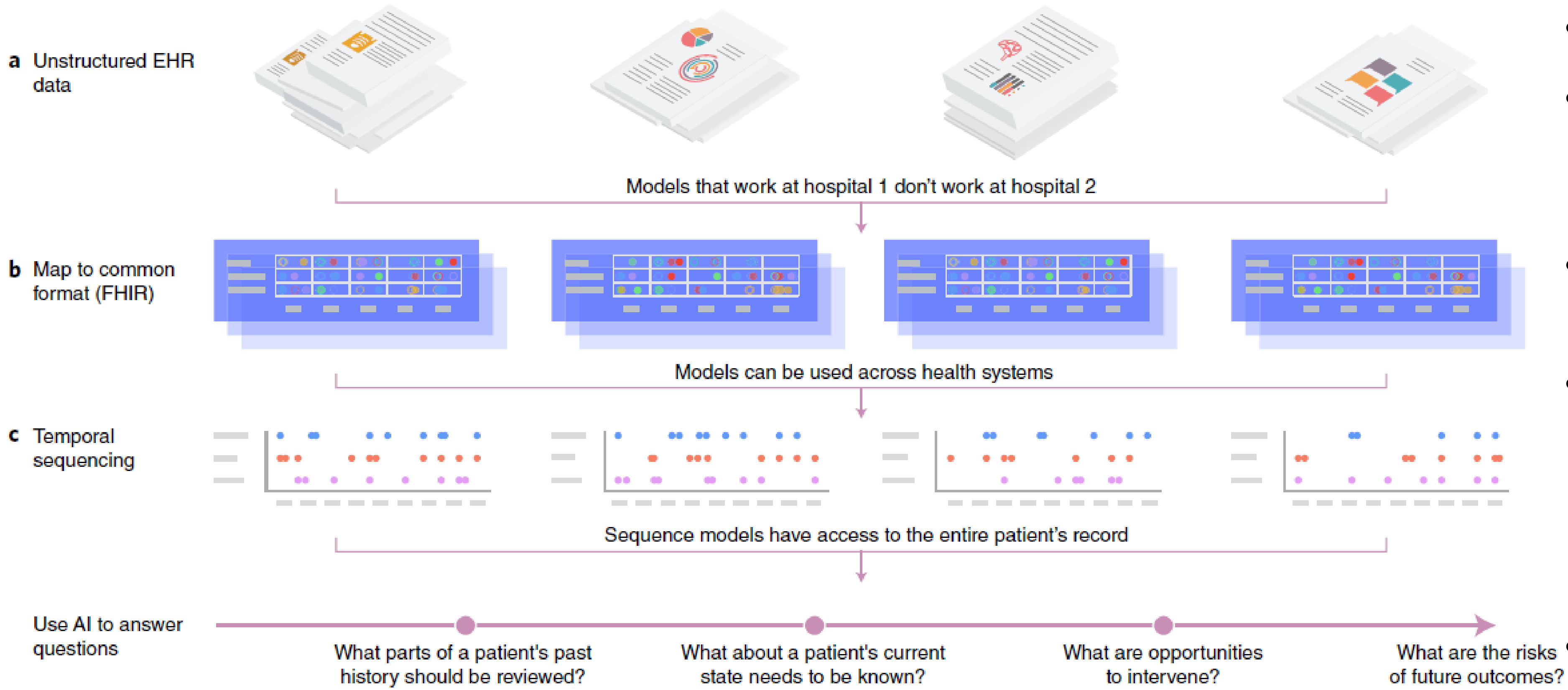
- Deep learning – subfield of machine learning.
- Resurgence largely driven by increases in computational power and availability of massive datasets.
- Striking advances in the ability of machines to understand and manipulate data: images, language, speech.
- Machine learning: Transforms inputs of an algorithm into outputs using statistical, data-driven rules.
- Deep learning: Form of representation learning (machine fed with raw data, develops its own representations needed for pattern recognition) that is composed of multiple layers of representations (**a**).
- Deep learning systems can accept multiple data types as input – relevant for heterogeneous healthcare data (**b**).

Computer vision



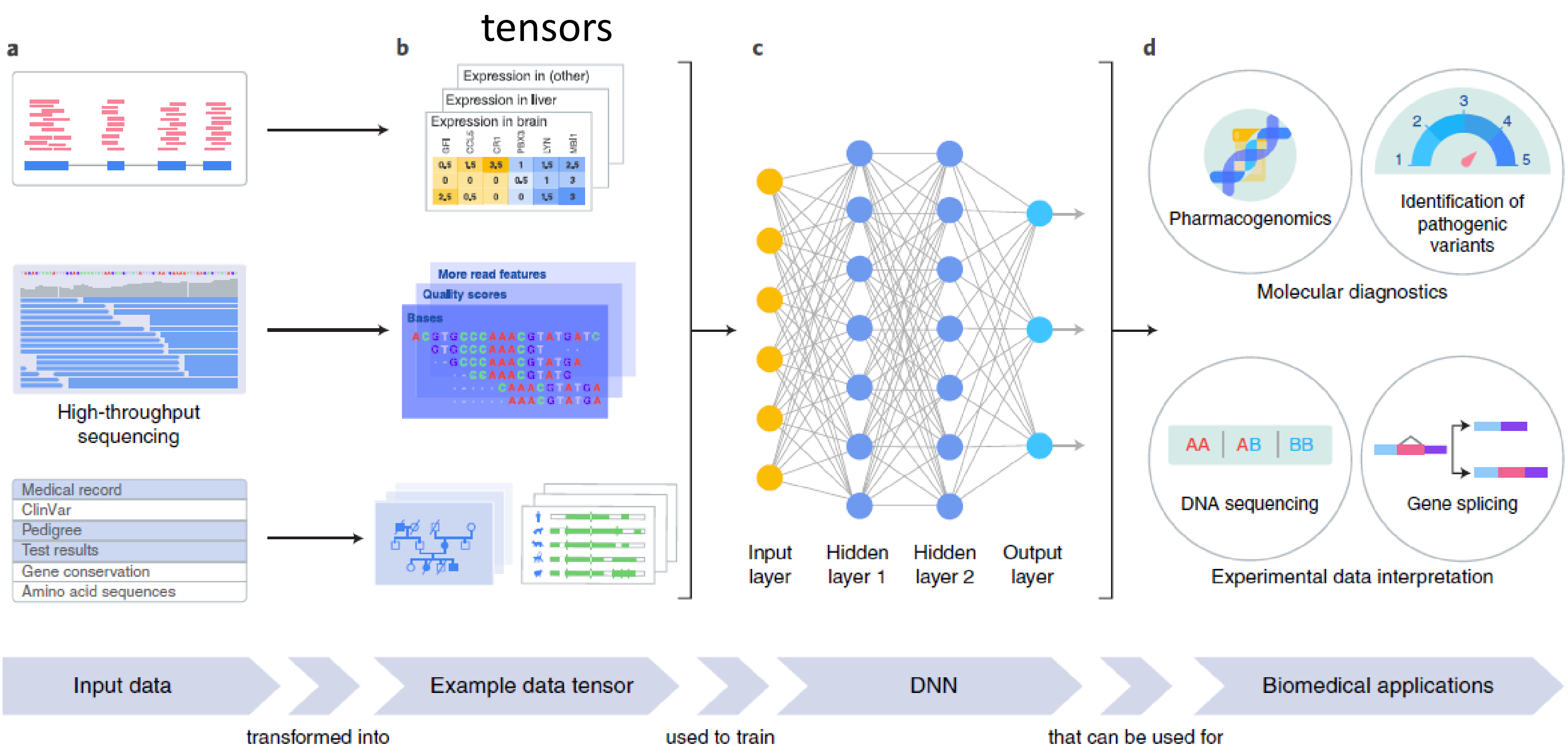
- Some of greatest success of deep learning in field of computer vision.
- First step: algorithm leverages large amounts of data to learn of the natural statistics in the images, such as straight lines, curves, coloration, …
- Second step: Higher-level layers of algorithms are retained to distinguish between diagnostic cases.

NLP

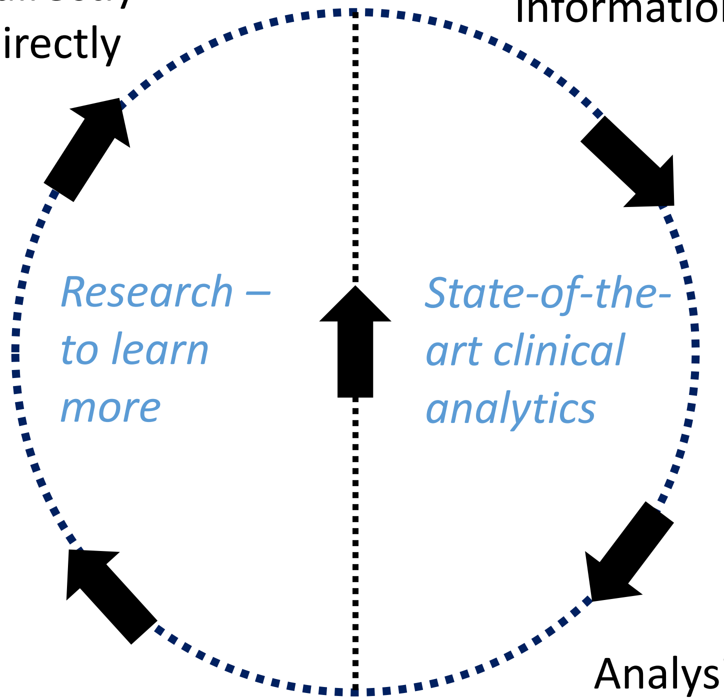


- Aggregate data.
- Build generalisable model.
- Standardise model.
- Use deep learning for health prediction based on wealth of input data.
- Lots of potential but current use is limited.

Generalised deep learning



Setting up of big medical data study



Privacy, data governance

Proper analytics

Evaluating the impact on society

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Dissecting racial bias in an algorithm used to manage the health of populations

- Growing concern that algorithms may reproduce racial and gender disparities via the people building them or through data used to train them.
    - Job search ads for highly paid positions less likely to be presented to women (Datta et al, 2015)
    - Searches for distinctively Black-sounding names are more likely to trigger ads for arrest receords (Sweeney, 2013)
    - Image searches for professions such as CEO procude fewer images of women (Kay et al, 2015)
    - Facial recognition systems (e.g. used in law enforcement) perform worse on recognising faces of women and Black individuals (Klare et al, 2012, Buolamwini et al, 2018)
    - Natural language processing algorithms encode language in gendered ways (Caliskan et al, 2019)

Dissecting racial bias in an algorithm used to manage the health of populations

- Algorithmic bias
  - Difficult to empirically study
  - **Algorithms deployed on large scales are typically proprietary**
  - Workarounds for researchers tidious
  - Disparities can be documented but, given the lack of insight into the algorithm itself, understanding why they arise remains mostly unclear

- Here, they investigated a rich dataset that provides insights into live, scaled algorithm deployed in the USA.
- Applied to roughly 200 million people in the USA each year.
- Target patients for «**high risk care management**» programs, programs that seek to improve care of patients with complex health needs.
- Such programs are considered **effective at improving outcomes** and satisfaction while **reducing costs**.

Dissecting racial bias in an algorithm used to manage the health of populations

- Aim: *Identify those patients who will benefit the most*!
- Challenging causal inference problem that requires estimation of individual treatment effects.
- Key assumption: Those with greatest care needs will benefit the most from the program.
- Predict who needs most care, identify who benefits most (→ «simple» prediction problem).

Past data $\xrightarrow{\text{Algorithm}}$ Predictor of future health care needs → Optimise care and cost

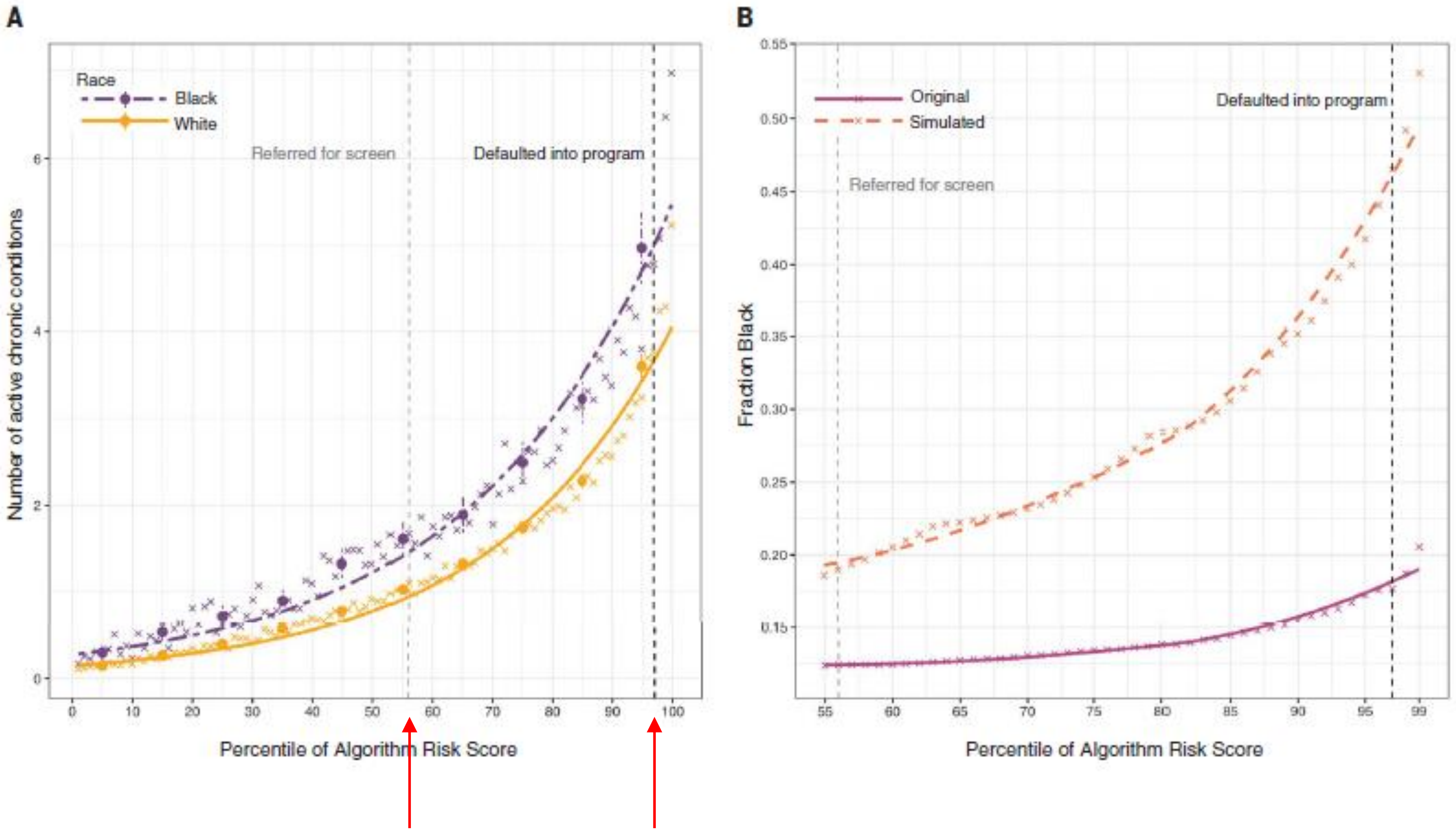
Dissecting racial bias in an algorithm used to manage the health of populations

Table 1. Descriptive statistics on our sample, by race. BP, blood pressure; LDL, low-density lipoprotein.

	White	Black
n (patient-years)	88,080	11,929
n (patients)	43,539	6079
Demographics		
Age	51.3	48.6
Female (%)	62	69
Care management program		
Algorithm score (percentile)	50	52
Race composition of program (%)	81.8	18.2
Care utilization		
Actual cost	$7540	$8442
Hospitalizations	0.09	0.13
Hospital days	0.50	0.78
Emergency visits	0.19	0.35
Outpatient visits	4.94	4.31
Mean biomarker values		
HbA1c (%)	5.9	6.4
Systolic BP (mmHg)	126.6	130.3
Diastolic BP (mmHg)	75.5	75.7
Creatinine (mg/dl)	0.89	0.98
Hematocrit (%)	40.7	37.8
LDL (mg/dl)	103.4	103.0
Active chronic illnesses (comorbidities)		
Total number of active illnesses	1.20	1.90
Hypertension	0.29	0.44
Diabetes, uncomplicated	0.08	0.22
Arrythmia	0.09	0.08
Hypothyroid	0.09	0.05
Obesity	0.07	0.18
Pulmonary disease	0.07	0.11
Cancer	0.07	0.06
Depression	0.06	0.08
Anemia	0.05	0.10
Arthritis	0.04	0.04
Renal failure	0.03	0.07
Electrolyte disorder	0.03	0.05
Heart failure	0.03	0.05
Psychosis	0.03	0.05
Valvular disease	0.03	0.02
Stroke	0.02	0.03
Peripheral vascular disease	0.02	0.02
Diabetes, complicated	0.02	0.07
Heart attack	0.01	0.02
Liver disease	0.01	0.02

- For this study, work with large academic hospital.
- 6079 patients who self-identified as Black and 43,539 patients who self-identified as White.
- Full algorithmic details available.
- Study differences between White and Black patients.
- First, calculate an overall measure of health status widely accepted in the field: number of active chronic conditions.

Dissecting racial bias in an algorithm used to manage the health of populations



Inform general practicioner
Auto-enrollment in special program

- At same level of algorithm predicted risk, Black people many m ore active chronic conditions.
- In other words, Black person has to be much more sick to be considered for the program.
- Simulate algorithm with no predictive gap between Blacks and Whites = without bias (replacing supramarginal healthier White with inframarginal sicker Black).
- At all risk thresholds above the 50th percentile, removing bias would increase the fraction of Blacks, from 17.7% to 46.5% at the 97th percentile.

Dissecting racial bias in an algorithm used to manage the health of populations

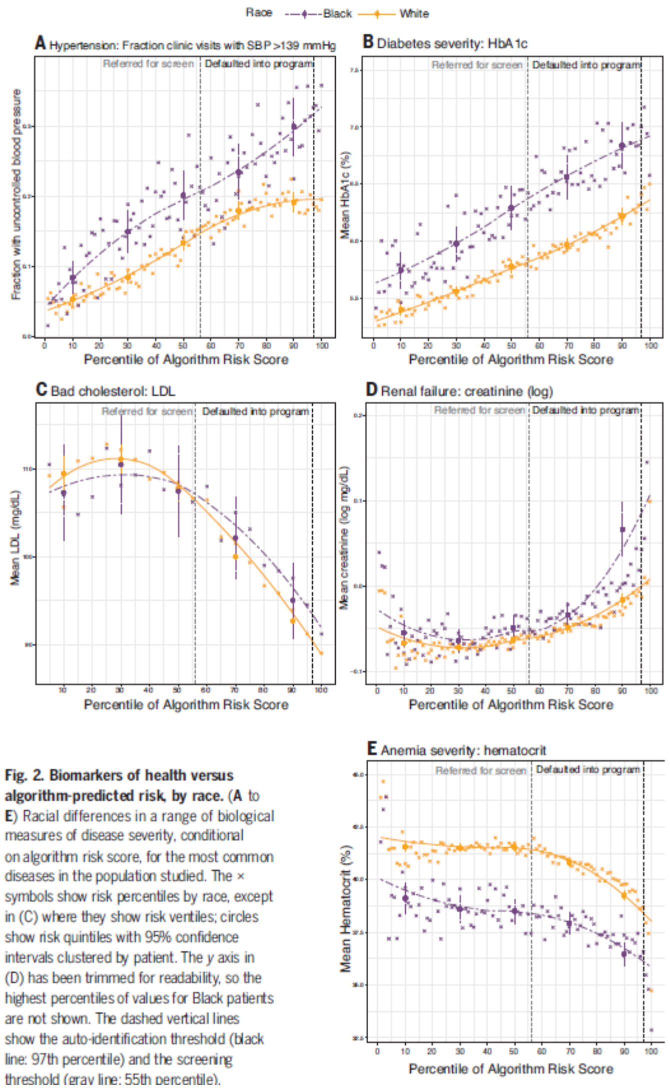
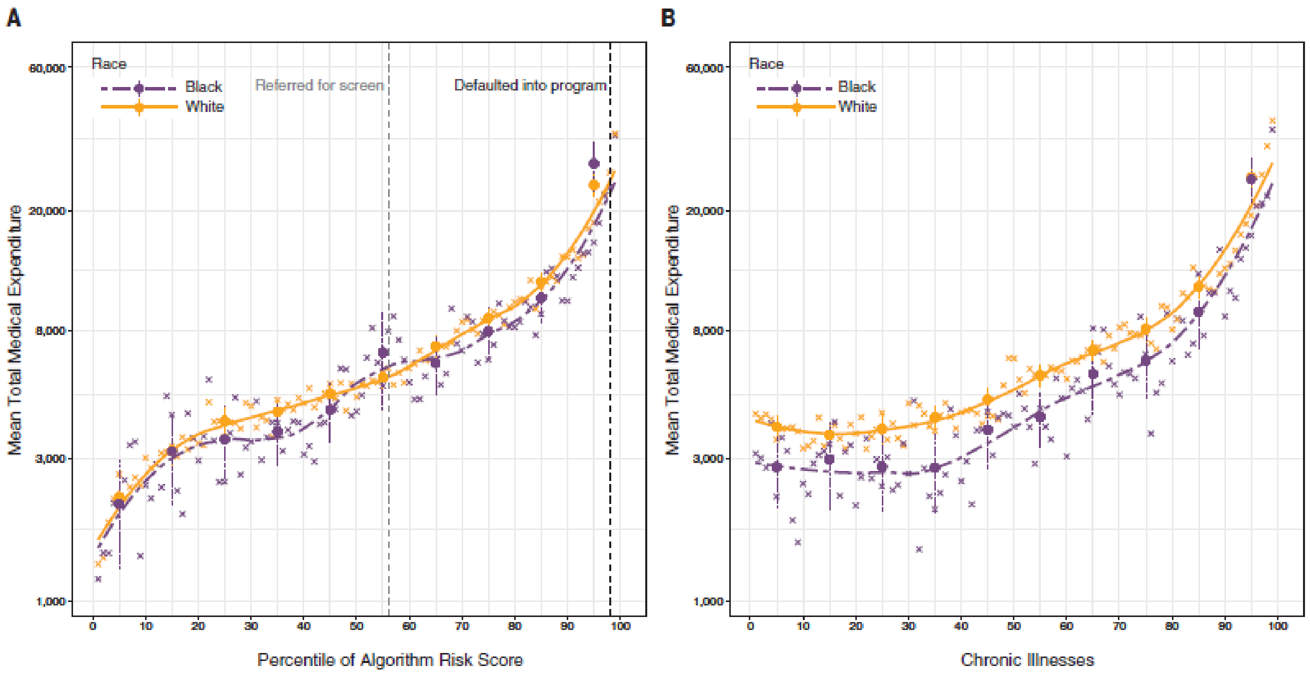


Fig. 2. Biomarkers of health versus algorithm-predicted risk, by race. (**A** to **E**) Racial differences in a range of biological measures of disease severity, conditional on algorithm risk score, for the most common diseases in the population studied. The × symbols show risk percentiles by race, except in (C) where they show risk ventiles; circles show risk quintiles with 95% confidence intervals clustered by patient. The *y* axis in (D) has been trimmed for readability, so the highest percentiles of values for Black patients are not shown. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile).

- Check scores for most common chronic illnesses separately.
- At any predicted risk level, Black people were significantly more sick for all biomarkers used.

Mechanism of bias



- In their setting, algorithm takes in a large set of raw insurance claims data of the previous year:
  - Age, sex
  - Insurance type
  - Diagnosis and procedure codes
  - Medications
  - Detailed costs
  - Algorithm specifically exludes race
- Algorithms prediction on health needs is, in fact, a prediction on health costs.
- At every level of predicted risk, Blacks and Whites effectuate same costs the following year. → Algorithm's predictions are well-calibrated across races (**A**).
- But much less money is spent on Blacks with the same disease severity than Whites (**B**).

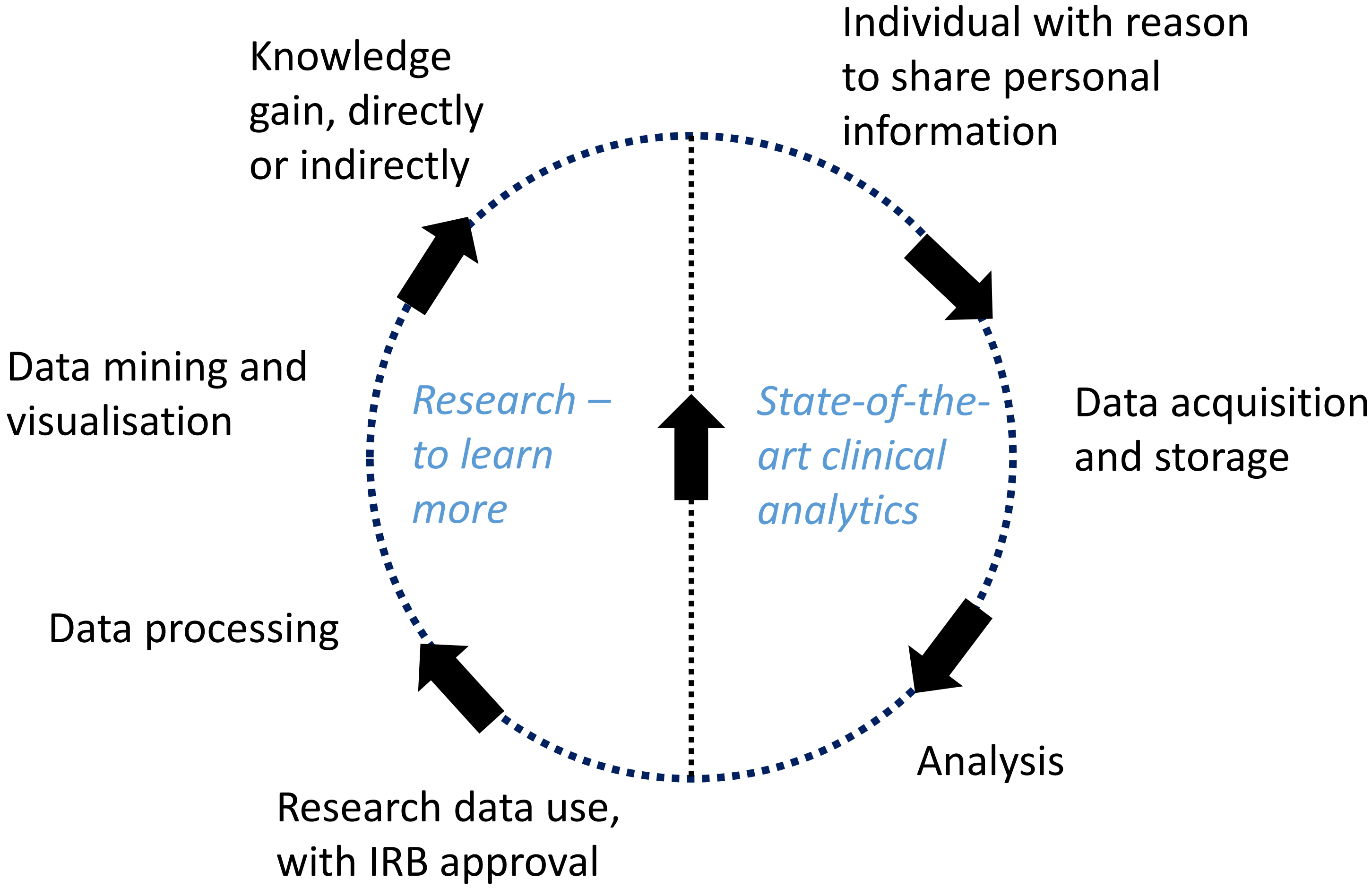
Mechanism of bias

- Results suggest that driving force behind bias is that Black patients generate lesser medical expenses, conditional on health.
- This leads to a racial health bias when accurately predicting costs.
- Poor patients face substantial barriers to accessing health care, even when enrollen in insurance plans.
- Race could have a direct effect (e.g. doctor-patient relationship) so that Black patients are less prone to seek medical aid (reduced trust in health care system).

Take home

- Importance of choice of label on which algorithm is trained.
- Labels are often measured with errors that reflect structural inequalities (ethnic minorities, only male patients, …).
- Careful choice can allow us to enjoy the benefit of algorithmic predictions while minimising their risks.

Setting up of big medical data study



Privacy, data governance

Proper analytics

Evaluating the impact on society

Thank you