Computing for Medicine

Monsoon 2025
Lecture 1

Today's Menu

Introduction	Welcome and Introductions
Learning Objectives	What you are expected to learn
Course Structure	Major Topics covered in this course
The Big Picture	Health Systems: Nuts, Bolts and Abstract

Course Structure

- **The Big Picture**
  - Computing and Healthcare
- **Data**
  - Open & FAIR Data
  - Syntactic Interoperability Standards (HL7, DICOM, FHIR)
  - Semantic Interoperability Standards (SNOMED, Ontologies)
- **Data Science, ML & AI**
  - Structured data - Statistics, Applied Machine Learning
  - Unstructured data - Natural Language Processing in Healthcare
  - **New Topic: Agentic AI in Healthcare**
- **Case Studies with Open Data**
  - Tindering Open Datasets
  - Antibiotic Resistance and its Trajectory Analysis
  - Predictive Modeling for Intensive Care Units
  - Issues with Inclusion, Safety and Ethics of Healthcare AI

Post conditions

You will

- build interoperable data exchange solutions for medicine and public health
- understand and critique existing modeling methods for medical and public health data
- produce models for medical and public health scenarios
- develop new interventions for real world healthcare challenges

Pedagogy

- Learning
  - **In-person** Lectures
  - Class participation: questions, discussions, activities
  - Online platform: Google Classroom
- Evaluation
  - Exams
  - Assignments
  - Quizzes (n-1 policy for quizzes applies)
- Learning by doing
  - Course project
  - 60 mins - 20 mins (in class discussions)

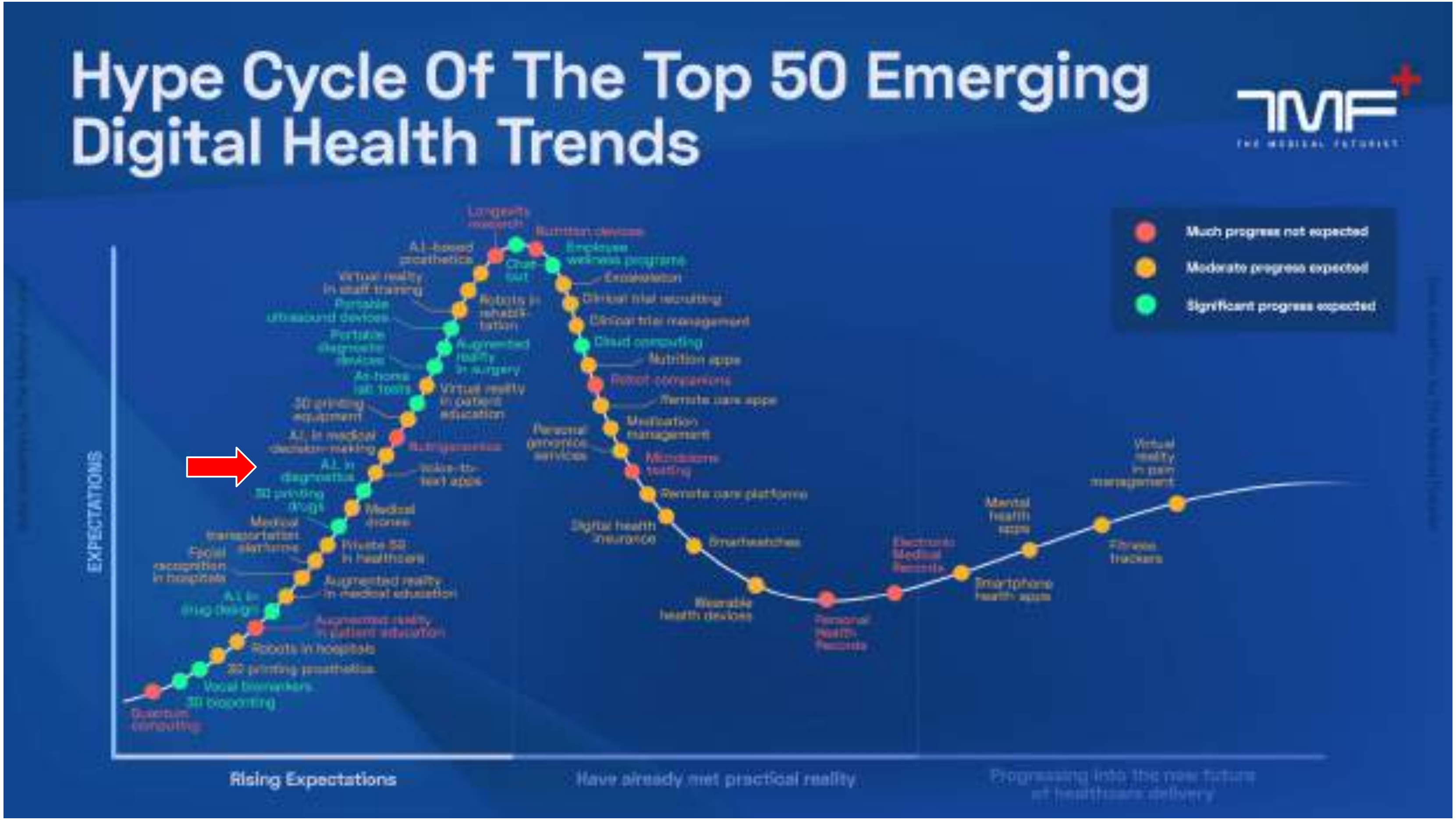
Grading

- Assignments: 15%
- Quizzes: 15%
- Mid sem Exam: 20%
- End sem Exam: 25%
- Project: 25%

Plagiarism

- What is it?
- Copying HWs
- Any content taken from another source without citation
- First time caught, zero on that submission
- Second time caught will be directed to academic committee
- Caught in final report, one grade lower
- Whatever policy from IIITD

Computing For Medicine: The Big Picture



https://medicalfuturist.com/healthcare-trends-curve/ March 11, 2025

What Do You See?

Computing for Medicine is Highly Interdisciplinary!

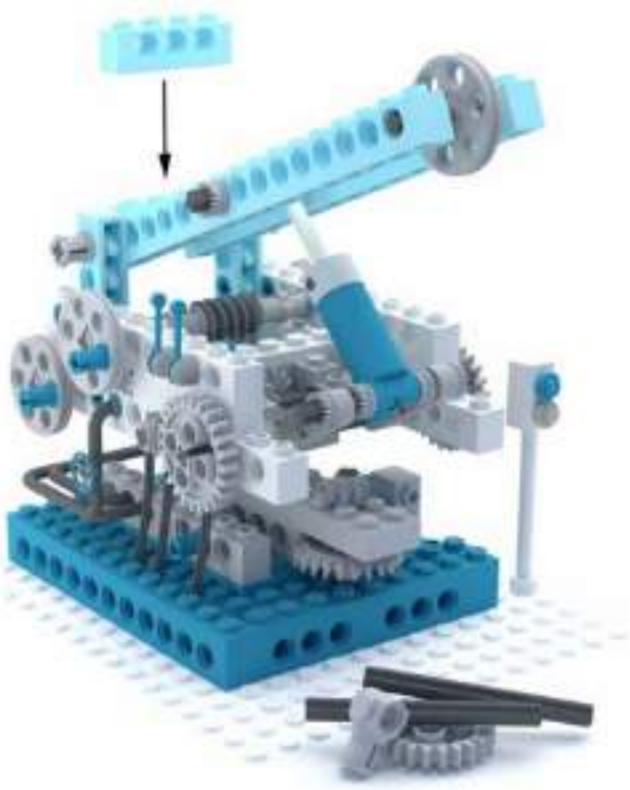
Some related programs and courses:-

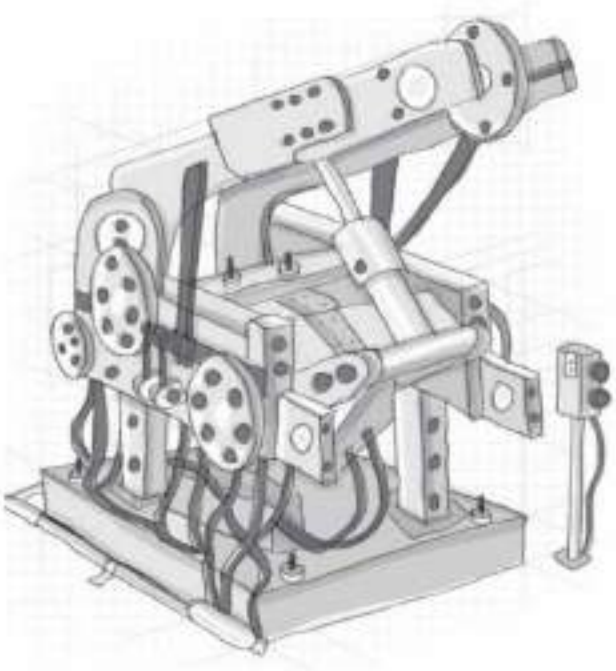
- Biomedical Informatics (Stanford University, Harvard University, UT Houston)
- Health Informatics (Melbourne Medical School)
- Computational Bioscience (University of Colorado, School of Medicine)
- Biomedical Data Science (Stanford University)
- Artificial Intelligence in Healthcare (Stanford University)
- Clinical and Translational Science (University of Vermont, USA)

C4M encourages polymathic learning by collaboration.

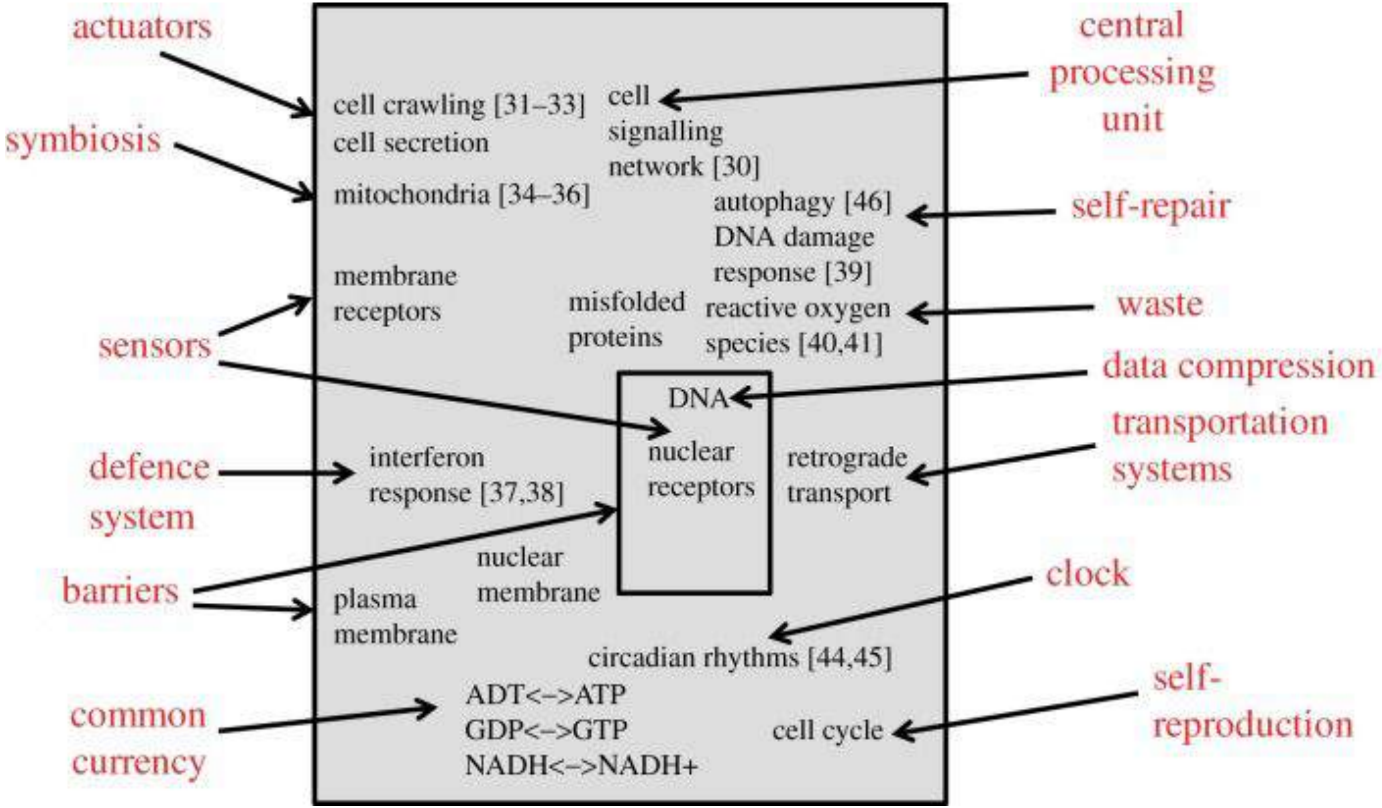
"Biology is the study of the complex things in the Universe. Physics is the study of the simple ones."

- Richard Dawkins





Human Cell: A Prototypical Complex System



Key characteristics of biological systems

Biological systems are:-

- Complex Adaptive Systems
- Exhibit Self Organizing behavior
- Are dynamically changing
- Have oscillatory properties
- Are homeostatic and homeokinetic
- Operate far from equilibrium
- Unpredictably sensitive
- Are not always optimum

Published: 12 February 2004

An optimal bronchial tree may be dangerous

B. Mauroy, M. Filoche, E. R. Weibel & B. Sapoval

Nature 427, 633–636(2004) | Cite this article

345 Accesses | 172 Citations | 0 Altmetric | Metrics

Abstract

The geometry and dimensions of branched structures such as blood vessels or airways are important factors in determining the efficiency of physiological processes. It has been shown that fractal trees can be space filling[1] and can ensure minimal dissipation[2,3,4]. The bronchial tree of most mammalian lungs is a good example of an efficient distribution system with an approximate fractal structure[5,6]. Here we present a study of the compatibility between physical optimization and physiological robustness in the design of the human bronchial tree. We show that this physical optimization is critical in the sense that small variations in the geometry can induce very large variations in the net air flux. Maximum physical efficiency therefore cannot be a sufficient criterion for the physiological design of bronchial trees. Rather, the design of bronchial trees must be provided with a safety factor and the capacity for regulating airway calibre. Paradoxically, our results suggest that bronchial malfunction related to asthma is a necessary consequence of the optimized efficiency of the tree structure.

How are Health Systems Different?

The Why, What, How? And Who Cares?

The Why? Questions "they" care about

- improving the health of populations through better
  - screening
  - diagnosis
  - treatment
  - prognosis
- reducing the per capita cost of healthcare
- improving the patient experience including quality and satisfaction
- Supporting healthcare providers by providing augmented decision support
- However, healthcare AI needs safety, interpretability

What? Same Data Different Purposes

Example:
Two recent studies have looked at the potential link between pioglitazone and bladder cancer.

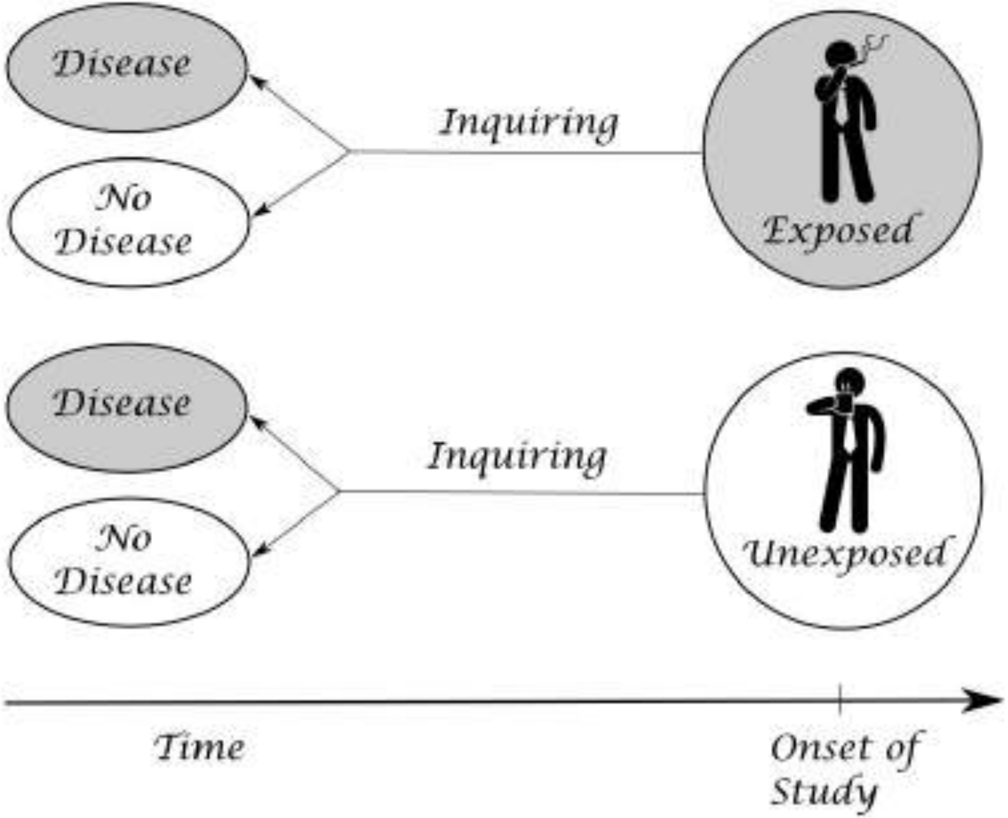
- In 2015, researchers used prescription data, cancer and mortality rates from people with Type 2 diabetes across six different regions from four different countries (three of the groups came from the UK) and did not find any evidence that pioglitazone increased the risk of bladder cancer.
- In 2016, a different study looked at nearly 150,000 people who were treated with antidiabetic drugs between 2000 and 2013, and followed up until 2014. This work did identify an increased risk of bladder cancer with the use of pioglitazone.

https://ourhealthydata.eu/how-is-health-data-used

How? Data Science Approaches

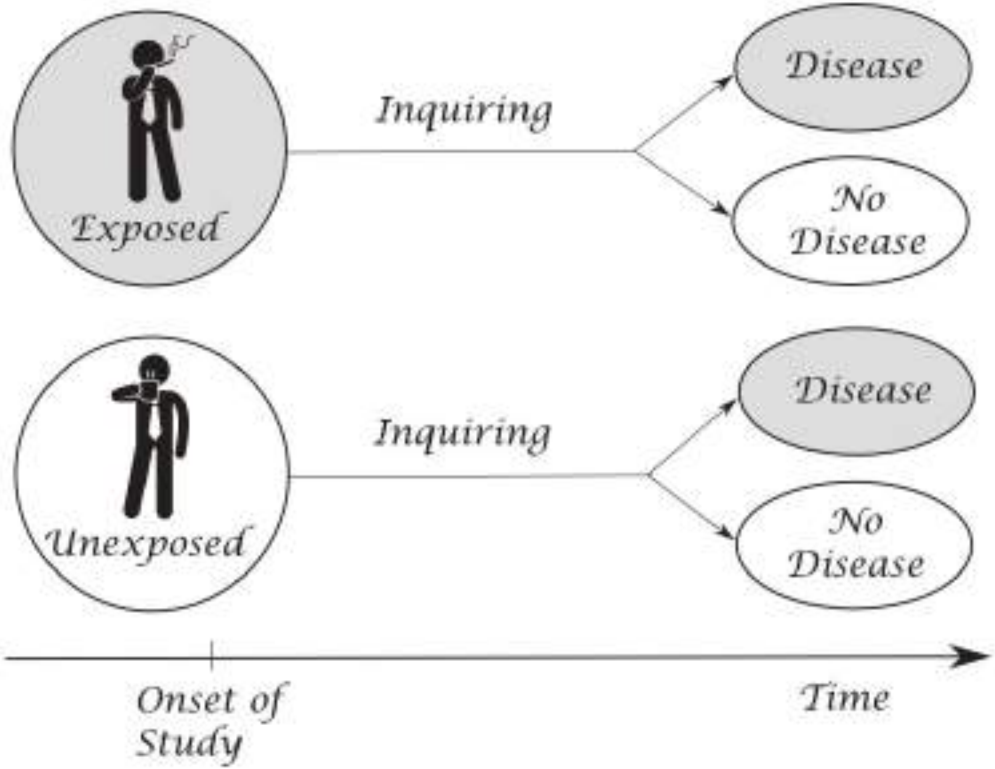
How were the Data Generated?

Example of Retrospective Study Design



How were the Data Generated?

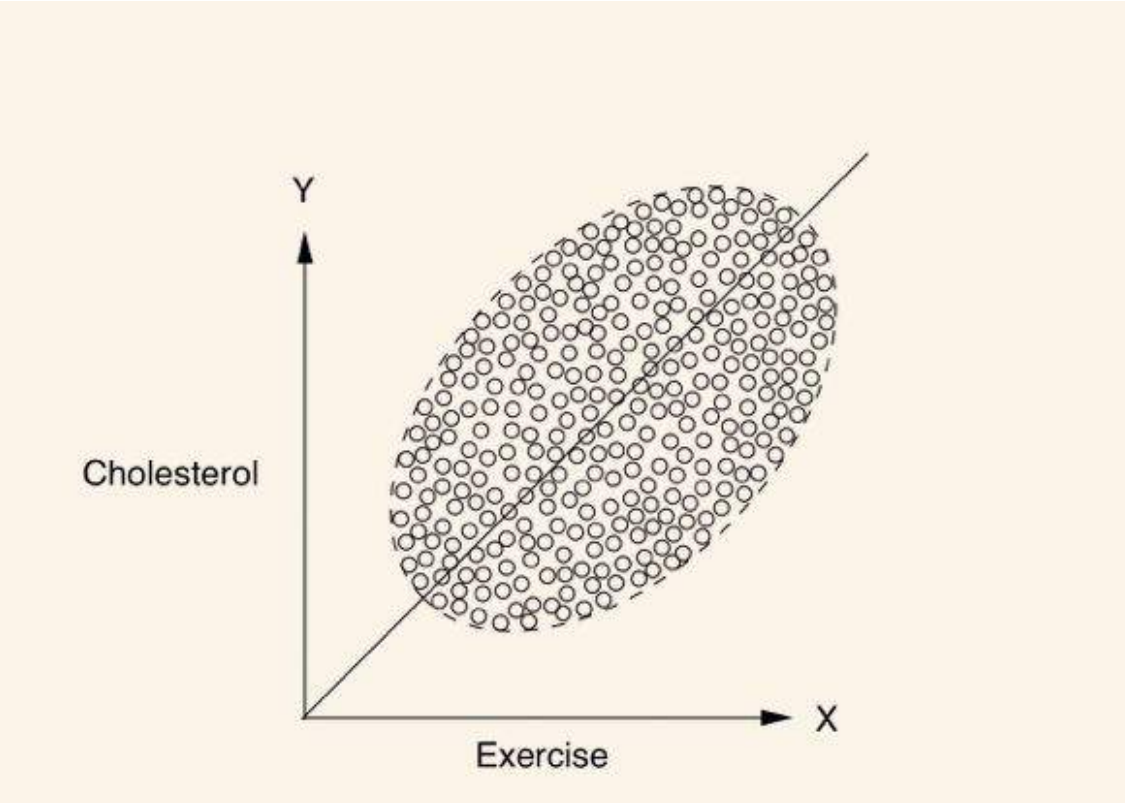
Example of Prospective Study Design



How? Need for Novel Data Science Approaches

You are a health policy maker:

How do you interpret this data?



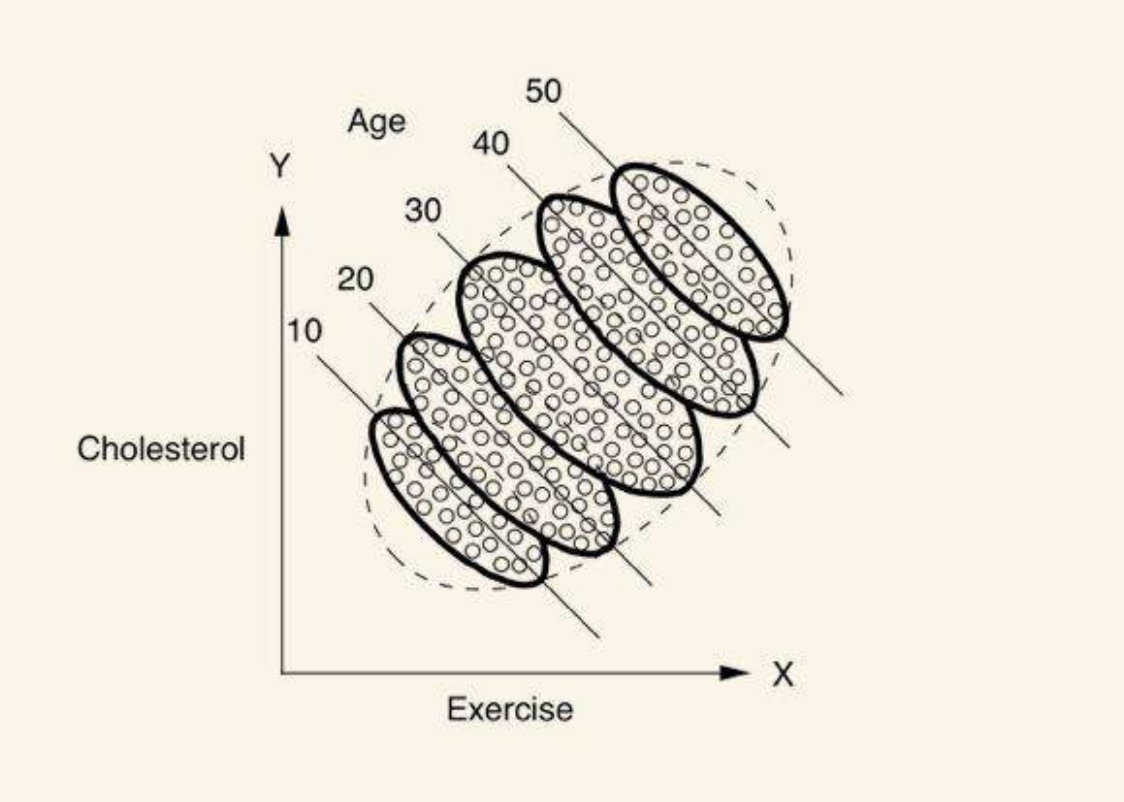
http://bayes.cs.ucla.edu/PRIMER/

How? Need for Novel Data Science Approaches

Findings on sub-groups of data may be opposite to that on the whole data!

Q: What is this phenomenon called?

Simpson's Paradox



http://bayes.cs.ucla.edu/PRIMER/

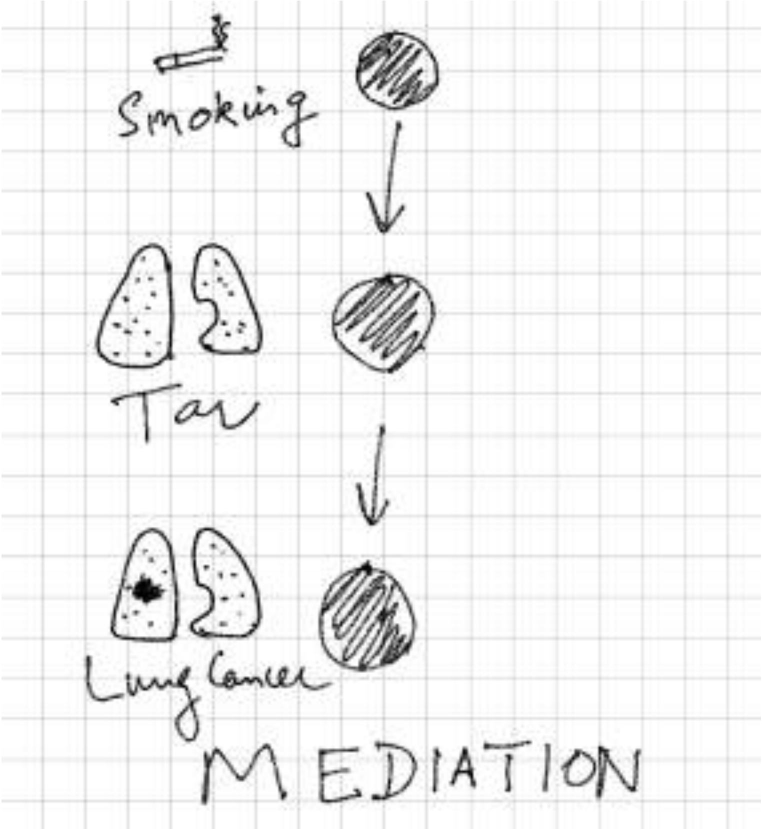
Israeli Data: August 15,2021

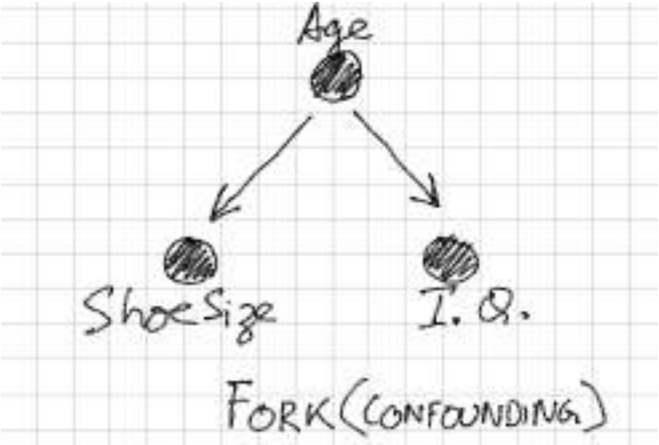
From: https://datadashboard.health.gov.il/COVID-19/general

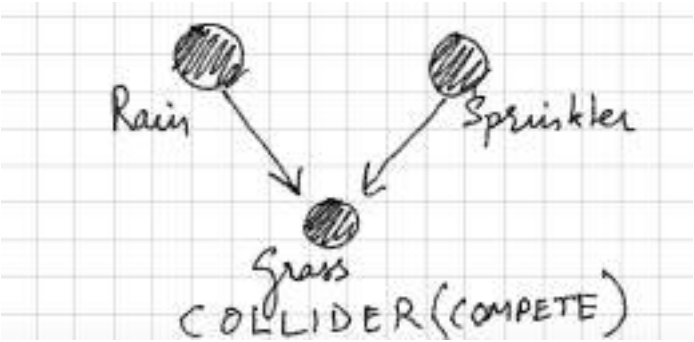
Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,170,563 90.4%	171 90.9	290 13.6	**85.2%**

This strange result illustrates something called **Simpson's Paradox**, in this case meaning you can have **very high efficacy in each group**, but the **overall efficacy looks much lower** because one group (older people) is ***more vaccinated*** and have a ***much higher risk of severe disease***.

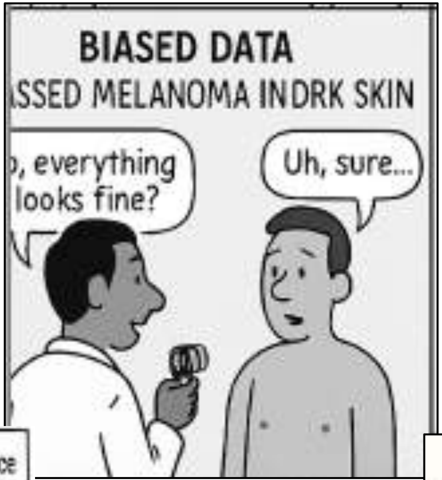
World of Conditional Probabilities

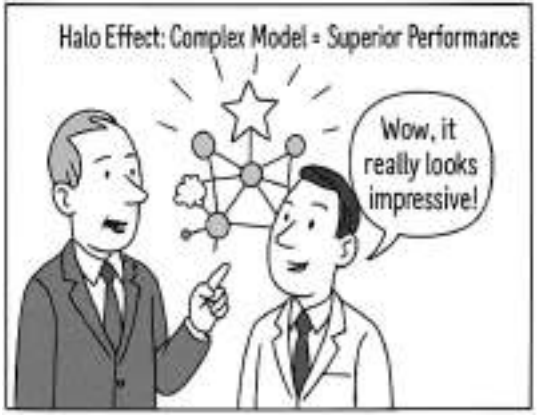






AI Failure Can Be of Many Types





Example: Classical Statistical Fail

RESEARCH-ARTICLE

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Authors: Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad Authors Info & Claims

KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining • Pages 1721 - 1730
https://doi.org/10.1145/2783258.2788613

Published: 10 August 2015 Publication History

Check for updates

Caruana, R., et al. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721–1730)*. Association for Computing Machinery.

Confounding

ing [1]. Although models based on rules were not as accurate as the neural net models, they were *intelligible*, i.e., interpretable by humans. On one of the pneumonia datasets, the rule-based system learned the rule "HasAsthama(x) ⇒ LowerRisk(x)", i.e., that patients with pneumonia who have a history of asthma have lower risk of dying from pneumonia than the general population. Needless to say, this rule is counterintuitive. But it reflected a true pattern in the training data: patients with a history of asthma who presented with pneumonia usually were admitted not only to the hospital but directly to the ICU (Intensive Care Unit).

Open Discussion (5 Minutes)

Who Benefits from Health Data?

Computing for Medicine Starts With Data!

Making Healthcare Data Science Reproducible and FAIR

To be **Findable**:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be **Accessible**:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be **Interoperable**:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be **Reusable**:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Subjects Clear

- [x] COVID-19 981
- [ ] Coronavirus 352
- [ ] SARS-CoV-2 283
- [ ] coronavirus 237
- [ ] covid-19 180
- [ ] Covid-19 148
- [ ] covid 141
- [ ] Coronaviridae 120
- [ ] pathogens 119
- [ ] CETAF-taskforce 118

File type

- [ ] PDF 4,293
- [ ] BIN 399
- [ ] ZIP 145

March 16, 2020 (2020-05-01) | Dataset | Open

COVID-19 Open Research Dataset (CORD-19)

Sebastian Kohlmeier; Kyle Lo; Lucy Lu Wang

Important: This dataset is updated regularly and the latest version for download can be found here. In response to the COVID-19 pandemic, the Allen Institute for AI has partnered with leading research groups to prepare and distribute the COVID-19 Open Research Dataset (CORD-19), a free resource of scholarly articles, including full text conten...

Uploaded on May 7, 2020

7 more versions exist for this record

13243 | 7422

April 11, 2020 (v1) | Preprint | Open

Immunological mechanisms explaining the role of IgE, mast cells, histamine, elevating ferritin, IL-6, D-dimer, VEGF levels in COVID-19 and dengue, potential treatments such as mast cell stabilizers, antihistamines, Vitamin C, hydroxychloroquine, ivermectin and azithromycin

Arumugham, Vinu

A novel coronavirus, SARS-CoV-2 was identified in Wuhan, China. The disease caused by the virus can range in severity from asymptomatic to acute respiratory distress syndrome (ARDS) and death. Primary dengue infection results in IgE mediated sensitization against dengue virus (DENV) proteins. These IgE bind to receptors on mast cells. Upon sub...

Uploaded on April 11, 2020

13229 | 7445

November 13, 2020 (2020-10-20) | Journal article | Open

COVIPENDIUM: information available to support the development of medical countermeasures and interventions against COVID-19

Martine DENIS; Valerie VANDEWEERD; Rein VERBEEKE

https://zenodo.org/search?q=coronavirus&f=subject%3ACOVID-19&l=list&p=1&s=10&sort=mostviewed

JOHNS HOPKINS
UNIVERSITY & MEDICINE

Welch Medical Library

Welch Medical Library Home / Welch Medical Library Guides / Finding Datasets for Secondary Analysis / Research Data Repositories & Databases

Finding Datasets for Secondary Analysis

Enter Search Words | Search

- About This Guide
- New to Hopkins?
- Research Data Repositories & Databases
  - NIH Data Repositories
  - Examples of NIH Data Repositories
  - Other Data Repositories/Consortium

NIH Data Repositories

In general, NIH does not endorse or require sharing in any specific repository and encourages researchers to select the repository that is most appropriate for their data type and discipline (though such specification does exist for particular initiatives). To help researchers locate an appropriate resource for sharing their data, as well as to promote awareness of resources where datasets can be located for reuse, Trans_NIH BioMedical Informatics Coordinating Committee (BMIC) maintains lists of several types of data sharing resources:

- Open NIH-supported domain-specific repositories that house data of a specific type or related to a specific discipline;
- Other NIH-supported domain-specific resources, including repositories and knowledgebases, that have limitations on submitting and/or accessing data; and
- Generalist repositories that house data regardless of type, format, content, or subject matter.

https://browse.welch.jhmi.edu/datasets/nih-data-repositories

Hopkins Initiatives

- COVID-19 Precision Medicine Analytics Platform Registry (JH-CROWN)
  The main data source is Johns Hopkins' electronic medical record, Epic. The registry is refreshed weekly with new and updated data and is available for Johns Hopkins investigators to analyze subsets of the COVID-19 patient population for retrospective analyses. CHSOR members Dr. Jodi Segal and Dr. Caleb Alexander have been using these data.
- COVID-19-specific Common Data Model
  PCORnet®, the National Patient-Centered Clinical Research Network, is creating a COVID-19-specific Common Data Model that will allow the use of information from patients across PCORnet's network At Johns Hopkins, Dr. Harold Lehmann leads these activities.
- Johns Hopkins COVID-19 Collaboration Platform
  Over 400 such trials have been registered on clinicaltrials.gov with dozens being added each day. Many of them are designed to answer similar questions and combining data or aggregating evidence could dramatically increase their efficiency and precision, getting answers to doctors faster and more reliably. *more...*
- National COVID Cohort Collaborative
  The National COVID Cohort Collaborative is the partnership among the NCATS-supported Clinical and Translational Science Awards (CTSA) Program hubs and the National Center for Data to Health (CD2H). At Johns Hopkins, Dr. Christopher Chute has taken the lead on this activity.

The Healthcare Data Revolution -> Transformation

- 4CE: Consortium for Clinical Characterization of COVID-19 by EHR
- Figshare: COVID-19 open data
- GitHub: COVID-19 Open Repo Data
- Harvard Dataverse: COVID-19 Data
- ICPSR: COVID-19 Data Repoasitory
- ImmPort: COVID 19
- Mendeley Data: Elsevier COVID-19 Research Environment
- National COVID Cohort Collaborative (N3C)
  OHDSI: Characterizing Health Associated Risks, and Your Baseline Disease In SARS-COV-2 (CHARYBDIS)
- Open-Access Data and Computational Resources to Address COVID-19.
- OpenSAFELY
  Tableau: COVID-19 Data Hub
- Vivli: Covid data
  Zenodo: Coronavirus Disease Research Community - COVID-19
- COVID-19 Research Database

MIMIC (Medical Information Mart for Intensive Care) Database

Medical Information Mart for Intensive Care III (MIMIC-III) is a large, freely-available database comprising de-identified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

The database includes information such as demographics, vital sign measurements made at the bedside (~1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital).

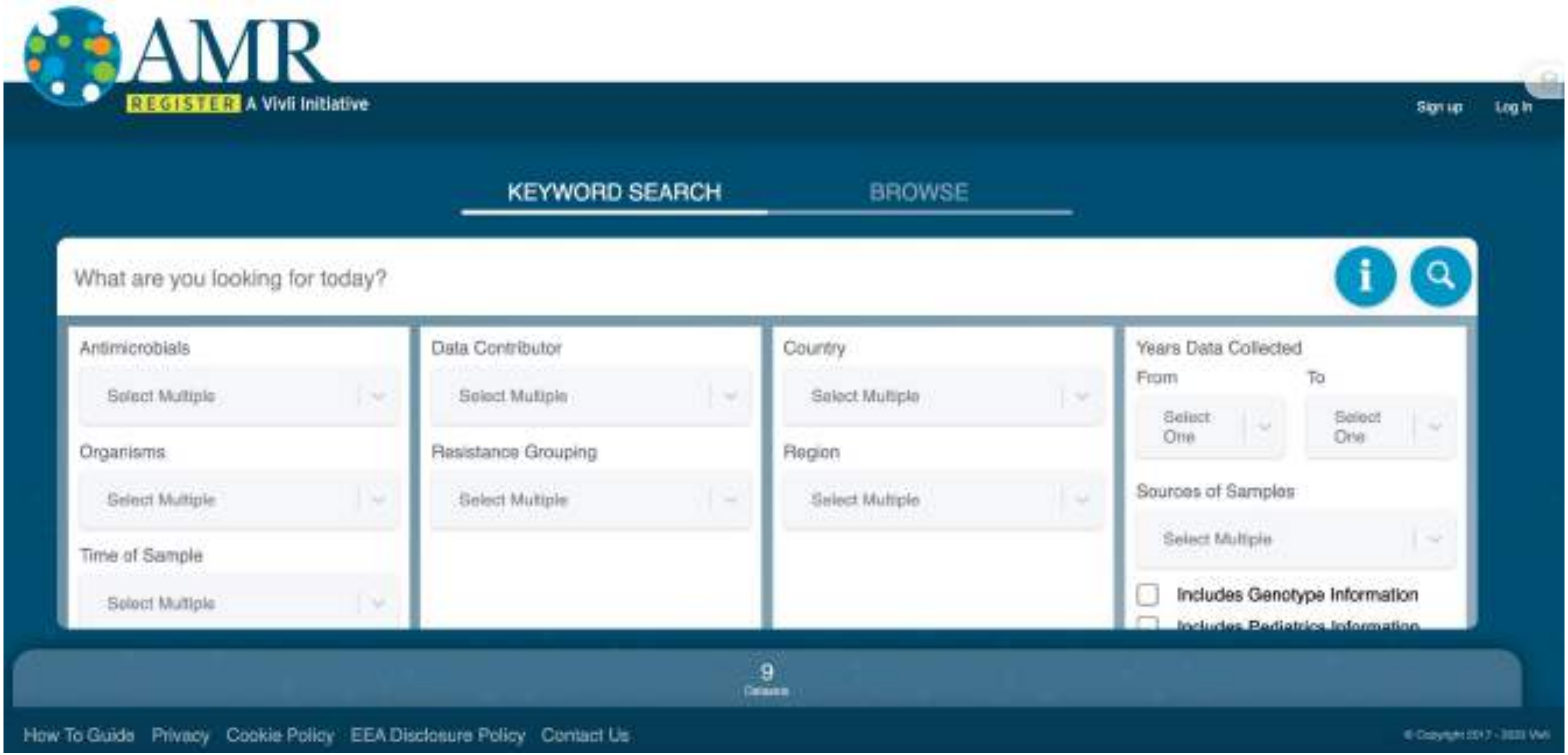
MIMIC-III supports a diverse range of analytic studies spanning epidemiology, clinical decision-rule improvement, and electronic tool development. It is notable for three factors:

- it is freely available to researchers worldwide
- it encompasses a diverse and very large population of ICU patients
- it contains high temporal resolution data including lab results, electronic documentation, and bedside monitor trends and waveforms

https://browse.welch.jhmi.edu/datasets/ehr-databases

Sources of Health Data

Sources of Health Data



Sources of Health Data

Published May 30, 2018 | Version v1.0

Software Open

SAFE-ICU/AMRsteward: stone

Tavpritesh; shubham14101; Tavpritesh@SAFE-ICU

Release version of AMR_Steward, the Bayesian Artificial Intelligence Dashboard for Stewarding Antimicrobial Stewardship at AIIMS, New Delhi, India.

Files

SAFE-ICU/AMRsteward-v1.0.zip

SAFE-ICU/AMRsteward-v1.0.zip

SAFE-ICU-AMRsteward-602c641

AMRsteward.Rproj	356 Bytes
DESCRIPTION	1.2 kB
LICENSE	35.1 kB

https://zenodo.org/records/1255584

Thanks for attending the class!