Problems in ML4H

Day 1 — What we're actually trying to solve

Peter Szolovits · MIT

June 22, 2026

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Why this week exists

Every health AI project lives or dies in one gap:

- an interesting idea → a system that actually helps patients.
- A strong notebook number is not clinical impact.
- The distance between the two is where projects quietly die.
- This week walks that gap end to end — each day closes one segment.

What do Clinicians Do?

- Interact with patient
- Gather information
  - Verbally
  - from instruments
  - lab records, meds
- determine diagnosis
- Plan next steps
- Prescribe drug, treatment, etc.
- Monitor
- Advocacy
- Bureaucratic stuff
- Expert witness
- Documentation

Traditional Roles of Healthcare

- **Prevention** — typical goal of public health
  - Classification into common categories
- **Diagnosis** — what is wrong with the patient?
  - Classification into common categories
  - Planning diagnostic steps
- **Prognosis** — what is likely to happen to the patient?
  - Sets expectations
  - Determines urgency and intensity of needed intervention
- **Therapy**
  - Tradeoff among risks and benefits
  - Compare expected outcomes of different possible therapies
  - Choose the next step and develop expectations for its results

WHO Constitution defines “health”

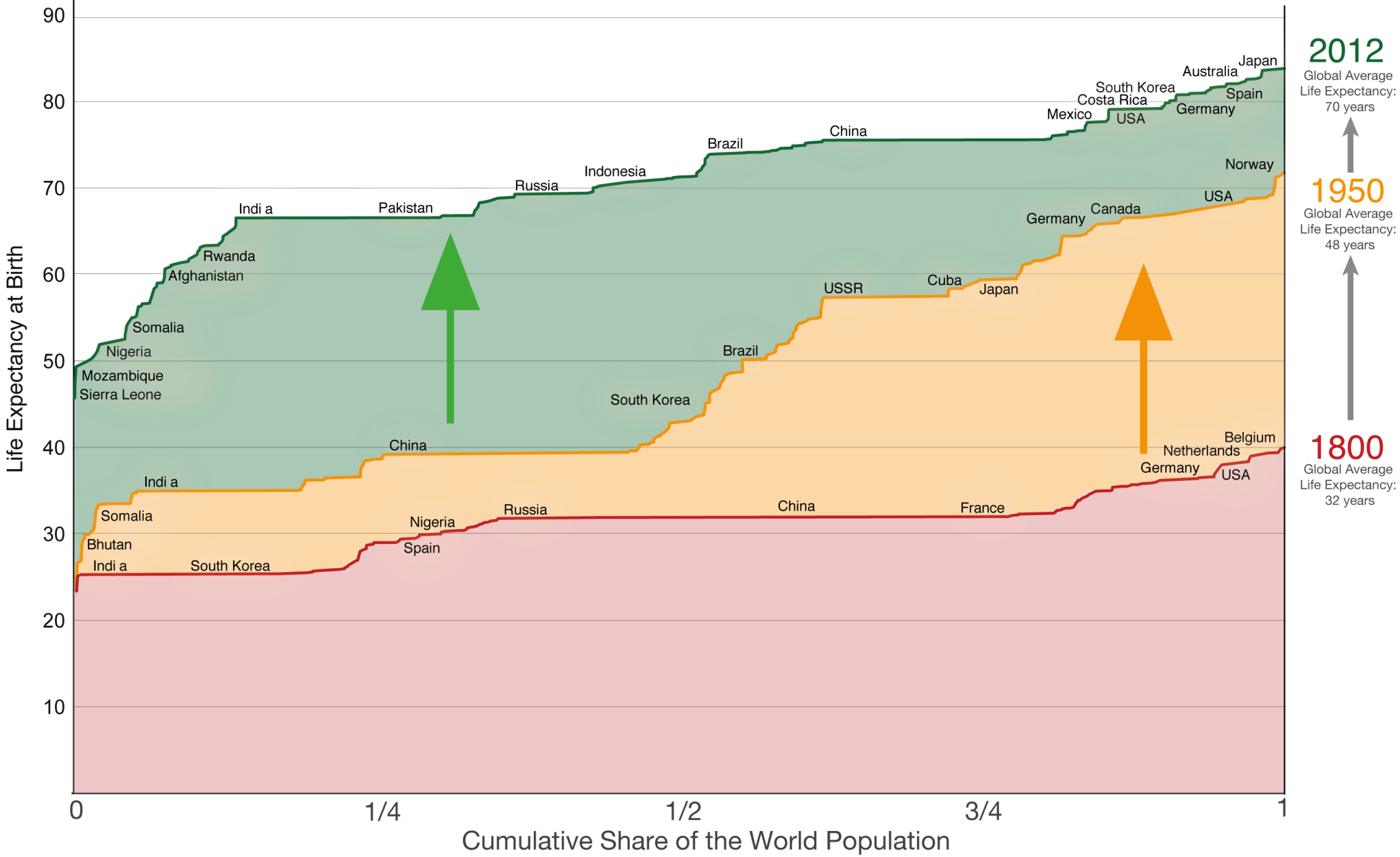
“a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity”

- **Physical**
- **Mental**
- **Social** (very hard to measure)
  - Are more or fewer people better?
  - Equity?
  - Government vs. private costs?

Our World in Data

Life Expectancy of the World Population in 1800, 1950 and 2012

Countries are ordered along the x-axis ascending by the life expectancy of the population. Data for almost all countries is shown in this chart, but not all data points are labelled with the country name.



Data source: The data on life expectancy by country and population by country are taken from Gapminder.org.

The interactive data visualisation is available at OurWorldinData.org. There you find the raw data and more visualisations on this topic.

ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Longevity at birth

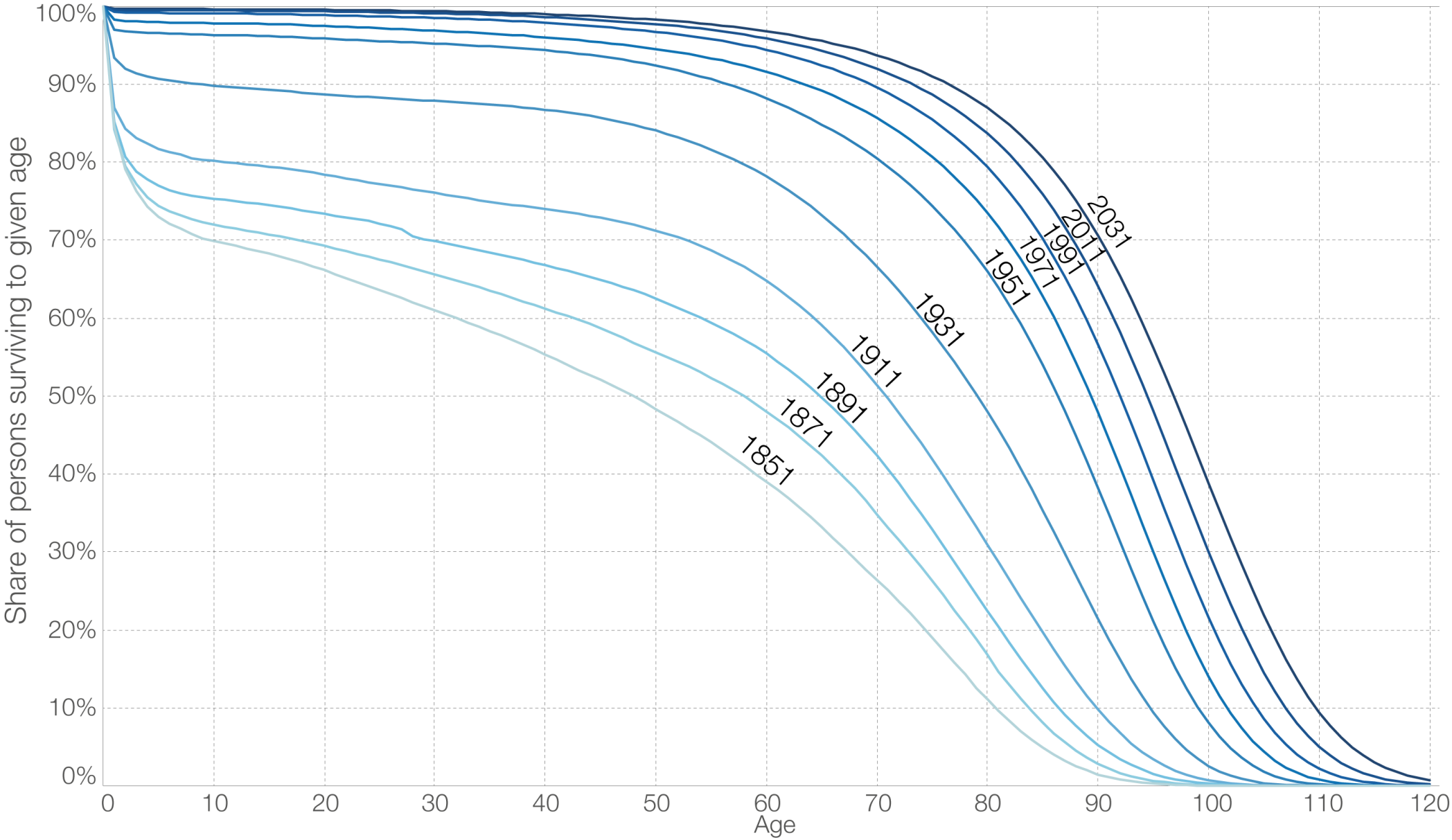
(CIA World Fact Book, 2001, 2024)

Country	*Male*		*Female*	
	2024	*2001*	*2024*	*2001*
Rwanda	64.6	38.4	68.6	39.7
Kenya	68.6	46.6	72.2	48.4
South Africa	70.3	47.6	73.5	48.6
Cambodia	69.6	54.6	73.3	59.1
Brazil	72.6	59.0	80.1	67.7
Russia	67.4	62.1	77.4	72.8
Turkey	74.4	68.9	79.2	73.7
Albania	77.3	69.0	82.8	74.9
USA	78.7	74.4	83.1	80.1
France	79.8	75.0	85.5	83.0
Israel	81.1	76.7	85.1	80.8
Japan	82.3	77.6	88.2	84.2

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Share of persons surviving to successive ages for persons born 1851 to 2031, England and Wales

according to mortality rates experienced or projected, (on a cohort basis)



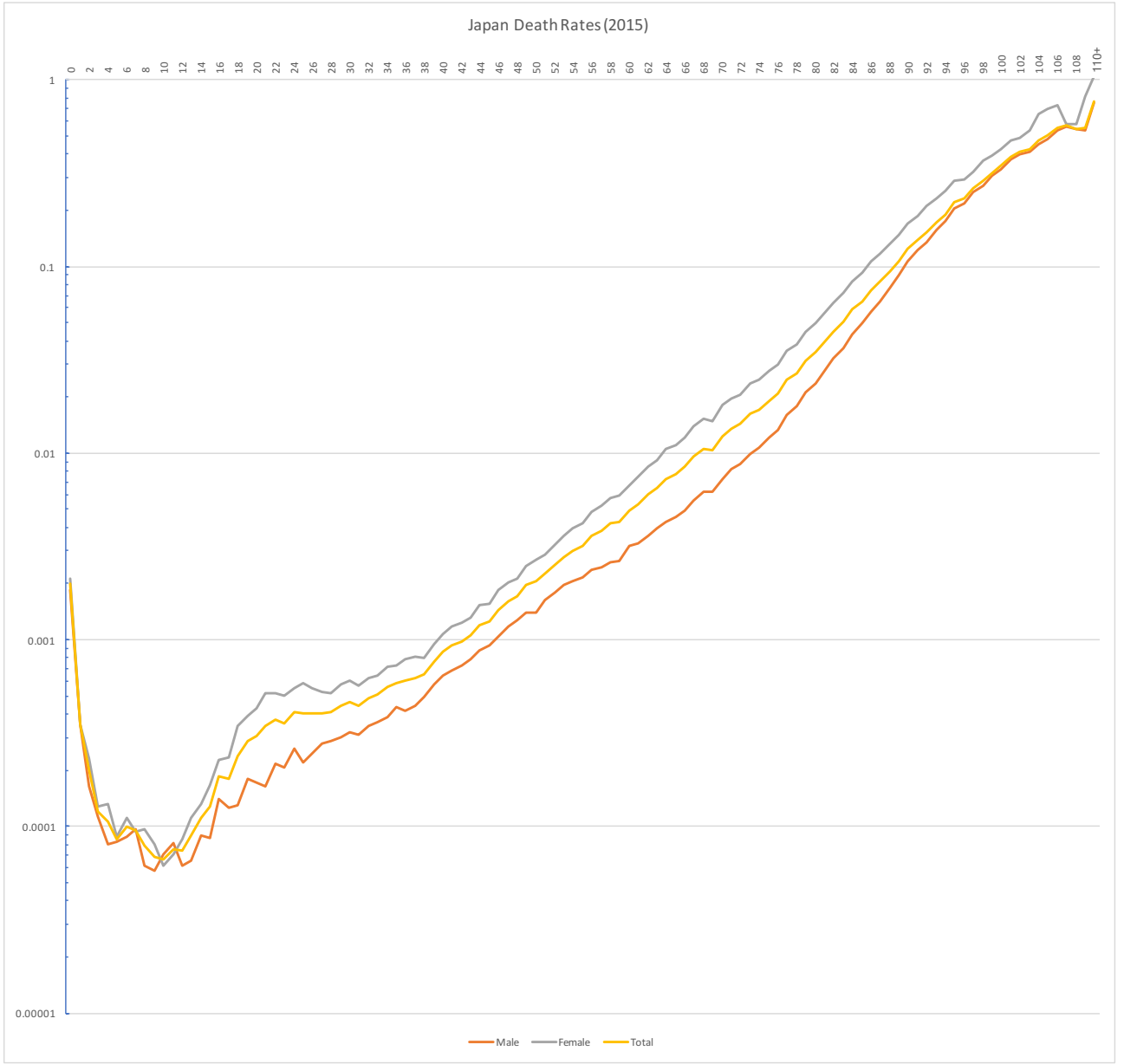
Data source: Office for National Statistics (ONS). Note: Life expectancy figures are not available for the UK before 1951; for long historic trends England and Wales data are used

The interactive data visualization is available at OurWorldinData.org. There you find the raw data and more visualizations on this topic.

Distribution of Death Rates by Age

- Life table deaths by year (Japan, 2015)

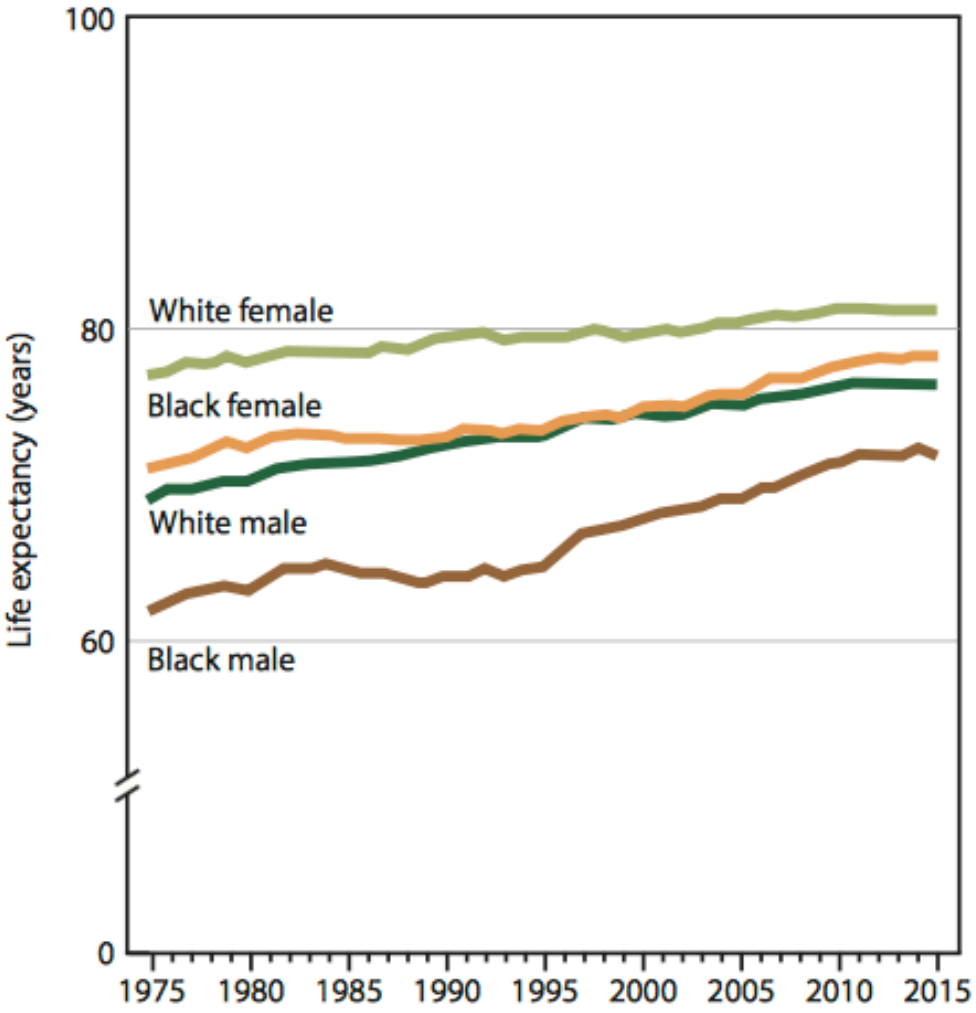
http://www.ipss.go.jp/p-toukei/JMD/00/STATS/Mx_1x1.txt

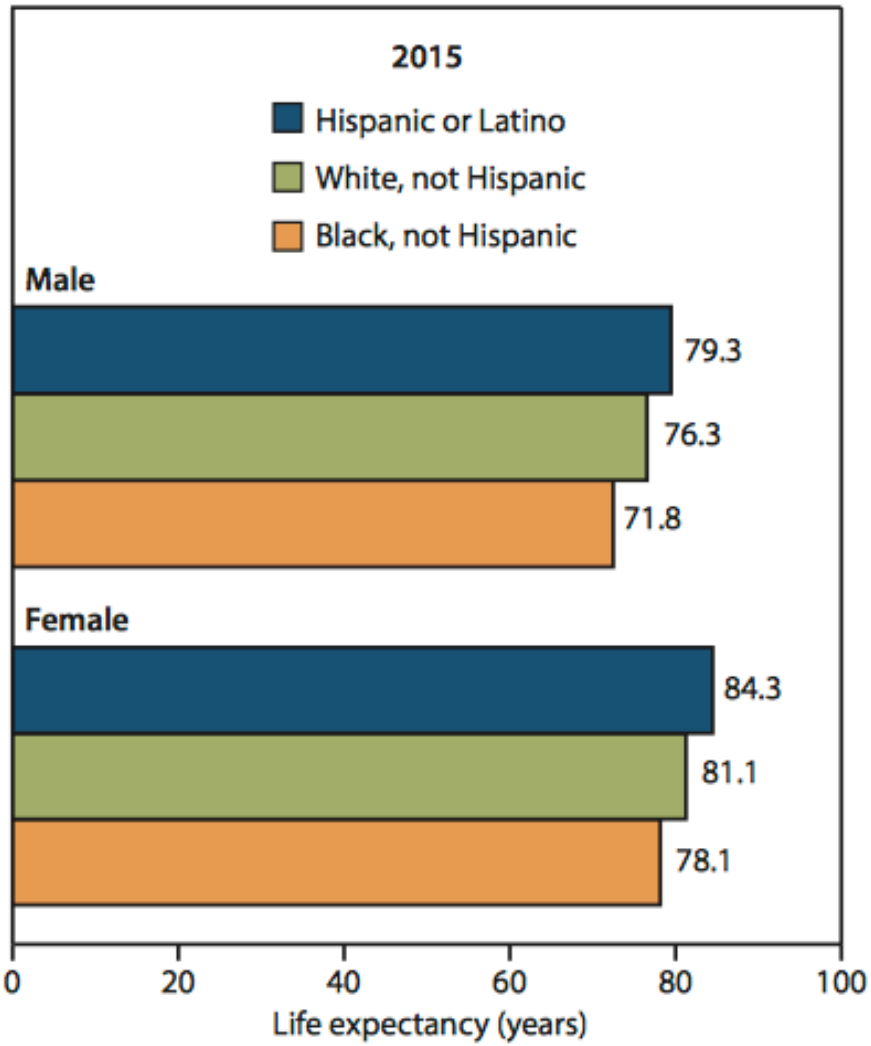


AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Ethnic Disparities

Figure 6. Life expectancy at birth, by sex, race and Hispanic origin: United States, 1975–2015





https://www.cdc.gov/nchs/data/hus/hus16.pdf#019

Causes of death (USA, 2014)

Cause	Deaths/100K	%
Heart disease	192.7	23.4
Cancer	185.6	22.5
Chronic lower respiratory disease	46.1	5.6
Accidents	42.7	5.2
Stroke	41.7	5.1
Alzheimer's disease	29.3	3.6
Diabetes	24.0	2.9
Influenza and pneumonia	17.3	2.1
Kidney disease	15.1	1.8
Suicide	13.4	1.6
OTHER	*215.8*	*26.2*
TOTAL	823.7	100.0

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

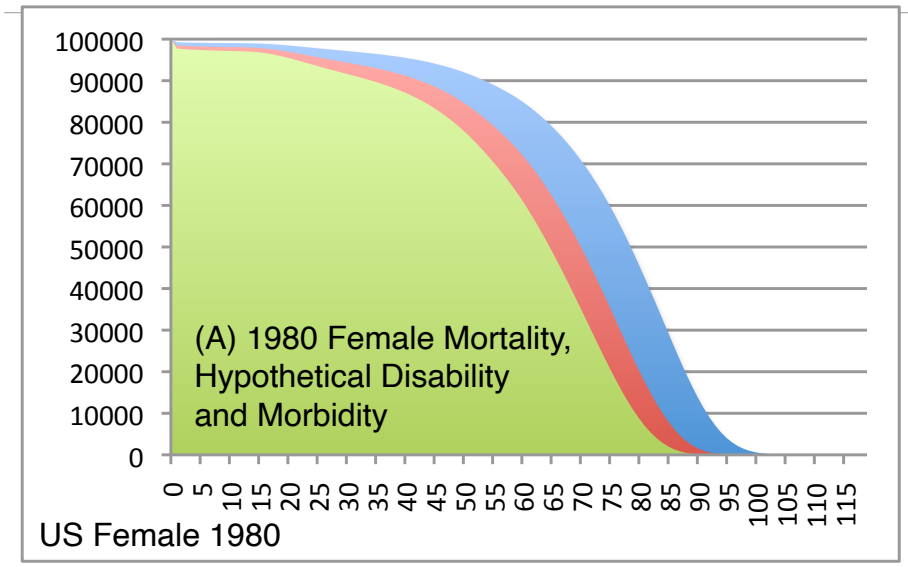
https://www.medicalnewstoday.com/articles/282929.php

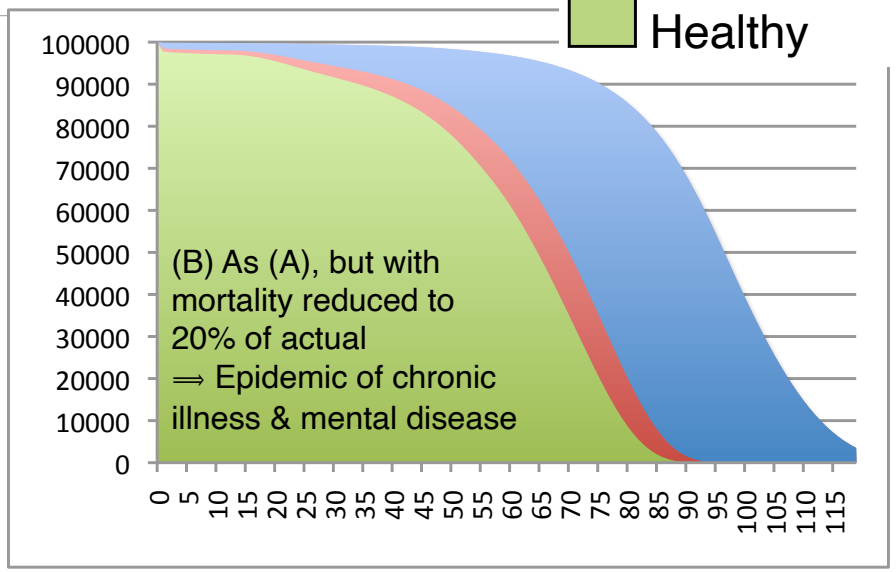
Morbidity: Top 10 Chronic Conditions
Persons aged ≥ 65

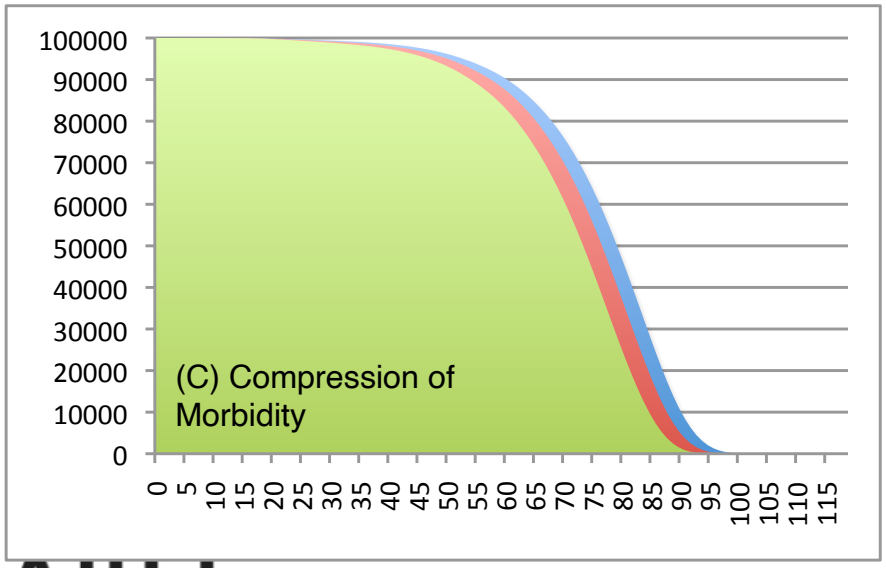
Condition	Both	Male	Female
Arthritis	49.6	40.7	55.7
Hypertension	39.0	33.0	43.2
Hearing impairment	30.0	35.2	26.3
Heart disease	25.7	26.9	24.9
Orthostatic impairment	16.8	15.7	17.8
Cataracts	15.5	11.3	18.4
Chronic sinusitis	15.2	13.7	16.2
Visual impairment	10.1	12.0	8.8
Genitourinary	9.9	11.3	8.9
Diabetes	8.9	7.8	9.7

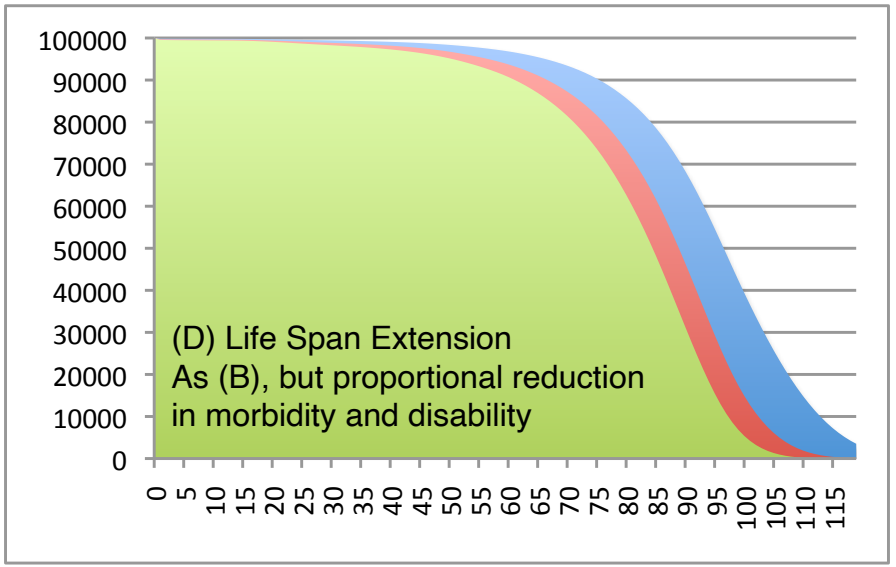
U.S. Nat'l Ctr Health Stat, *Vital and Health Statistics*, 1985 (1982 data)

Mortality, Disability, Morbidity









Mortality
Disability
Morbidity
Healthy

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

The first question:
What *is* ML4H?

Is health AI just "machine learning, applied to hospitals"?

Our central claim

ML4H is not just a domain application of ML — it is a distinct field.

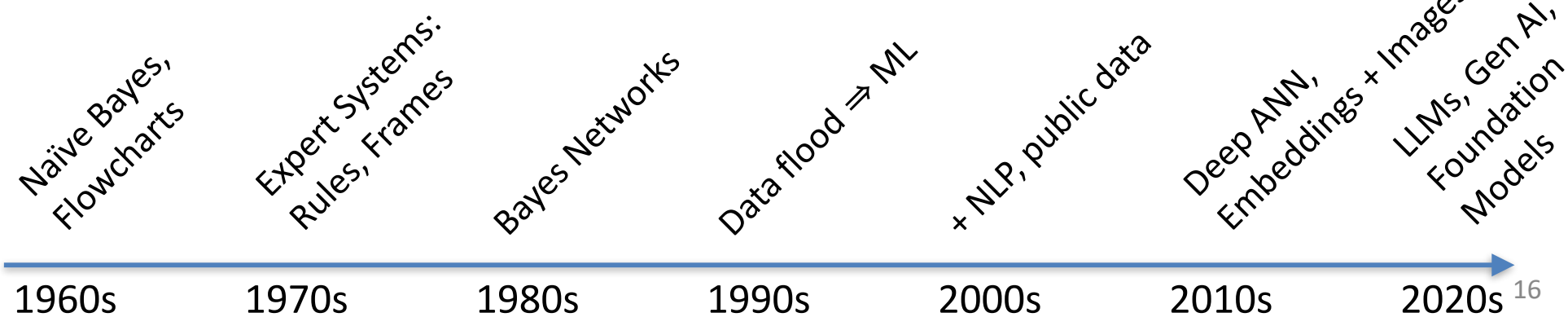
The steelman against this — *"isn't it just sequence/tabular ML with domain quirks?"* — fails on several structural points:

- Outcomes for different patients are **not fungible**.
- **The label is endogenous** to the system you predict: it is produced by the same care process that generates the features.
- **Deployment changes the data-generating process** — prediction in health is performative , not passive (Perdomo et al., 2020).

These aren't quirks: they **complicate identifiability** for causal and decision claims, and violate the **i.i.d./stationarity** assumptions predictive workflows lean on.

Fifty years of perspective

- What is **genuinely new** now — and what is **recycled** under a new name.
- From **MYCIN** and **INTERNIST** to today's transformers, the cycle repeats: each era promised (some) autonomy and delivered, at its best, useful **decision support**.



AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Promise and Problems

- Clinical artificial intelligence (AI) **promises** to offer new abilities in clinical decision support, diagnostic reasoning, precision medicine, clinical operational support, and clinical research.
- In practice, it can be hard to determine how one can **effectively use ML/AI** techniques for real-world problems
  - often explored in **solely academic settings**, without considering the nuances of true clinical deployments
  - clinicians deploying ML algorithms in practice may have different backgrounds and have different stages of **familiarity** with ML methodology and assumptions than those in the research community
  - many **non-technical barriers** exist preventing the widespread use of ML in healthcare, limiting practical examples of its usage

McDermott, Nestor, Szolovits. Clinical Artificial Intelligence: Design Principles and Fallacies. Clin Lab Med . 2023 March ; 43(1): 29–46. doi:10.1016/j.cll.2022.09.004.

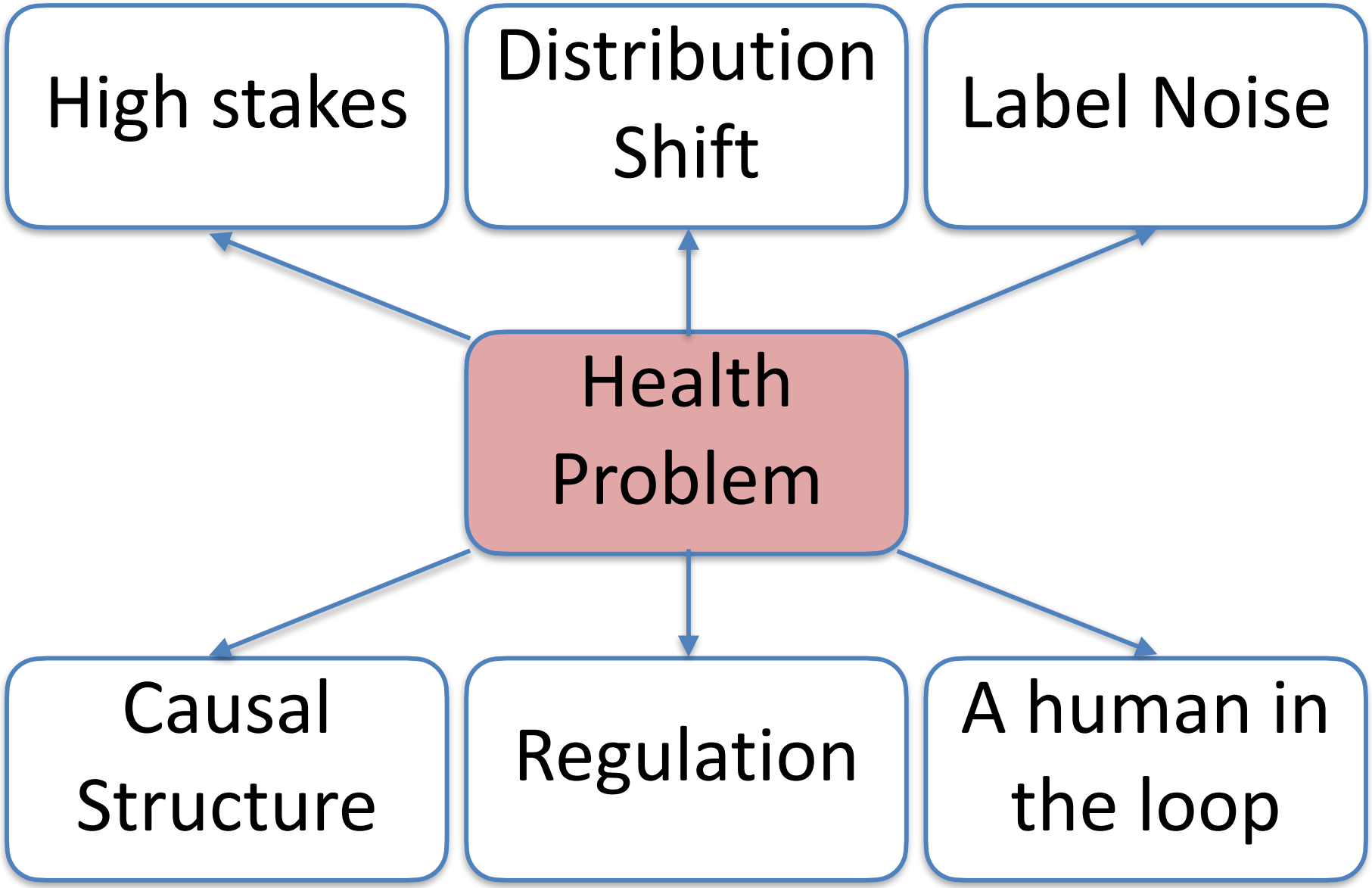
The common structure

What every health problem shares

Six features hold across all health problems

Regardless of domain or modality:

- **High stakes** — decisions affect lives.
- **Distribution shift** — patients, practice, coding all change.
- **Label noise** — "ground truth" is a clinical judgment.
- **Causal structure** — treatment and outcome entangle.
  - Value of the randomized controlled trial.
- **Regulation** — health AI is governed, not free.
- **A human in the loop** — a clinician mediates the decision.



AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

High stakes → asymmetric, patient-specific cost

In health, the two error types are **not** interchangeable.

- A missed sepsis case and a false sepsis alarm cost wildly different things.
- The cost is **patient-specific** — and it is rarely the loss your model optimizes.
- "Accuracy" averages over a cost structure that is anything but uniform.

The metric you train on is a stand-in for a clinical cost you have not written down. → Day 3: estimands

Distribution shift → the data moves under you

The IID assumption is a convenience, not a fact. Two points past the textbook:

- Shift has a **causal direction**. Predicting effect-from-cause degrades differently than cause-from-effect (anticausal); what is invariant across sites depends on the causal graph, not the marginal (Schölkopf et al., 2012).
- **More sites ≠ more general**. Transportability is a property of which mechanism is stable; naively pooling biased sources can compound bias rather than average it out.

Concretely: an ICD-9 → ICD-10 cutover moves label prevalence with no change in biology — the model learns the **calendar**, not the disease.

→ Day 2: data fusion → Day 5: transport & monitoring

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Label noise → "ground truth" is a judgment

The label is usually a proxy , recorded for another purpose — so the open problem isn't handling noise, it's constructing labels.

- The outcome you can measure ≠ the outcome you care about; the noise is often **systematic and label-dependent**, not symmetric.
- The live methodology: **programmatic weak supervision** and **anchor learning** — build labels from noisy sources you can reason about, rather than trusting one column.
- Inter-rater disagreement isn't noise to average away; it can encode the **subgroup** the model will later fail.

→ Day 2: the label is a byproduct of care

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Causal structure → treatment and outcome entangle

Most clinical data is generated under **treatment** — so the question is identifiability, not just confounding.

- Confounding by indication is the symptom; the disease is that the effect you want is **not identified** from the observational distribution without assumptions.
- Naming those assumptions explicitly — and emulating the **target trial** you can't run — is the discipline (Hernán & Robins).
- A predictor fit on treated data can be **anti-useful** for deciding whether to treat: it has learned the current policy.

This is why some health questions are **not** prediction problems at all (Kleinberg 2015).

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Confounding by Indication

Consider an observational study investigating whether taking a strong blood pressure medication increases the risk of suffering a stroke.

- The Exposure: Taking blood pressure medication.
- The Outcome: Experiencing a stroke.
- The Clinical Indication: High blood pressure (hypertension).

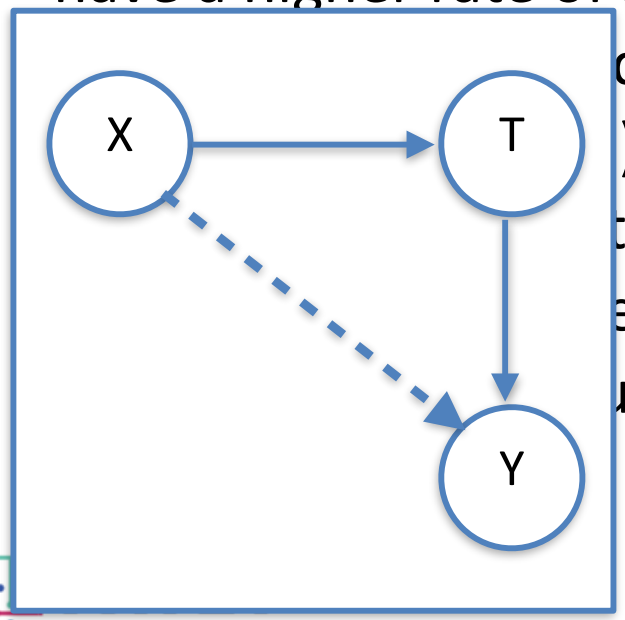
At first glance, the study might find that people taking the medication have a higher rate of strokes than those who do not. However, this is distorted by confounding by indication. The underlying condition (severe hypertension) is the exact reason the medication was prescribed in the first place, and it is also the true cause of the strokes. The medicine is being unfairly blamed for the poor health outcomes of the naturally sicker individuals who take it.

Confounding by Indication

Consider an observational study investigating whether taking a strong blood pressure medication increases the risk of suffering a stroke.

- The Exposure: Taking blood pressure medication.
- The Outcome: Experiencing a stroke.
- The Clinical Indication: High blood pressure (hypertension).

At first glance, the study might find that people taking the medication have a higher rate of strokes than those who do not. However, this is ding by indication. The underlying condition) is the exact reason the medication was t place, and it is also the true cause of the e is being unfairly blamed for the poor health urally sicker individuals who take it.



ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Confounding by Indication

Consider an observational study investigating whether taking a strong blood pressure medication increases the risk of suffering a stroke.

- The Exposure: Taking blood pressure medication.
- The Outcome: Experiencing a stroke.
- The Clinical Indication: High blood pressure (hypertension).

At first glance, the study might find that people taking the medication have a higher rate of strokes than those who do not. However, this is

din

derlying condition

) is

nedication was

t pl

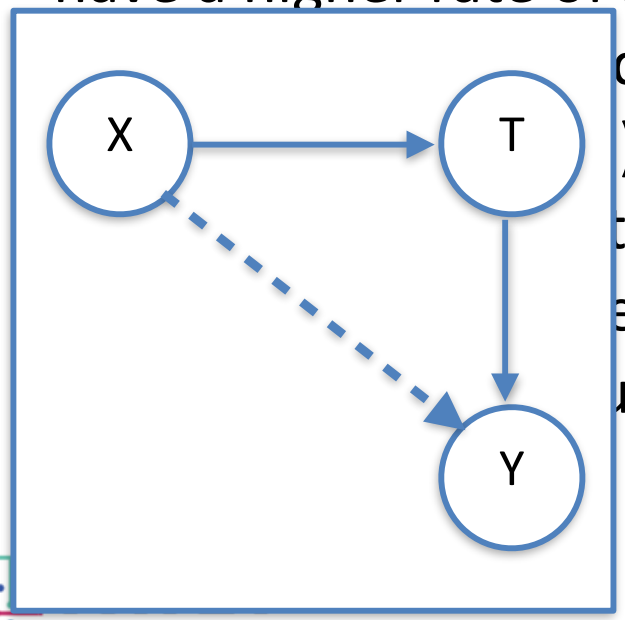
rue cause of the

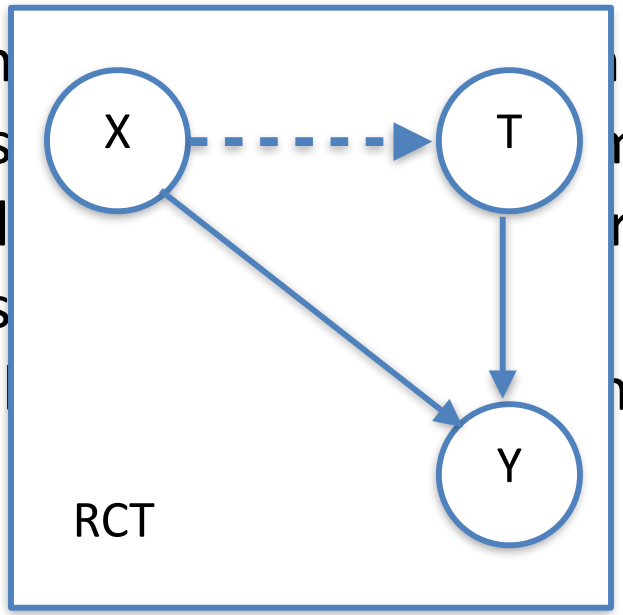
e is

for the poor health

ura

no take it.





ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Confounding vs. Identifiability

	Identifiability	Confounding
Definition	The mathematical possibility of calculating the exact causal effect from your data.	An outside (extraneous) variable that influences both the independent and dependent variables, creating a false association.
Core Issue	Whether distinct causal effects map to distinct, measurable probability distributions.	Baseline differences between groups that mix with the effect you are actually trying to measure.
Fix	Make structural assumptions (e.g., randomization, instrumental variables).	Control, stratify, or adjust for the confounding variable in the study design or analysis.

https://pmc.ncbi.nlm.nih.gov/articles/PMC2745408/

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Why health breaks generic ML assumptions

The intellectual core of the lecture:

- **IID is rarely true** — see distribution shift.
- **The label is often a proxy** — see label noise.
- **"Performance" ≠ "benefit"** — a better AUROC need not help a patient.
- **Error cost is asymmetric and patient-specific** — see high stakes.

Generic ML assumes these away. **Health does not let you.**

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Promise and limits

What AI can — and cannot — do in health

Where AI is genuinely well-suited

The honest optimistic half:

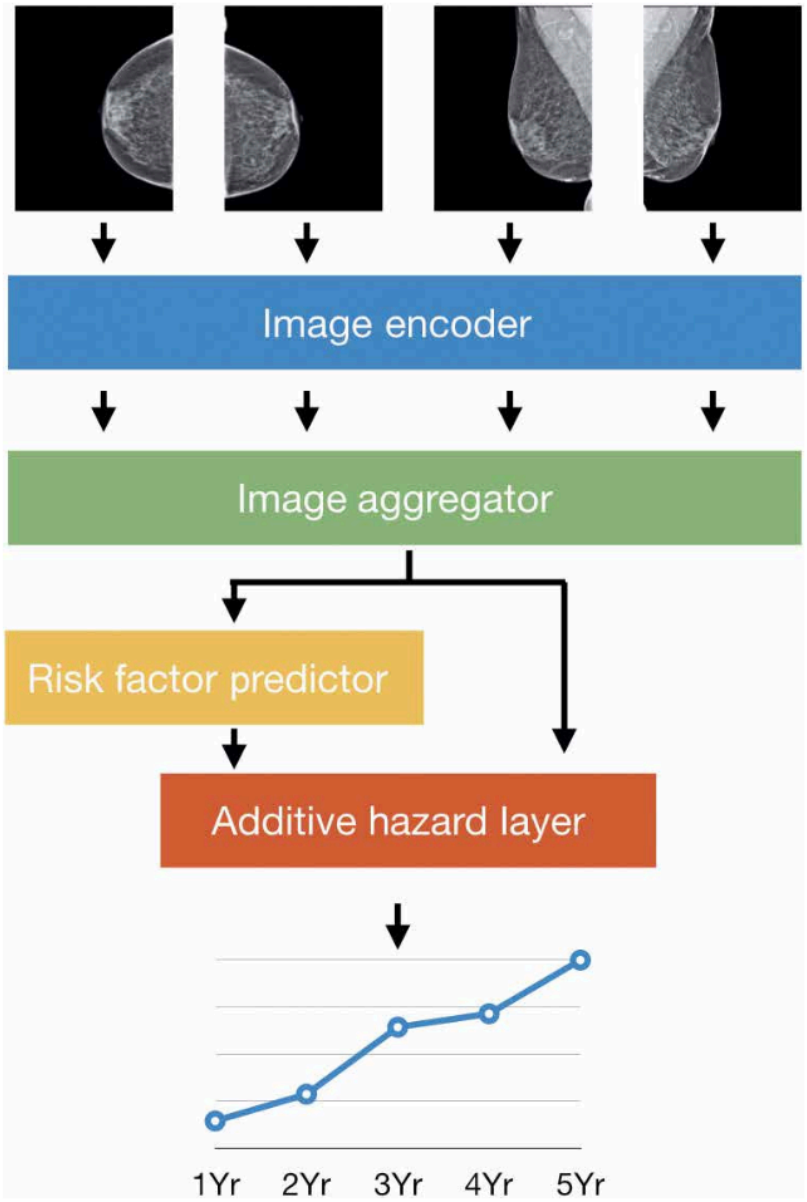
- **Pattern recognition at scale** — especially high-dimensional data.
- **Triage and prioritization** — directing scarce human attention.
- **Surfacing signal** a human reader would miss or lack time to find.
- Tasks with **abundant, well-defined labels**.

Worked example — mammography risk modeling

A real, credible contribution (Yala et al., 2021):

- A model that estimates **future** breast-cancer risk from a mammogram.
- Beats traditional risk factors; works across populations when built carefully.
- Pattern recognition at scale, on a task with **clear labels** and a **real decision** (screening interval).

Mirai Architecture



The four standard views of an individual mammogram were fed into Mirai. The image encoder mapped each view to a vector, and the image aggregator combined the four view vectors into a single vector for the mammogram. In this work, we used a single shared ResNet-18 as an image encoder, and a transformer as our image aggregator. The risk factor predictor module predicted all the risk factors used in the Tyrer-Cuzick model, including age, detailed family history, and hormonal factors, from the mammogram vector. The additive hazard layer combined information from both the image aggregator and risk factors (predicted or given) to predict coherent risk assessments across 5 years. Yala, et al., 2021.

Where AI is poorly suited — or cannot help

- Problems that are **actually causal**, not predictive.
- Settings with **no credible label** to learn from.
- Problems where the binding constraint is **not prediction** but action, access, or trust.

The most common mistake is reaching for a predictor when the decision needs a causal effect — or needs a policy change, not a model.

Summary

- AI does not establish causation from observational data alone → Day 3/4
- does not transfer across settings without evidence → Day 5
- does not absorb clinical accountability; a person remains responsible → Day 5

Prediction, causation, decision

Three different questions. Too many projects conflate them.

- **Prediction:** What will happen? Forecast the outcome.
- **Causation:** What if we act? Estimate an effect.
- **Decision:** What should we do? Choose under a loss.

Kleinberg et al. (2015) — the prediction policy problem: some decisions need an accurate forecast; others need the effect of acting.

The harder point: the **boundary moves once you deploy**. A forecast that triggers action becomes an intervention — so even "pure prediction" projects must write down an **estimand** (ICH E9(R1)), not just a metric.

→ Day 3: estimands

Worked contrast — readmission

The same clinical situation, two different questions:

- **Prediction**: "Which patients will be readmitted within 30 days?" → a forecast; useful for triage and resource planning.
- **Causation**: "Does this intervention reduce readmission?" → an effect; a target-trial question (population, treatment, contrast, outcome, time-zero).

A perfect readmission **predictor** tells you nothing about whether your **program** works — and the readmission penalty era shows the policy gap is not hypothetical.

The signature move

A problem is not a method

Start from the need, not the technique

For this room the trap is subtler than "I want to use <method>":

- Much of ML4H optimizes **benchmarks that lack construct validity** — a leaderboard is not evidence the underlying problem is real or well-posed (Raji et al., 2021).
- "State-of-the-art on <dataset>" can be a method in search of a problem, dressed as progress.

The test: can you state your problem without naming a method or a benchmark? If you can't, you have a technique — or a leaderboard — in search of a problem.

Problem formulation is itself a choice

- One clinical situation can be posed as many ML problems — and the framing has consequences **before any model exists**.
- The choice of target, population, and label encodes values (Passi & Barocas, 2019).
- **Fairness and harm are baked in at formulation** — not added later.
- A method cannot rescue a badly chosen problem.

Worked case — "reduce healthcare cost"

The same goal, two formulations, two very different systems:

- **Predict cost**
  - Target = future spending. Optimizes for who is expensive.
- **Predict need**
  - Target = unmet clinical need. Optimizes for who is sick.

Cost is a **proxy** for need — and the proxy is biased by who has had access to care. Same data, opposite consequences.

Obermeyer et al. 2019 — returns on Day 5

Stakeholders and the unmet need

A well-posed problem names the people.

- **Patients** — bear the outcome.
- **Clinicians** — act on or override the output.
- **Health systems** — absorb cost and workflow change.
- **Payers** — decide what is reimbursed.
- **Regulators** — decide what is permitted.
- **The public** — bears externalities, grants trust.

Each has different incentives — and a different definition of "success."

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

New vs. recycled — a senior, candid read

- What in *this* moment is genuinely new — and what is recurring hype?
- **Plausibly new**
  - Scale of routinely-collected EHR data
  - Modern representation learning
  - Foundation models as reusable substrate
- **Recycled / overclaimed**
  - "AI will replace clinicians" — said of expert systems in the 1980s
  - Benchmarks mistaken for deployment
  - New names for old shortcuts

The failure modes recur too: misspecification, uninterpretability, and irresponsibility are old fallacies in new clothes.

Design principles & fallacies — McDermott et al. 2023 (https://www.sciencedirect.com/science/article/pii/S0272271222000622)

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

Open problems — what we genuinely don't know

- **Identification at scale** — when can an effect be trusted from observational health data without a trial?
- **Foundation models for health** — transferable clinical structure, or laundered dataset shortcuts? → Day 4
- **Evaluation that predicts deployment value** — our benchmarks rarely do. → Day 3
- **Generative & agentic clinical systems** — no ground truth, a new failure surface.

The field's hardest questions are open. Your project can move one of them.

AHLI
ASSOCIATION FOR HEALTH LEARNING AND INFERENCE

What makes a good problem

The rubric for today's deliverable:

- A **clear unmet need** — not a technique looking for a use.
- **Identifiable stakeholders** and an explicit definition of success.
- A **plausible path** — real and pursuable within ~a year.
- **The test**: can you state it without naming a method?