Building intelligent Global Health and
Humanitarian Response Technologies

Week 11:
Trustworthy Evidence

Context Proven ≠ Trustworthy, or *Ethical*
Engineering *Uncertainty* as Architecture

LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

EPFL ARIADNE LABS ASHOKA UNIVERSITY WORLD ECONOMIC FORUM Rwanda Centre for the Fourth Industrial Revolution

When is it ethical to run a trial, and when can we trust its result?

- Equipoise: do we honestly not know which treatment is better?
- Who protects the people inside the trial?
- How would you know what was left out of the paper?

EPFL

Proven ≠ Trustworthy, or *Ethical*

Last week: Plausible ≠ Proven
This week: Proven ≠ Trustworthy…or Ethical

Prof. Mary-Anne "Annie" Hartley



LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Outline

1 Protect

Is the trial fair to the people in it?

1.1 Consent
1.2 Equipoise
1.3 Oversight

2 Design

Could chance or bias explain the result?

2.1 Randomisation
2.2 Blinding
2.3 Outcomes

+ Weigh

What does all the evidence say, taken together?

Systematic review
Meta-analysis
The evidence pyramid

CONTEXT

How did patients gain a say in research?

1942

At Dachau, Nazi physicians forced prisoners into a low-pressure chamber to simulate extreme altitude.



As the air pressure fell, the prisoner lost consciousness from lack of oxygen.

Dachau concentration camp. A prisoner subjected to a high-altitude experiment. Archival photograph via Wikimedia Commons.

CONTEXT

How did patients gain a say in research?

1942 At Dachau, Nazi physicians forced prisoners into a low-pressure chamber to simulate extreme altitude.

1947 **Nuremberg Code**. The Doctors' Trial judges made voluntary consent the first of ten principles.

Nuremberg Code

Doctors' Trial judgment, 1947

"The voluntary consent of the human subject is absolutely essential."

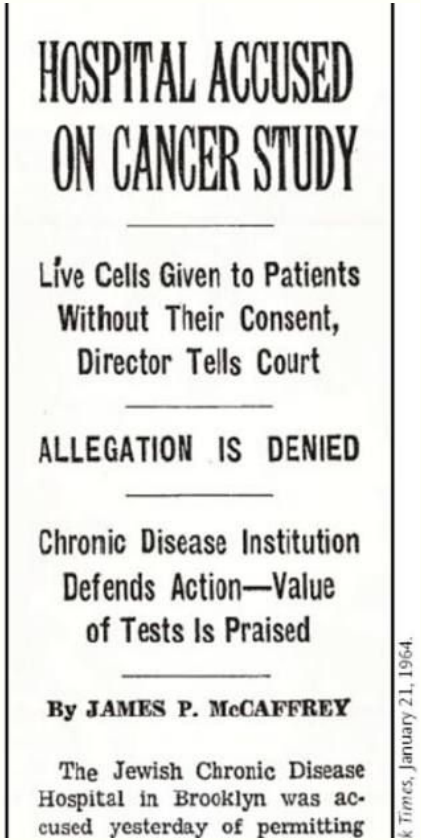


Doctors' Trial defendants, photographed by Raymond D'Addario. Code excerpt from the 1947 judgment, reproduced by the U.S. Department of Energy.

CONTEXT

How did patients gain a say in research?

1942	At Dachau, Nazi physicians forced prisoners into a low-pressure chamber to simulate extreme altitude.
1947	**Nuremberg Code**. The Doctors' Trial judges made voluntary consent the first of ten principles.
1963	In Brooklyn, Chester Southam and colleagues injected 22 patients with live cancer cells without telling them what the cells were.



HOSPITAL ACCUSED ON CANCER STUDY

Live Cells Given to Patients Without Their Consent, Director Tells Court

ALLEGATION IS DENIED

Chronic Disease Institution Defends Action—Value of Tests Is Praised

By JAMES P. McCAFFREY

The Jewish Chronic Disease Hospital in Brooklyn was accused yesterday of permitting

k Times, January 21, 1964.

LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Jewish Chronic Disease Hospital, Brooklyn, 1963. Patients received cancer cells without being told what was injected.

CONTEXT

How did patients gain a say in research?

1942	At Dachau, Nazi physicians forced prisoners into a low-pressure chamber to simulate extreme altitude.
1947	**Nuremberg Code**. The Doctors' Trial judges made voluntary consent the first of ten principles.
1963	In Brooklyn, Chester Southam and colleagues injected 22 patients with live cancer cells without telling them what the cells were.
1964	**Declaration of Helsinki**. The World Medical Association set out physicians' duties to research participants.

Declaration of Helsinki

World Medical Association, 1964

"It is the mission of the doctor
to safeguard the health of the people."

Read the WMA's original June 1964 declaration

CONTEXT

How did patients gain a say in research?

1942	At Dachau, Nazi physicians forced prisoners into a low-pressure chamber to simulate extreme altitude.
1947	**Nuremberg Code. The Doctors' Trial judges made voluntary consent the first of ten principles.**
1963	In Brooklyn, Chester Southam and colleagues injected 22 patients with live cancer cells without telling them what the cells were.
1964	**Declaration of Helsinki. The World Medical Association set out physicians' duties to research participants.**
2024	**Engage communities. Share burdens and benefits fairly. Consider exclusion harms and safeguards.**



WMA delegates in Helsinki for the 2024 revision.

1 Protect

Is the trial fair to the people in it?

1.1 Consent

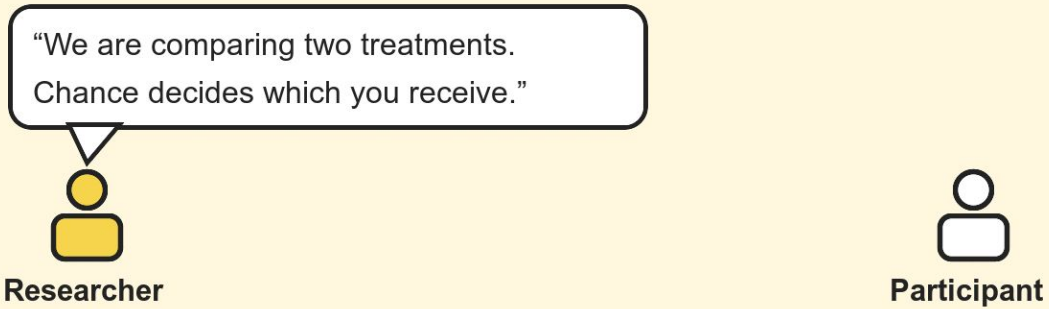
1.2 Equipoise

1.3 Oversight

Informed consent

A process in which a participant or legally acceptable representative voluntarily agrees to trial participation after receiving and discussing the information relevant to that decision.

1 Explain the study



CONTEXT

Tuskegee: what did the men agree to?

Where	Macon County, Alabama.
Who	600 Black men enrolled. Of these, 399 had syphilis and 201 did not.
Consent	Researchers described examinations as care for “bad blood” and offered a spinal tap as “special free treatment.”



CONTEXT

Tuskegee: 40 years

1932 Study begins: 399 Black men with syphilis are told they have bad blood

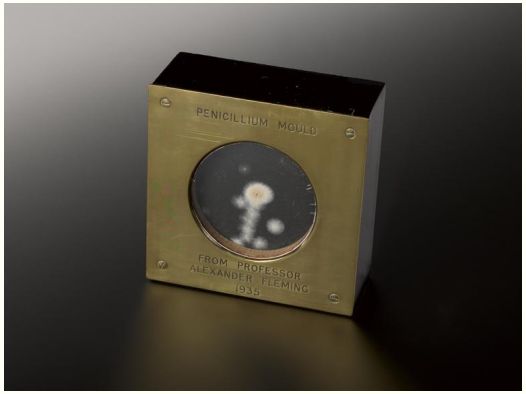


Image: U.S. National Archives. Sources: CDC study timeline and Brandt, Racism and Research (1978).

CONTEXT

Tuskegee: 40 years

1932	Study begins: 399 Black men with syphilis are told they have bad blood
1947	Penicillin becomes standard and is withheld from the men



Fleming's penicillin mould, 1935. By 1947, penicillin was standard treatment for syphilis.

Image: U.S. National Archives. Sources: CDC study timeline and Brandt, Racism and Research (1978).

CONTEXT

Tuskegee: 40 years

1932	Study begins: 399 Black men with syphilis are told they have bad blood
1947	Penicillin becomes standard and is withheld from the men
1972	By 1969, at least 28 men had died of syphilis. In 1972, a whistleblower helped bring the study to public attention.





Men who took part in the Tuskegee study. Archival photograph via KQED.

Image: U.S. National Archives. Sources: CDC study timeline and Brandt, Racism and Research (1978).

CONTEXT

Tuskegee: 40 years

1932	Study begins: 399 Black men with syphilis are told they have bad blood
1947	Penicillin becomes standard and is withheld from the men
1972	By 1969, at least 28 men had died of syphilis. In 1972, a whistleblower helped bring the study to public attention.
1974	National Research Act Congress strengthened review requirements for federally supported human research.



A survivor at the White House apology, 1997.



The New York Times

Syphilis Victims in U.S. Study Went Untreated for 40 Years

By JEAN HELLER
The Associated Press

WASHINGTON, July 25—For 40 years the United States Public Health Service has conducted a study in which human beings with syphilis, who were induced to serve as guinea pigs, have gone without medical treatment for the disease and a few have died of its late effects, even though an effective therapy was eventually discovered.

The study was conducted to determine from autopsies what the disease does to the human body.

Officials of the health service who initiated the experiment have long since retired. Current officials, who say they have serious doubts about the morality of the study, also say that it is too late to treat the syphilis in any surviving participants.

Doctors in the service say they are now rendering whatever other medical services they can give to the survivors while the study of the disease's effects continues.

Dr. Merlin K. DuVal, Assistant Secretary of Health, Education and Welfare for Health and Scientific Affairs, expressed shock on learning of the study. He said that he was making an immediate investigation.

The experiment, called the Tuskegee Study, began in 1932 with about 600 black men,

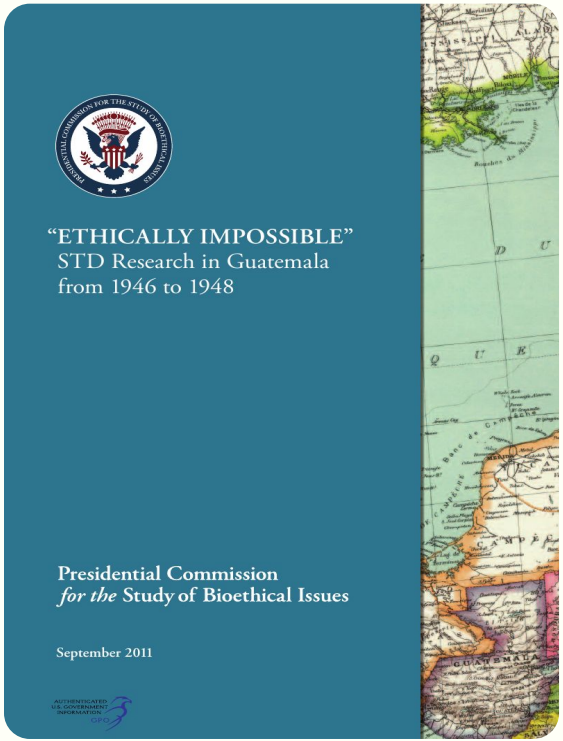
LIGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Image: U.S. National Archives. Sources: CDC study timeline and Brandt, Racism and Research (1978). By 1969, at least 28 men, and possibly as many as 100, had died from syphilis.

CONTEXT

Guatemala: exposure without consent

1946–48	Researchers arranged sexual exposure or directly inoculated people.
People	At least 1,308 people were exposed without consent.
Care	Treatment was documented for only 678 people.
Deaths	83 deaths were recorded. Attribution remains uncertain.



LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

US Presidential Commission, "Ethically Impossible" (2011). Official investigation.

1 Protect

Is the trial fair to the people in it?

1.1 Consent

1.2 Equipoise

1.3 Oversight

LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Clinical equipoise

Clinical **experts** genuinely disagree about which **strategy** offers these patients the best balance of benefits and harms.

Is the comparison fair?

Which experts?	Which clinicians know the evidence and care for these patients?
Compared with what?	What is the best proven treatment? What is available locally?
Is it fair…whom?	Could those clinicians reasonably offer either trial arm?

CONTEXT

Zidovudine in pregnancy

Where	Antenatal clinics in Bangkok, Thailand.
Who	Pregnant women living with HIV. Few could access the proven longer zidovudine regimen.
Comparator	A shorter course might prevent HIV transmission to infants. Which control should researchers use?

Vote: compare it with placebo or the proven long course?



Antenatal care in Thailand. Illustrative photograph, The Coverage.

ACTG 076 (1994). Shaffer et al., Lancet 1999.

CONTEXT

The Bangkok trial used placebo

Pregnant women living with HIV were randomly assigned to one of two groups.

Short-course zidovudine	Placebo
9.4%	18.9%

Estimated proportion of infants infected with HIV by six months.

Shaffer et al., Lancet 1999;353:773–780. Comparator debate: Lurie & Wolfe and Varmus & Satcher, NEJM 1997.

CONTEXT

FEAST: fluids for severe infection

Where	Hospitals in Kenya, Tanzania and Uganda.
Who	Children with severe infection and poor circulation, without severe hypotension.
Comparator	Rapid saline or albumin bolus versus no initial bolus. All groups received maintenance fluids.

Vote: would you approve a no-bolus control group?



Soroti Hospital, Uganda. Malaria Consortium.

Maitland et al., NEJM 2011.

CONTEXT

FEAST: more deaths after a fluid bolus

Death within 48 hours, by randomised group

Albumin bolus	Saline bolus	No initial bolus
10.6%	10.5%	7.3%

All groups received maintenance fluids. Mortality was higher in both bolus groups.

Maitland et al., NEJM 2011. Main trial stratum.

1 Protect

Is the trial fair to the people in it?

1.1 Consent

1.2 Equipoise

1.3 Oversight

Independent oversight

Independent oversight protects participants before and during a trial.

Ethics committees review the protocol. Sponsors monitor safety. An independent monitoring board, when appointed, reviews emerging data and can recommend stopping.

1. PROTECT: OVERSIGHT

A safe pill for morning sickness

THE LOGIC

Thalidomide calmed nausea, and no dose could be found that killed a rat

CONTEXT

Therefore:
give it to pregnant women!



Image: Contergan package, via Wikimedia Commons.

CONTEXT

Thalidomide: from warning to withdrawa

1957 Sold in Germany without a prescription



LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Image: Wikimedia Commons. Lenz W. Lancet 1962;279:45. McBride WG. Lancet 1961;278:1358. Kefauver-Harris Amendment, 1962.

CONTEXT

Thalidomide: from warning to withdrawal

1957	Sold in Germany without a prescription
1961	Clinicians Widukind Lenz and William McBride independently raised the alarm about limb differences after prenatal exposure.



Foot differences after prenatal exposure to thalidomide. National Museum of Health and Medicine, Otis Historical Archives.

Image: Wikimedia Commons. Lenz W. Lancet 1962;279:45. McBride WG. Lancet 1961;278:1358. Kefauver-Harris Amendment, 1962.

CONTEXT

Thalidomide: from warning to withdrawal

1957	Sold in Germany without a prescription
1961	Clinicians Widukind Lenz and William McBride independently raised the alarm about limb differences after prenatal exposure.
By 1962	More than 10,000 babies affected in 46 countries

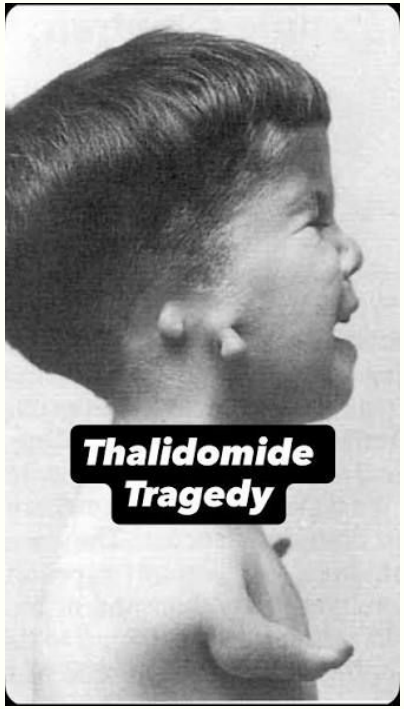


Image: Wikimedia Commons. Lenz W. Lancet 1962;279:45. McBride WG. Lancet 1961;278:1358. Kefauver-Harris Amendment, 1962.

CONTEXT

Thalidomide: from warning to withdrawal

1957	Sold in Germany without a prescription
1961	Clinicians Widukind Lenz and William McBride independently raised the alarm about limb differences after prenatal exposure.
By 1962	More than 10,000 babies affected in 46 countries
1962	The US strengthens drug approval requirements: manufacturers must demonstrate efficacy as well as safety.

Image: Wikimedia Commons. Lenz W. Lancet 1962;279:45. McBride WG. Lancet 1961;278:1358. Kefauver-Harris Amendment, 1962.

Who challenged the safety claim?

Scale of harm	About 10,000 children born with birth defects worldwide. An estimated 2,000 deaths.
Who challenged it?	German paediatrician Widukind Lenz and Australian obstetrician William McBride independently linked the defects to thalidomide.
Monitoring in practice	A clinician reports an unusual limb defect. Regulators compare similar reports and check for a shared exposure in early pregnancy.

Further reading: Suffer the Children, Sunday Times Insight Team. Harm estimates: BMJ 2016;352:i353.

2 Design

Could chance or bias explain the result?

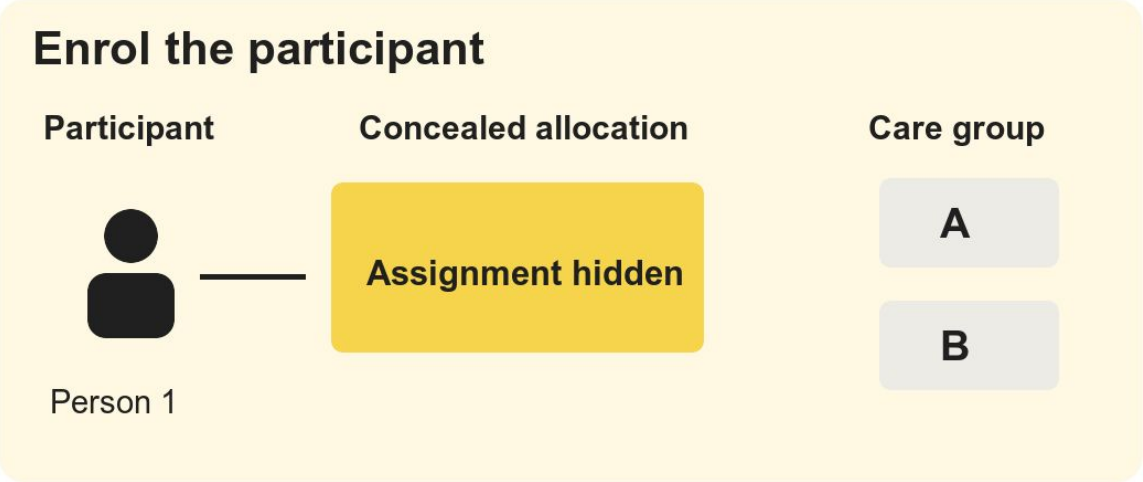
2.1 Randomisation

2.2 Blinding

2.3 Outcomes

Randomisation

After enrolment, chance determines which care a participant receives. Conceal the next assignment until enrolment is complete.



CONTEXT

Randomisation: three ways to let chance decide

Simple

Fresh draw for each person

Blocked

Two A and two B per block

Stratified

Separate lists for each risk group

LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Illustrative sequences: A = intervention; B = comparator. Cochrane Handbook, chapter 8.

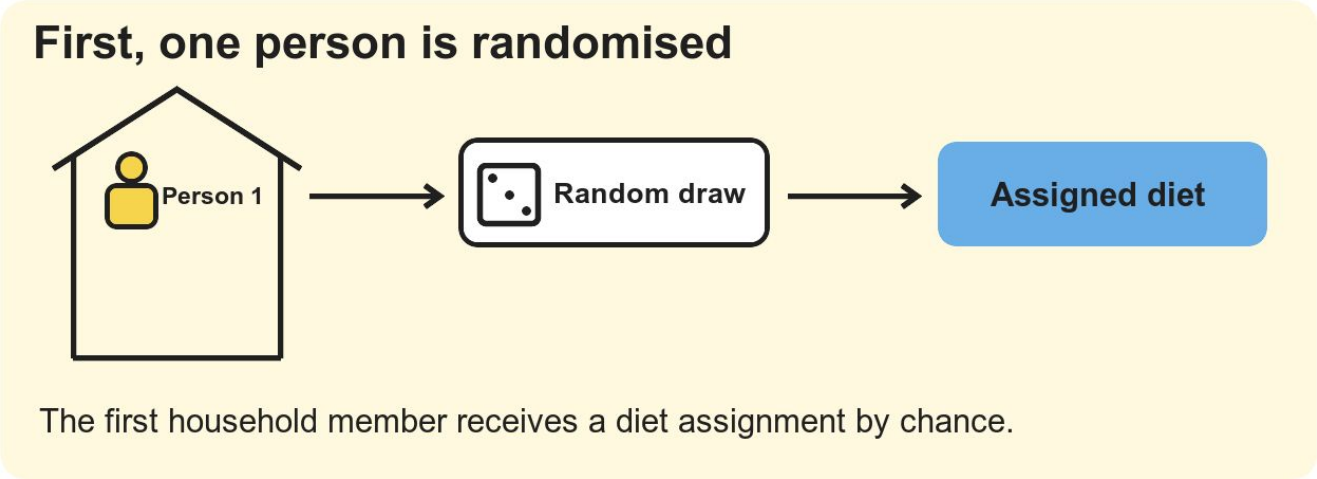
Which randomisation scheme would you choose?

Method	When to choose it	Watch for
Simple	Large trial: give each participant an independent 50:50 draw.	Small samples can end up with uneven group sizes.
Blocked	Small trial: shuffle two A and two B assignments within each block of four.	Known block sizes can reveal the final assignment.
Stratified	Risk matters: randomise separately within low- and high-risk groups.	Too many factors create strata with very few participants.

You can use blocks within each stratum to balance group sizes.

CONTEXT

PREDIMED: two people, one random draw



Individual randomisation was planned. Some relatives received the first household member's diet, so their assignments were linked.

Estruch et al., NEJM 2018;378:e34. Household example.

CONTEXT

Allocation concealment

Terminology

Allocation concealment

Keep the next treatment assignment hidden until enrolment is irrevocable.



MRC 1948: tuberculosis;
streptomycin + bed rest versus bed rest.

AI-generated teaching image

CONTEXT

Predict the next allocation—then find the bias

Selection bias, Predictability, Concealment

Alternation

Block of four

Two A and two B

Vote: could a recruiter use this knowledge to change who enters next?

Classroom challenge, illustrative sequences, not an actual trial. Cochrane Handbook, chapter 8.

2 Design

Could chance or bias explain the result?

2.1 Randomisation

2.2 Blinding

2.3 Outcomes

LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Blinding

Keep group assignments hidden from the people whose knowledge could influence care or outcome assessment.

Open-label

Who can see the treatment assignment?

Participant	Clinician	Outcome assessor
A / B	A / B	A / B
Assignment visible	Assignment visible	Assignment visible

CONTEXT

Who knows the treatment assignment?

Single-blind	Common example: participants do not know which treatment they receive.
Double-blind	Common example: both participants and treating clinicians are unaware of the assignment.
Outcome assessor	Report this role separately. An assessor can be blinded even when the participant and clinician know the treatment.

Trial reports should name each blinded role. The labels alone are ambiguous.

CONTEXT

Vitamin C: participants could taste the difference

What participants saw

Matching capsules contained either vitamin C or a lactose placebo.

What gave it away

Some volunteers opened a capsule and tasted the contents. They could then guess their group.

If you think you received vitamin C, might you rate the same cold symptoms differently?

Karlowski et al., JAMA 1975. How much unblinding affected the result remains disputed.

When can a clinician break the blind?

Urgent care	A participant has life-threatening bleeding. If knowing the treatment would change urgent reversal, obtain their assignment through the emergency procedure.
Fatal car crash	Report the fatal event promptly to the sponsor under the protocol, even if it first appears unrelated. Seriousness and causality are separate judgments.
Investigate	Could a heart attack have preceded the crash? Obtain clinical or autopsy findings, assess relatedness and update the report. Keep others blinded where possible.

ICH E6 (R3), §§2.7.2 and 2.11. Protocol-defined exceptions may apply to immediate reporting. Document the reason for any code break.

2 Design

Could chance or bias explain the result?

2.1 Randomisation

2.2 Blinding

2.3 Outcomes

TERMINOLOGY

CONTEXT

Outcomes

Decide in advance what to measure and when.
Choose outcomes that matter to patients.

One trial, three kinds of outcome

Primary	Does the intervention reduce 30-day mortality? *This is the main question used to plan the sample size.*
Secondary	How many days are spent alive and out of hospital by day 30? *Prespecify this additional benefit and its analysis.*
Exploratory	Does the mortality effect vary by baseline heart rate? *Treat this as a hypothesis for a future trial.*

Heart drugs after a heart attack

THE LOGIC

Extra heartbeats after a heart attack predict sudden death, and these drugs suppress them

CONTEXT

Therefore:
suppress the beats, save the patient!



Image: ECG with a premature ventricular beat, via Wikimedia Commons.

CONTEXT

CAST: testing the obvious

1980s	Extra ventricular beats after a heart attack predict risk. Drugs that suppress them are widely prescribed.
1987	CAST asks whether suppressing these beats saves lives. Participants receive active drug or matching placebo.
1989	Excess deaths trigger early stopping of the encainide and flecainide arms.
Result	All-cause mortality: 56/730 (7.7%) on active drug versus 22/725 (3.0%) on placebo.

CAST Investigators. N Engl J Med 1989;321:406–412. Echt DS et al. N Engl J Med 1991;324:781–788.

CONTEXT

ENGINEERING

Randomising is not a train/test split

Random allocation	Train / test split
Assign care → groups comparable in expectation.	Partition records → evaluate generalisation.

A random train/test split preserves confounding already present in treatment assignment.

Different operations: assigning treatment versus partitioning observed data.

CONTEXT

Weigh

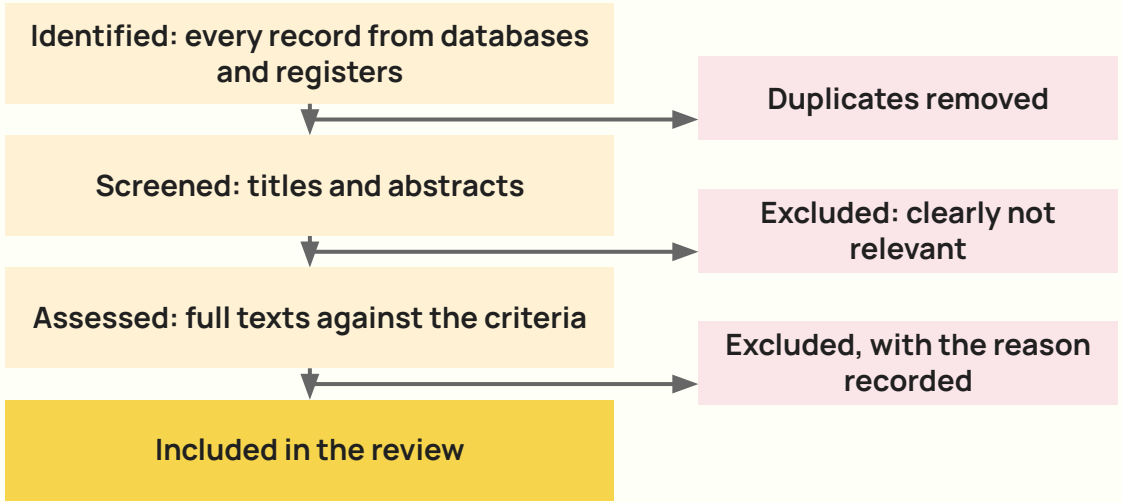
What does all the evidence say, taken together?

Systematic review

Meta-analysis

The evidence pyramid

A search for every study on one question, using rules written down in advance.



A PRISMA diagram counts records at each stage and gives reasons for exclusions after full-text assessment.

CONTEXT

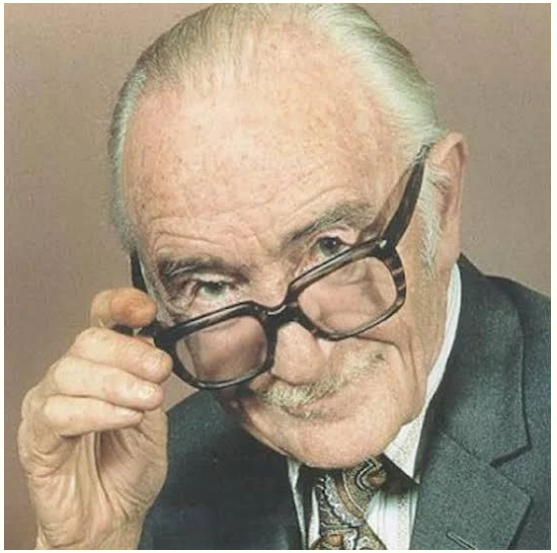
Example: Cochrane's challenge

CARDIFF, 1979

Archie Cochrane: medicine keeps no organised summary of its trials

Cochrane, founded in 1993: an international non-profit network that writes and updates systematic reviews of health care

Therefore: one trial is never the whole story



CONTEXT

Good practice: show your search

Protocol registered in advance (PROSPERO).
Search, criteria and flow diagram reported (PRISMA).

CONTEXT

Systematic review: define eligibility first

Include	Exclude
Adult smokers	Wrong population
Programme vs usual support	Wrong comparison
Randomised trials	Wrong study design

Illustrative eligibility criteria; Cochrane Handbook, chapter 3.

CONTEXT

Systematic review:

Research in context

Evidence before this study

We searched PubMed from Jan 1, 2017, to April 30, 2026, for human studies with study designs including clinical trials and observational studies using the terms “obesity” AND (“semaglutide” OR “tirzepatide” OR “GLP-1”) AND (“discontinuation” OR “withdrawal” OR “switching” OR “continuation” OR “bariatric surgery”) in the title or abstract. One randomized clinical trial and one extension study established that discontinuation of these medications reverses their beneficial effects on body weight and cardiometabolic parameters. Several observational studies documented that discontinuation of semaglutide or tirzepatide is common in clinical practice, and a few described the alternative obesity treatments patients pursue in routine care. However,

Illustrative eligibility criteria; Cochrane Handbook, chapter 3.

CONTEXT

Weigh

What does all the evidence say, taken together?

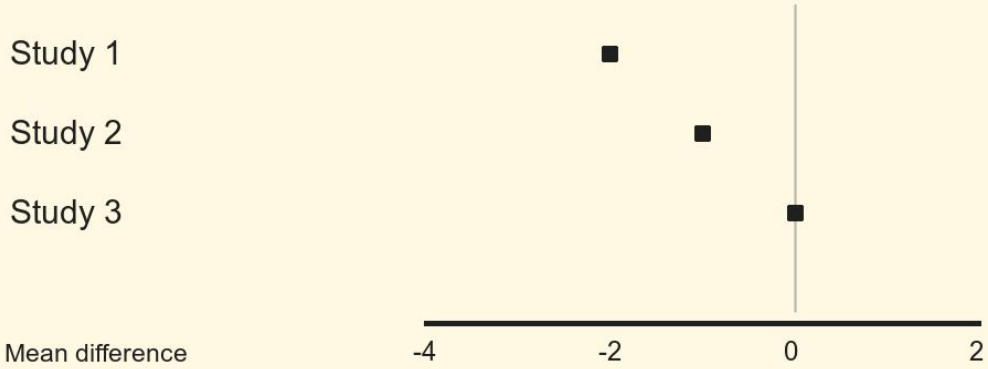
Systematic review

Meta-analysis

The evidence pyramid

Meta-analysis combines estimates from several studies.

Plot each study estimate



CONTEXT

Example: a review of coin flips

A review of every published study of coin flips

Every paper reports a coin that lands heads. The tails results were never sent to a journal

Therefore: coins land heads?

Illustrative example.

CONTEXT

Limitation of a systematic review

Publication bias

Where it failed

Imagine that researchers publish every coin flip that lands heads and shelve every tails result. A review could find every published study and still make the coin look biased.

Illustrative example.

CONTEXT

Example: steroids before birth

SEVEN TRIALS, 1972 TO 1991

Each trial alone was too small to be convincing
Pooled together, the answer was clear years before practice changed

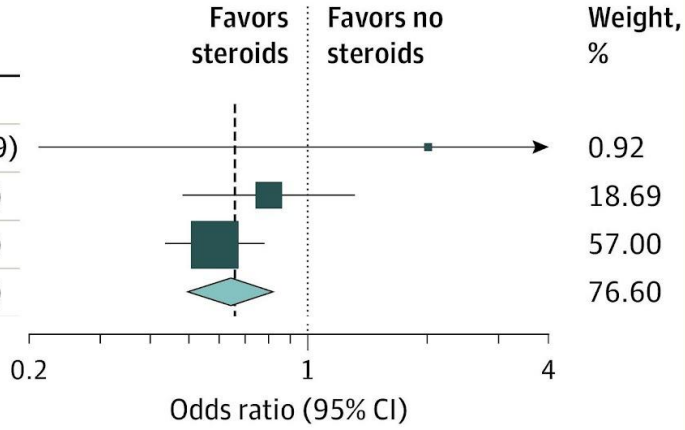
Therefore: steroids cut deaths in premature babies by 30 to 50%

CONTEXT

Forest plot: dexamethasone and 28-day mortality

Odds ratio, 95% confidence interval, Subgroup pooled estimate

Drug and trial	ClinicalTrials.gov identifier	No. of deaths/total No. of patients: Steroids	No. of deaths/total No. of patients: No steroids	Odds ratio (95% CI)	Weight, %
Dexamethasone					
DEXA-COVID 19	NCT04325061	2/7	2/12	2.00 (0.21-18.69)	0.92
CoDEX	NCT04327401	69/128	76/128	0.80 (0.49-1.31)	18.69
RECOVERY	NCT04381936	95/324	283/683	0.59 (0.44-0.78)	57.00
Subgroup fixed effect		166/459	361/823	0.64 (0.50-0.82)	76.60



LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

WHO REACT, JAMA 2020, Figure 2 excerpt: dexamethasone subgroup; dose column omitted.

CONTEXT

What a meta-analysis can do

Find an effect too small for any one trial

Seven small trials of steroids before preterm birth looked inconclusive one by one. Pooled, they showed far fewer babies dying.

Show when one more trial is one too many

Pooling trials in date order showed that clot-busting drugs for heart attack worked by 1973. Placebo-controlled trials ran for another 15 years.

Catch harm early

Pooling the rofecoxib (Vioxx) trials in date order showed the excess of heart attacks by 2000. The drug was withdrawn in 2004.

Pool observational studies, confounding included

Pooled cohort studies suggested hormone therapy protected the heart. A randomised trial then found the opposite.

Crowley et al. BJOG 1990;97:11-25. Lau et al. NEJM 1992;327:248-254. Juni et al. Lancet 2004;364:2021-2029. Stampfer & Colditz. Prev Med 1991;20:47-63. WHI. JAMA 2002;288:321-333.

CONTEXT

Example: an average temperature

Head in the oven, feet in the freezer

On average, the temperature is perfectly comfortable

Therefore: the patient is fine?

Illustrative example.

CONTEXT

Limitation of a meta-analysis

Heterogeneity

Where it failed

A person with their head in an oven and feet in a freezer can have a comfortable average temperature. Likewise, pooling very different treatment effects can hide harm in one group and benefit in another.

Illustrative example.

CONTEXT

Weigh

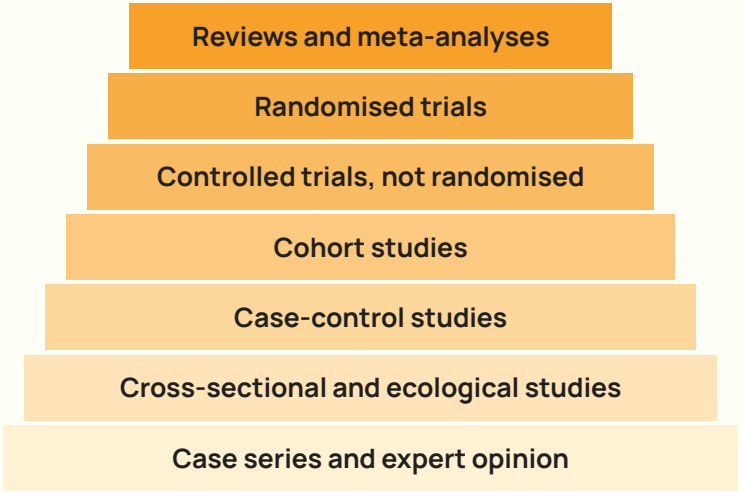
What does all the evidence say, taken together?

Systematic review

Meta-analysis

The evidence pyramid

A ranking of study designs by how well each one protects against bias.



Stronger at any level

More patients

More sites and populations

Longer follow-up

Outcomes that matter to patients

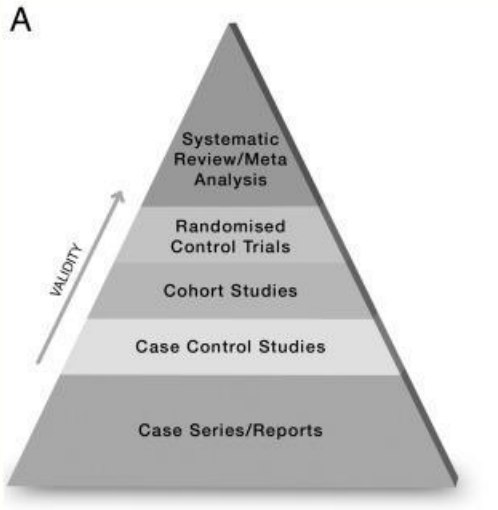
Plan declared in advance

Repeated by others

For treatment effects, study conduct and risk of bias matter at every level.

CONTEXT

The evidence pyramid



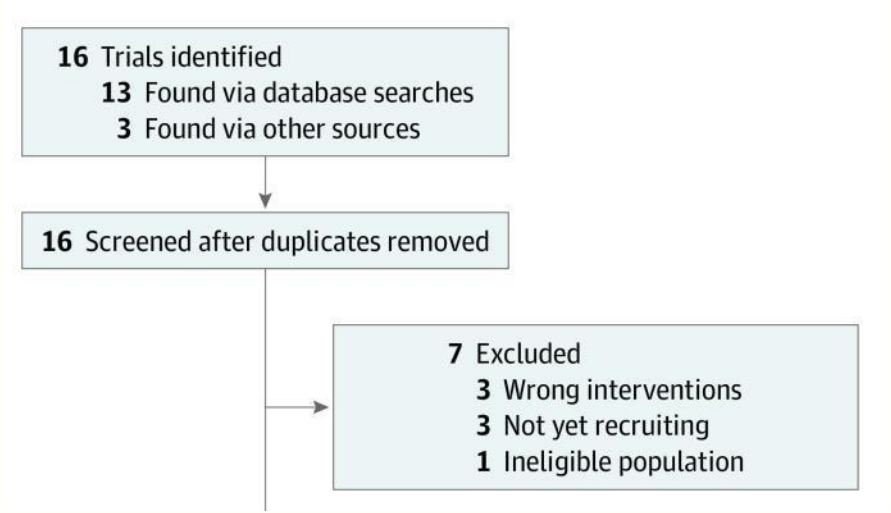
Systematic review / meta-analysis
Randomised trial
Cohort
Case–control
Case series / report

Murad et al., BMJ EBM 2016, Figure 1A. Traditional hierarchy; quality still requires appraisal.

CONTEXT

Study selection: identify and screen records

PRISMA-style selection flow, Eligible trials

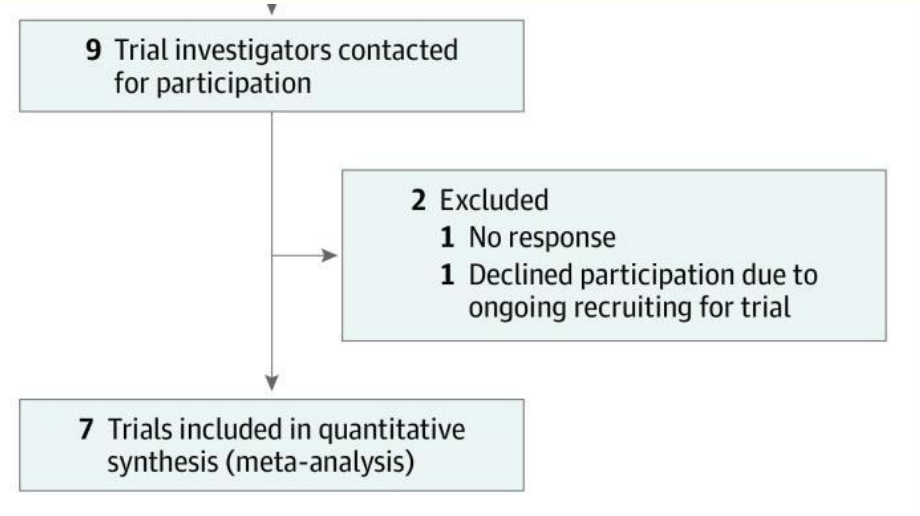


WHO REACT Working Group, JAMA 2020, Figure 1. Original trial-selection flow, shown in two parts.

CONTEXT

Study selection: explain exclusions and inclusion

Eligibility, Exclusion reasons, Included trials



9 trial teams contacted
2 did not contribute
7 trials in the meta-analysis

Why were trials excluded?

WHO REACT Working Group, JAMA 2020, Figure 1. Original trial-selection flow, shown in two parts.

CONTEXT

Many models, few trials

AI IN HEALTH, 2018 TO 2023

Tens of thousands of AI-in-health papers every year

but only 86 randomised trials of AI in the clinic, in six years

Therefore: published is not the same as proven in patients

IMAGE:
XY plot,
AI-in-health papers
versus
AI clinical trials, by year

Han R et al. Lancet Digit Health 2024;6:e367-e373. Plana D et al. JAMA Netw Open 2022;5:e2233946.

CONTEXT

EAGLE: AI-enabled ECG screening

An ECG algorithm flags possible weak heart pumping.

Clinicians can order an echo.



AI-generated teaching image

CONTEXT

EAGLE: randomise the team that changes care

Cluster randomisation, Unit of allocation, Diagnostic outcome

22,641 adults undergoing ECG; 120 clinical teams.

AI result visible	Usual care
ECG flag → clinician may order echocardiography	AI result not shown to the clinical team

New low-ejection-fraction diagnosis by 90 days: 2.1% vs 1.6%.

Yao et al., Nature Medicine 2021. More diagnoses; survival benefit was not established.

CONTEXT

Manz: prompting serious-illness conversations

Whole intervention, Process outcome, End-of-life care

Oncology clinics: mortality-risk predictions + clinician prompts versus usual care; stepped-wedge cluster randomisation.

Conversation documented	High-risk encounters: 3.4% → 13.5%
Therapy near death	10.4% → 7.5%; other outcomes did not clearly improve

Manz et al., JAMA Oncology 2023. 20,506 patients; 41,021 encounters.

CONTEXT

Downing: testing a sepsis alert

Comparator, Process outcome, Statistical power

1,123 adult ward patients:
rule-based sepsis alert shown to the team versus the same alert kept silent.

New antibiotic orders within 3 hours

35.0% active alert	36.7% silent alert

p = 0.53; stopped at 1,123 of 3,000 planned. Null ≠ equivalence.

LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Downing et al., BMJ Quality & Safety 2019;28:762–768. This was rule-based CDS, not ML.

CONTEXT

Design an evaluation of model-guided care

Intervention, Comparator, Outcome, Follow-up	
Intervention	A deterioration score triggers a named clinical response.
Comparator	The same eligible patients receive usual escalation.
Outcome / time	Choose a patient-important outcome and a time horizon.

Pairs: which labels and allocation unit would test clinical benefit?

CONTEXT

ENGINEERING

Clinical terminology → machine-learning concepts

Confounder	Common causes → choose adjustment variables using causal structure.
Counterfactual outcome	Missing potential-outcome label → causal assumptions needed.
Random allocation	Assignment mechanism ≠ train/test partitioning.
Clinical benefit	Deployment target → patient outcomes after model-guided care.

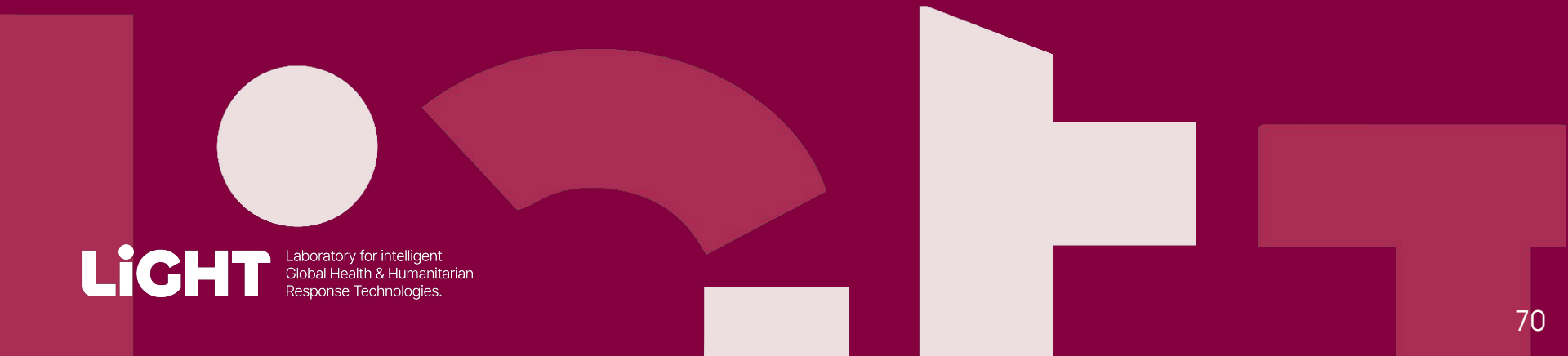
Connections to the engineering session: causal structure, identification assumptions and effect estimation.

Questions?

EPFL

Uncertainty as Architecture

Lecturer: *Tim Arni*



LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

ENGINEERING

Outline

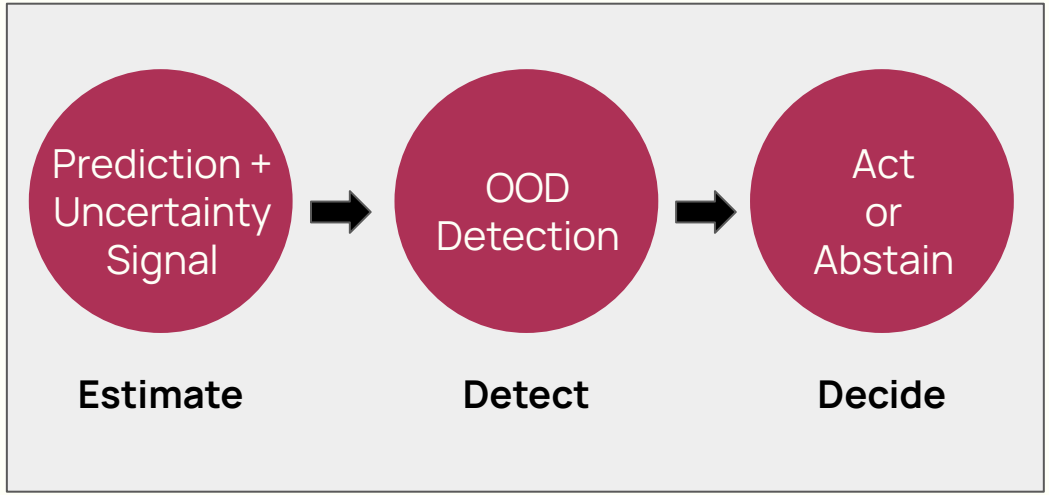
01 Introduction

02 Uncertainty & Calibration

03 OOD Detection

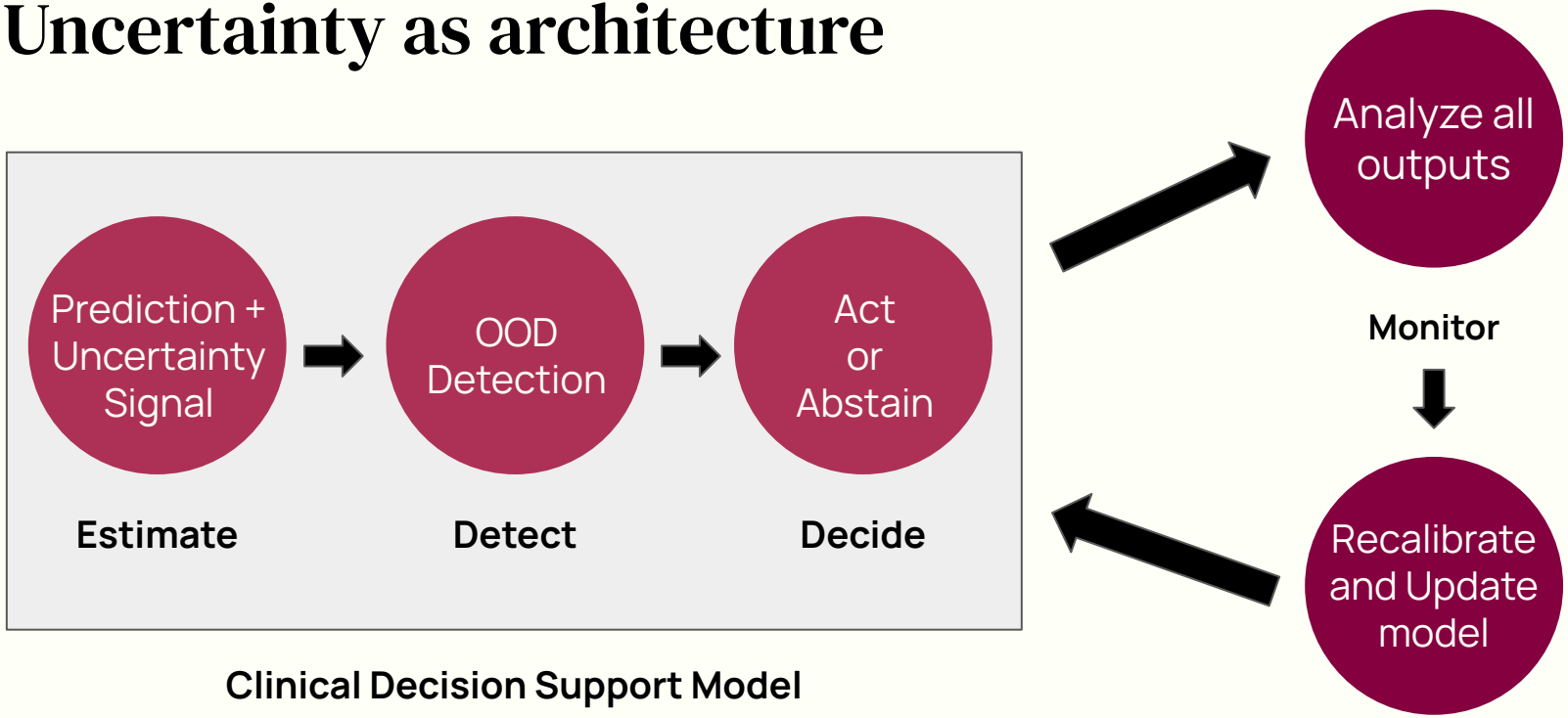
04 Back to your Projects

Uncertainty as architecture



Clinical Decision Support Model

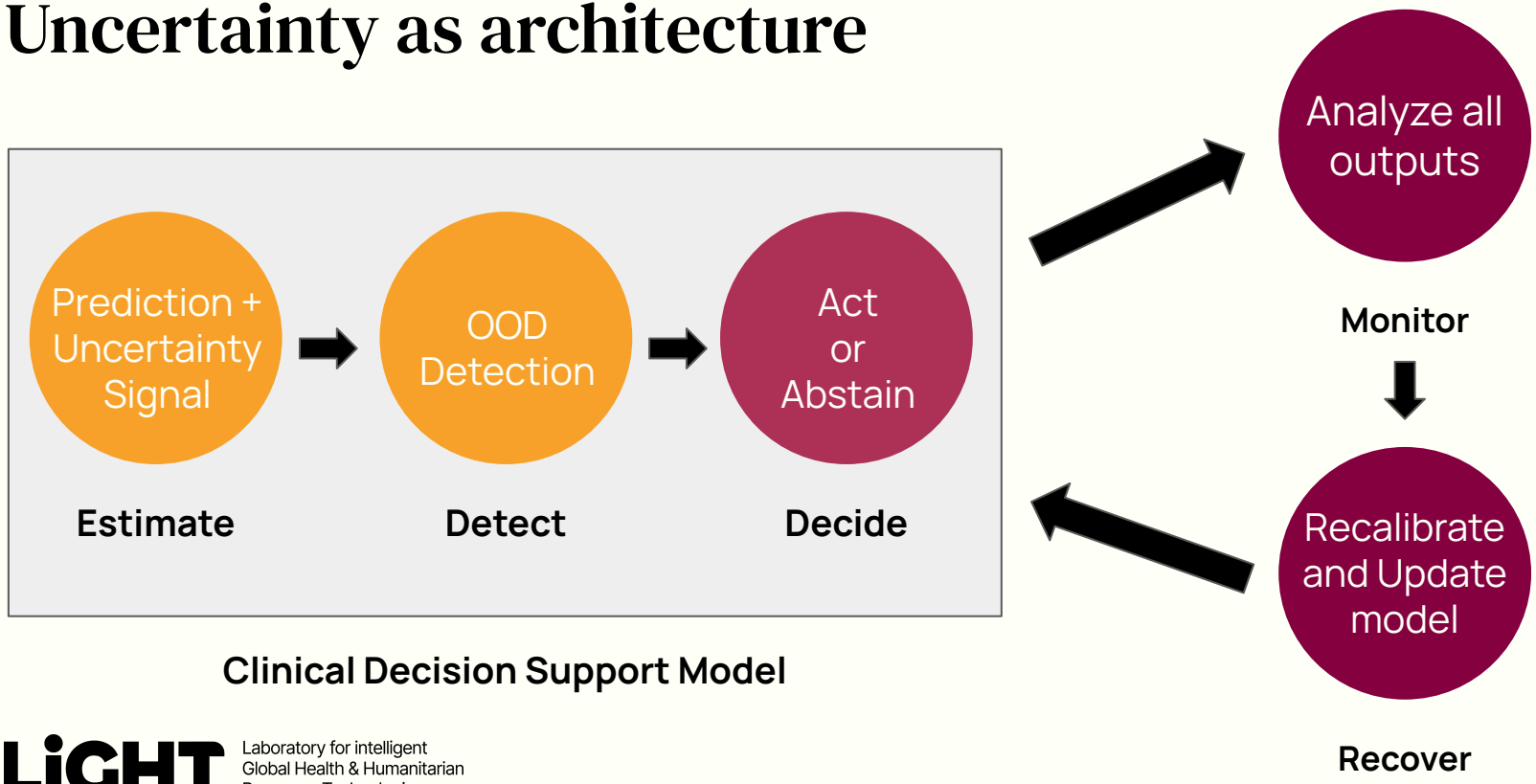
Uncertainty as architecture



Clinical Decision Support Model

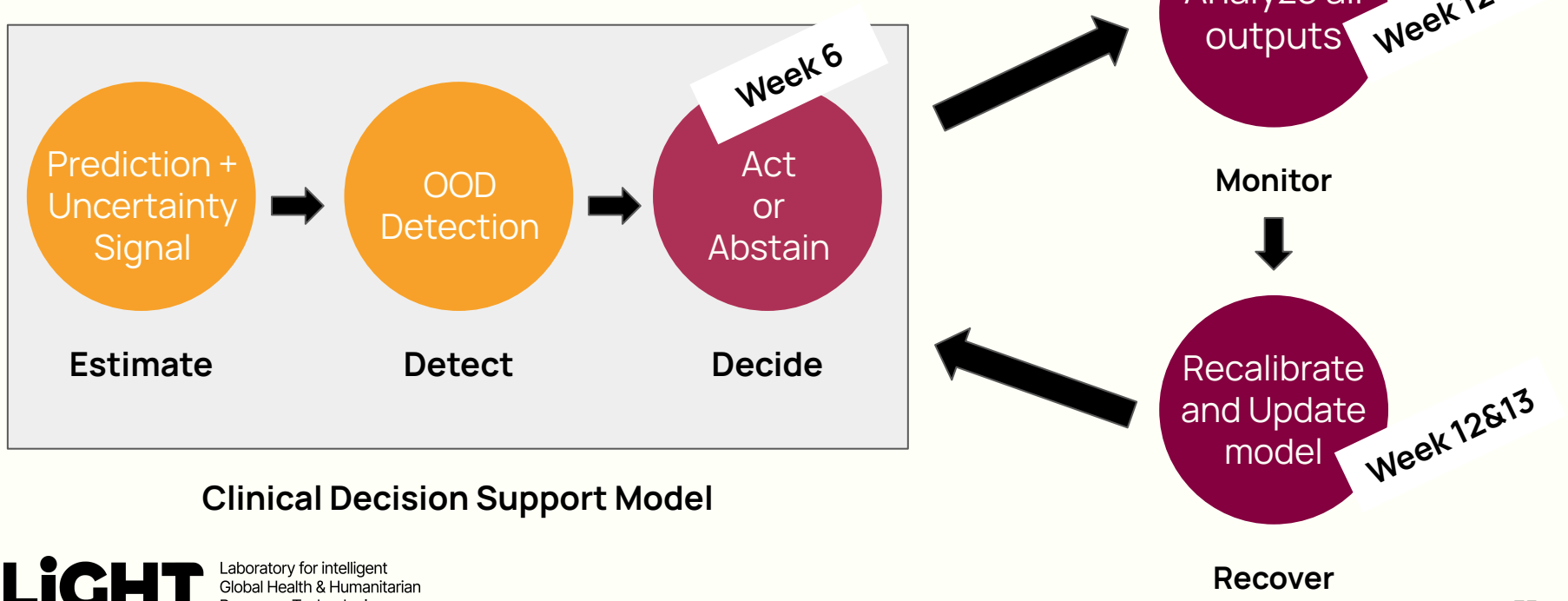
LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Uncertainty as architecture



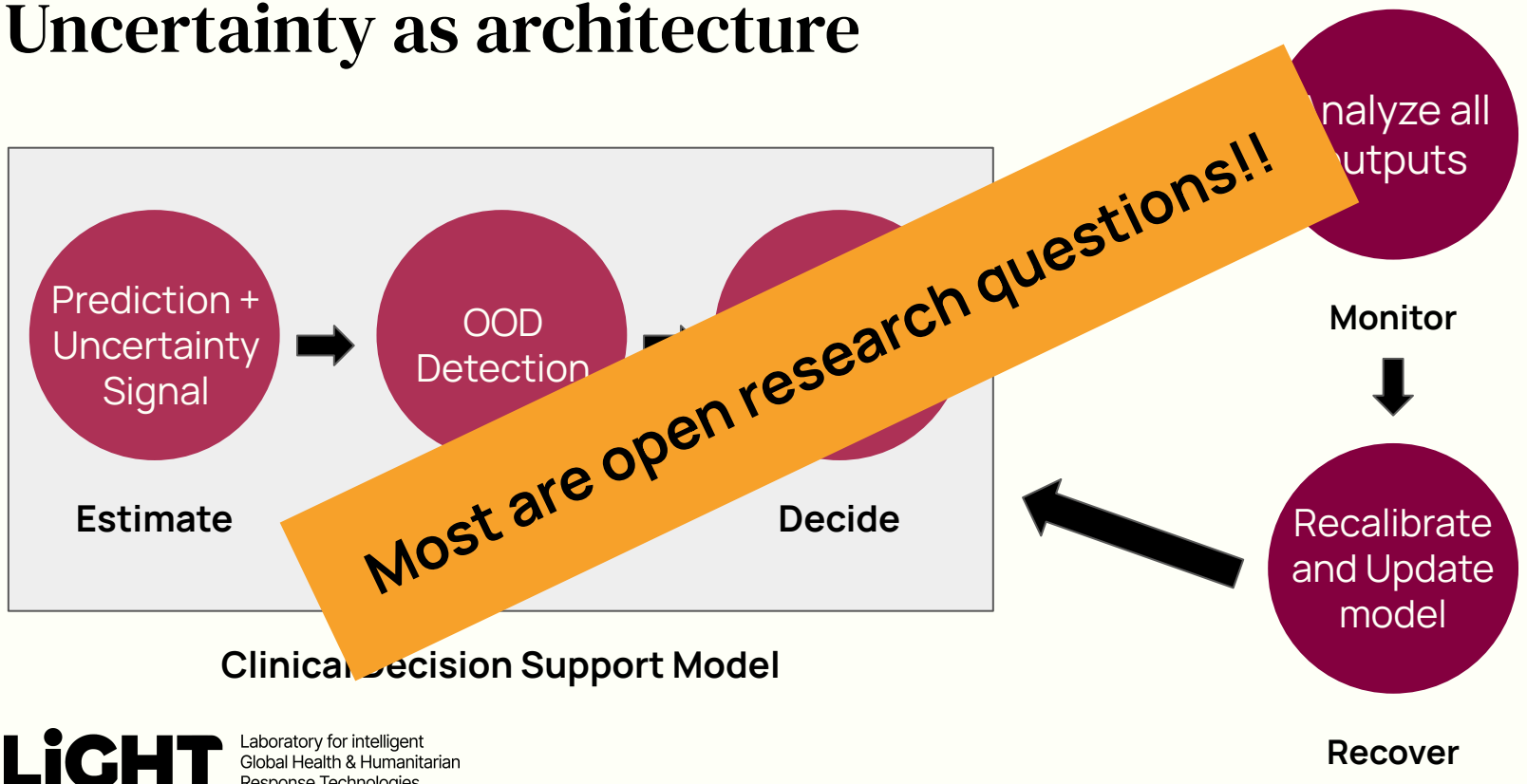
ENGINEERING

Uncertainty as architecture



LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Uncertainty as architecture



Simple Models

→ Today: Uncertainty calibration and OOD detection in (M)LLM

[1]Hernandez-Garcia et al., ICML 2026
Irresponsible AI: big tech's influence on AI research and associated impacts

Simple Models

→ **Today: Uncertainty calibration and OOD detection in (M)LLM**

But:

- Small models can be very effective on a well defined task
- Uncertainty estimation, OOD detection etc. are generally easier with smaller models
- Resist the general push to “scale” in the ML field[1]

→ **Use simple models where possible**

[1]Hernandez-Garcia et al., ICML 2026
Irresponsible AI: big tech’s influence on AI research and associated impacts

Uncertainty

→ Quantifying uncertainty is crucial in medical applications

Uncertainty

→ Quantifying uncertainty is crucial in medical applications

Where does the uncertainty come from?

- **Aleatoric Uncertainty** = intrinsic randomness of the phenomena
- **Epistemic Uncertainty** = lack of knowledge (something we could in principle know for sure) → can be reduced with more data, better models etc.

Aleatoric vs. Epistemic Uncertainty

Example from *"Active Learning"*

Is it a fork or a spoon?

Robot has to decide if it is worth moving closer to get a better look



yes!
epistemic uncertainty



no!
aleatoric uncertainty

Example from: CIS 7000: Trustworthy Machine Learning (Spring 2024)

Aleatoric vs. Epistemic Uncertainty

$$y - f_{\hat{\beta}(Z)}(x) = \underbrace{\left(y - f_{\beta^*}(x)\right)}_{\text{Aleatoric uncertainty}} + \underbrace{\left(f_{\beta^*}(x) - f_{\hat{\beta}(Z)}(x)\right)}_{\text{Epistemic uncertainty}}$$

- **Aleatoric uncertainty:** Error of best possible model f_{β^*}
- **Epistemic uncertainty:** Error of our model $f_{\hat{\beta}(Z)}$ vs. f_{β^*}

Confidence Calibration

- **Calibration:** predicted probabilities = observed frequency
- Makes confidence scores trustworthy
- How for MLLMs?

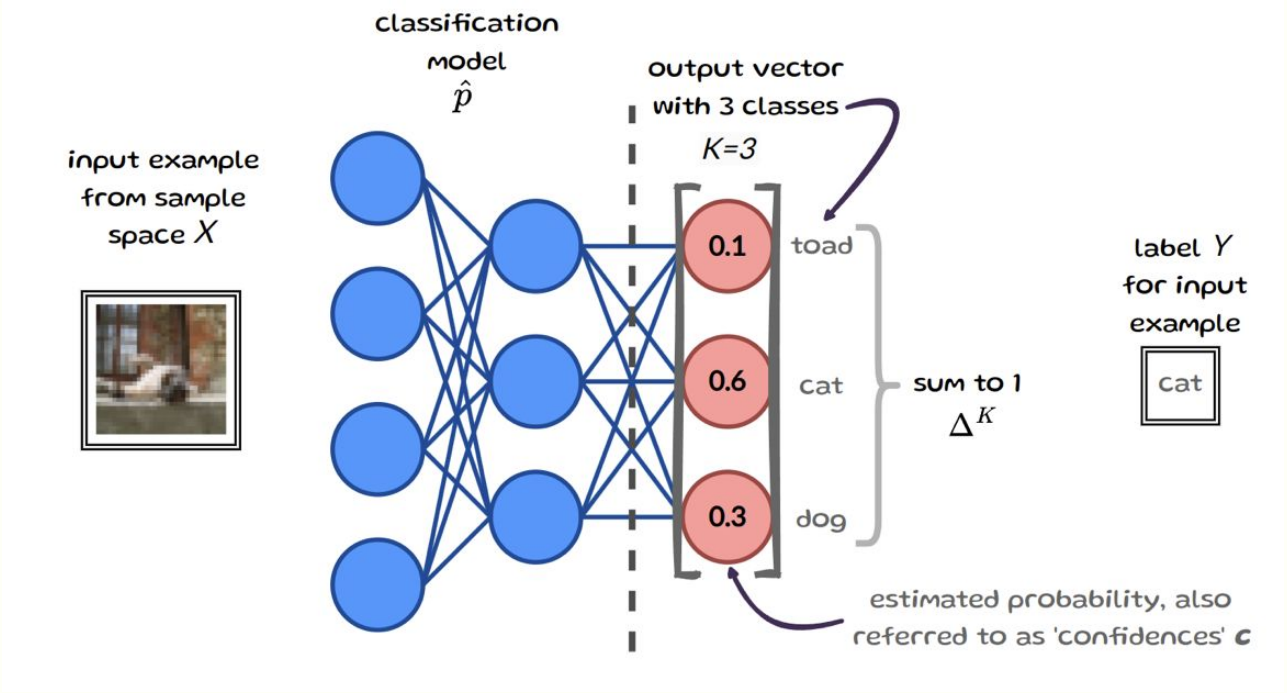
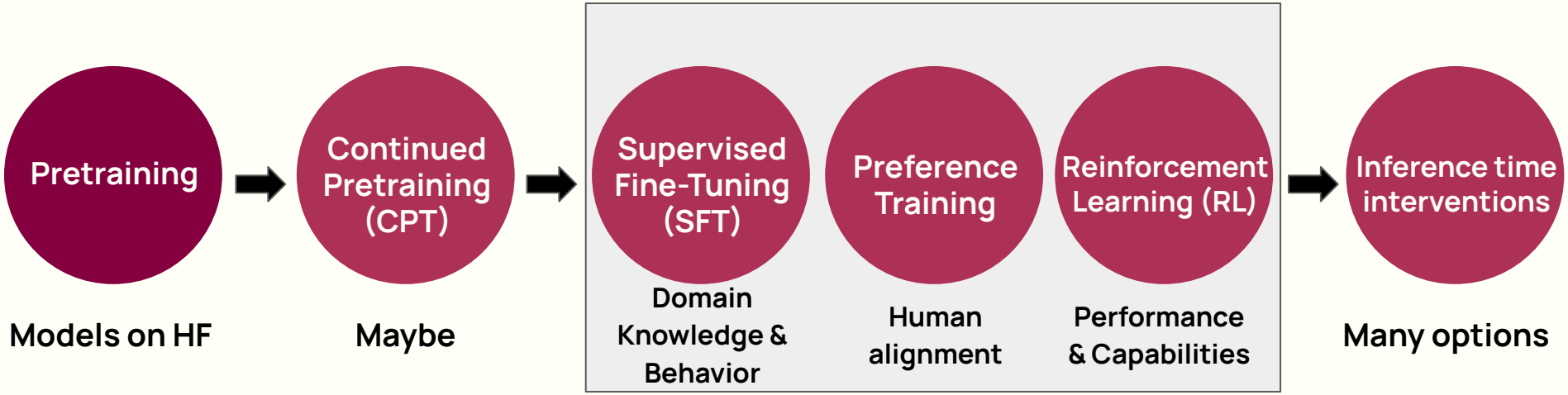
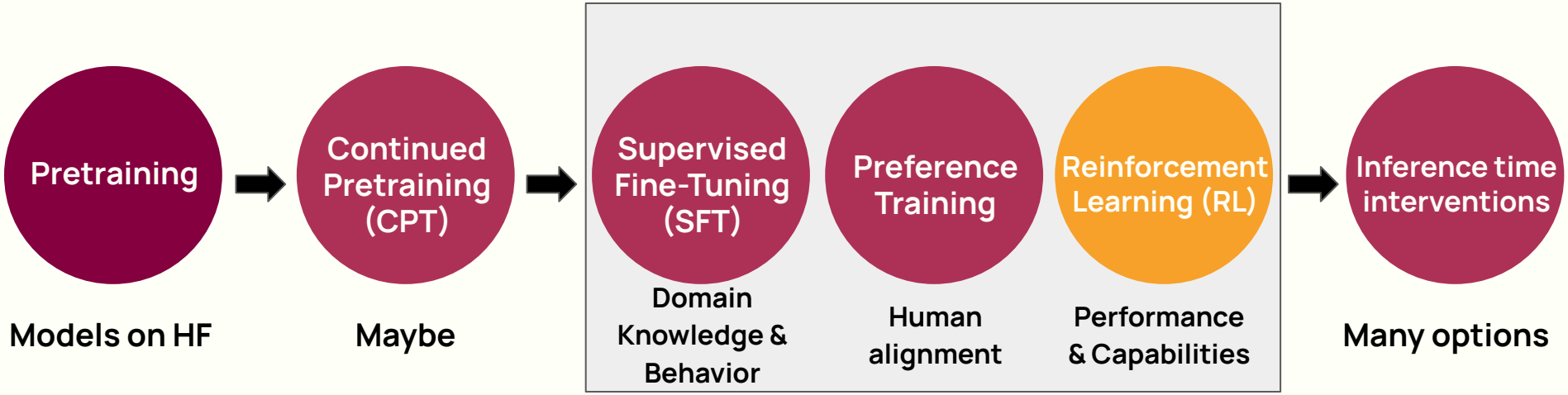


Figure from: [Understanding Model Calibration - A gentle introduction and visual exploration of calibration and the expected calibration error (ECE) | ICLR Blogposts 2025]

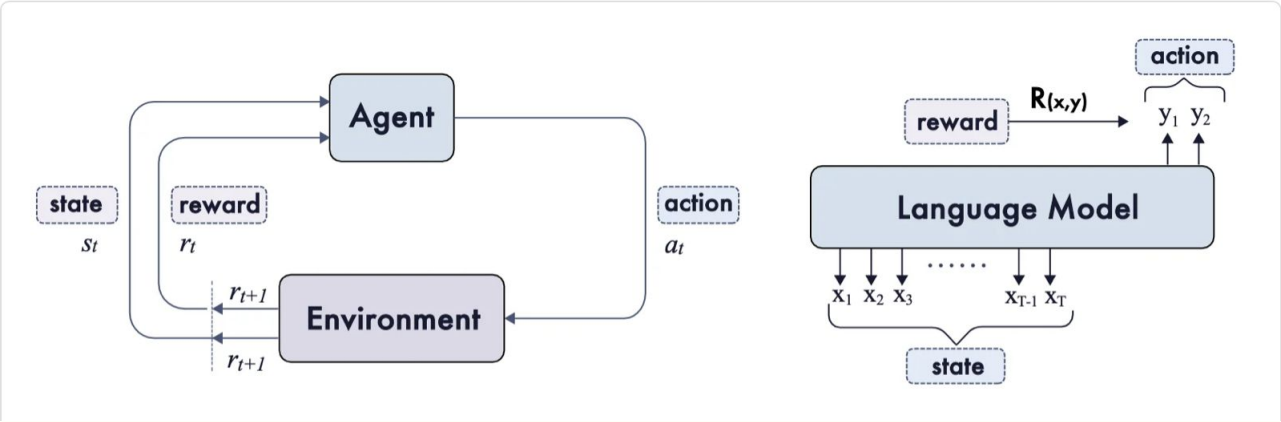
LLM Training pipeline



LLM Training pipeline



Reinforcement Learning for LLMs



Tokens (x, y) = Text or function/tool calls (agentic scenarios)

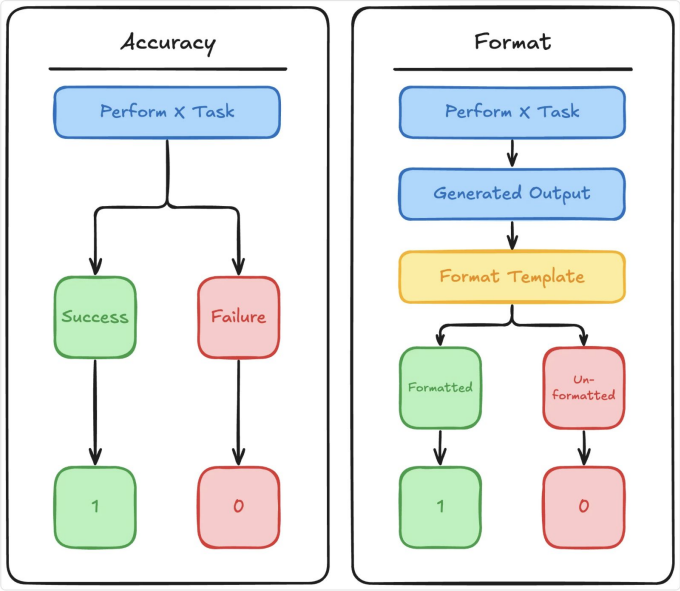
Reward = signal to update weights of the policy (LLM)

Figure from: Reinforcement Learning with Verifiable Rewards for LLMs

Reinforcement Learning from Verifiable Rewards (RLVR)

Reward functions:

- **Accuracy:** Successfully order lab test, correct diagnosis, pass unit test etc.
- **Format**: reasoning (<think> token), json format, length constraints etc.
- **Other**: e.g. LLM-as-a-judge variants



$$R = R_A + R_F$$

Figure from: Reinforcement Learning with Verifiable Rewards for LLMs

Reinforcement Learning from Verifiable Rewards (RLVR)

Issues with RLVR (for uncertainty calibration):

- Same reward for confident answer and lucky guess
- Same penalty for “I don’t know” and wrong answer

→ **Consequence**: RL-trained models tend to be overconfident (poor calibration)

→ **Research question**: Craft a reward that optimizes for accuracy and calibration?

Reinforcement Learning with Calibrated Rewards (RLCR[1])

Idea: Add a calibration score term to the reward

Brier score (BS):

- proper scoring rule
- rewards calibrated and informative confidence
- in $[0, 1]$
- 0 = perfect predictions

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

$$f_t = predicted\ probability$$

$$o_t = outcome$$

[1] Damani&Puri et al., Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty, ICLR 2026

Reinforcement Learning with Calibrated Rewards (RLCR[1])

(a) Example Output

<think> The question asks for the song with which Lulu represented the UK in the 1969 Eurovision Song Contest. […] I need to recall the specific song that Lulu performed for the UK in 1969. </think>
<answer> To Sir With Love </answer>
<analysis> There is a high level of uncertainty in this answer because while Lulu did represent the UK in the 1969 Eurovision Song Contest, the specific song she performed is not widely known or easily recalled. […] but without more specific information, it is difficult to pinpoint the exact song.} </analysis>
<confidence> 0.3 </confidence>

[1] and Figures from Damani&Puri et al., Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty, ICLR 2026

Reinforcement Learning with Calibrated Rewards (RLCR[1])

$$R_{\text{RLCR}}(y, q, y^*) = R_{\text{correctness}}(y, y^*) + R_{\text{Brier}}(y, q, y^*)$$
$$= \mathbb{1}_{y \equiv y^*} - (q - \mathbb{1}_{y \equiv y^*})^2 .$$

y = predicted answer
q = confidence
y* = ground truth

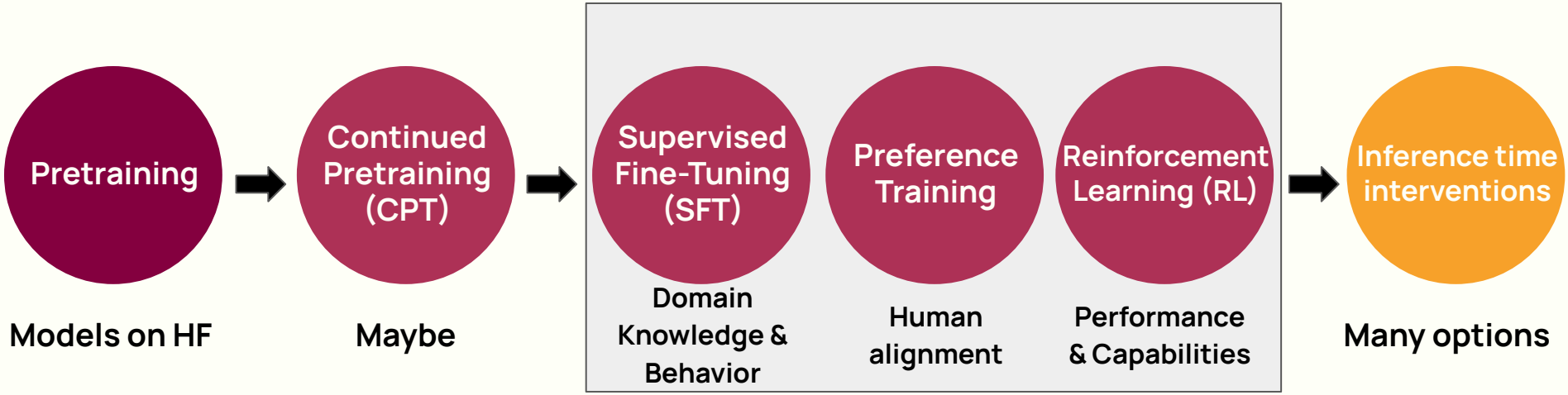
→ Only R_{Brier} leads to **Reward Hacking**: learn to predict a wrong answer with 0% confidence

(a) Example Output

<think> The question asks for the song with which Lulu represented the UK in the 1969 Eurovision Song Contest. […] I need to recall the specific song that Lulu performed for the UK in 1969. </think>
<answer> To Sir With Love </answer>
<analysis> There is a high level of uncertainty in this answer because while Lulu did represent the UK in the 1969 Eurovision Song Contest, the specific song she performed is not widely known or easily recalled. […] but without more specific information, it is difficult to pinpoint the exact song.} </analysis>
<confidence> 0.3 </confidence>

[1] and Figures from Damani&Puri et al., Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty, ICLR 2026

LLM Training pipeline



Uncertainty Quantification for MLLMs

Research question: Training-free uncertainty measure that works for any input modality?

Solution: UMPIRE[1] (Uncertainty using Model Probability Indicators and Response Embeddings)

Key assumption: Sample repeatedly, diversity in outputs is higher if model is unsure



[1] Lau&Dao et al., UMPIRE: Uncertainty Quantification for Multimodal Large Language Models with Incoherence-adjusted Semantic Volume, NeurIPS 2026

ENGINEERING

Uncertainty Quantification for MLLMs

What do RLCR and UMPIRE lack for low-resource medical AI settings?

Uncertainty Quantification for MLLMs

What do RLCR and UMPIRE lack for low-resource medical AI settings?

→ **No subgroup disparity analysis**: Not hard to imagine that for underrepresented populations the models are more uncertain or more confidently wrong

→ Only optimizing for aggregated calibration can **lead to fairness issues** (more safety vs. restricted benefit)

Distribution shifts

A benchmark typically measures:

$$Acc(\theta; \mathcal{D}) = \mathbb{E}_{(x,y^*)\sim\mathcal{D}}[p_\theta(y^*|x)]$$

with

- x, y^*= dataset
- $\mathcal{D}$ = distribution from which the dataset is drawn
- θ = model weights (e.g. the LLM)

→ Distribution shift: $\mathcal{D}$ **changes**

Terminology

Medical	Machine learning
validation data	test set
deployment data	inference time input

Types of distribution shifts

ML Research:

- **OOD (Out-of-distribution)**: e.g. age in train 20 - 45, age in test 10 - 18
- **Out-of-domain**: e.g. RLVR on math data, evaluate on medical data

Types of distribution shifts

ML Research:

- **OOD (Out-of-distribution)**: e.g. age in train 20 - 45, age in test 10 - 18
- **Out-of-domain**: e.g. RLVR on math data, evaluate on medical data

Medical Research:

- **Sub-group shift**: e.g. proportion of socio-economic status changes (not OOD!)
- **Population shift**: e.g. university vs. rural hospital
- **Time shift**: e.g. change in imaging methods (usually unavoidable)
- **Drift**: data-generating process evolves continuously during deployment
- Others: Label/prevalence shift, concept shift, covariate shift…

How to detect shifts?

- **Post-market surveillance** is enforced by regulators (for medical devices)
- **Difficulty**: Detect shift **without access to labels** (outcome often only available too late)
- Shift detection as a **hypothesis test**: H_0: $P = Q$ vs. H_1: $P \neq Q$, with $\alpha = 0.05$, with P = validation data and Q = deployment data
  → many methods to do so
- **Fix shifts: drift-triggered transfer learning, continual learning etc.**
  → Challenge: catastrophic forgetting

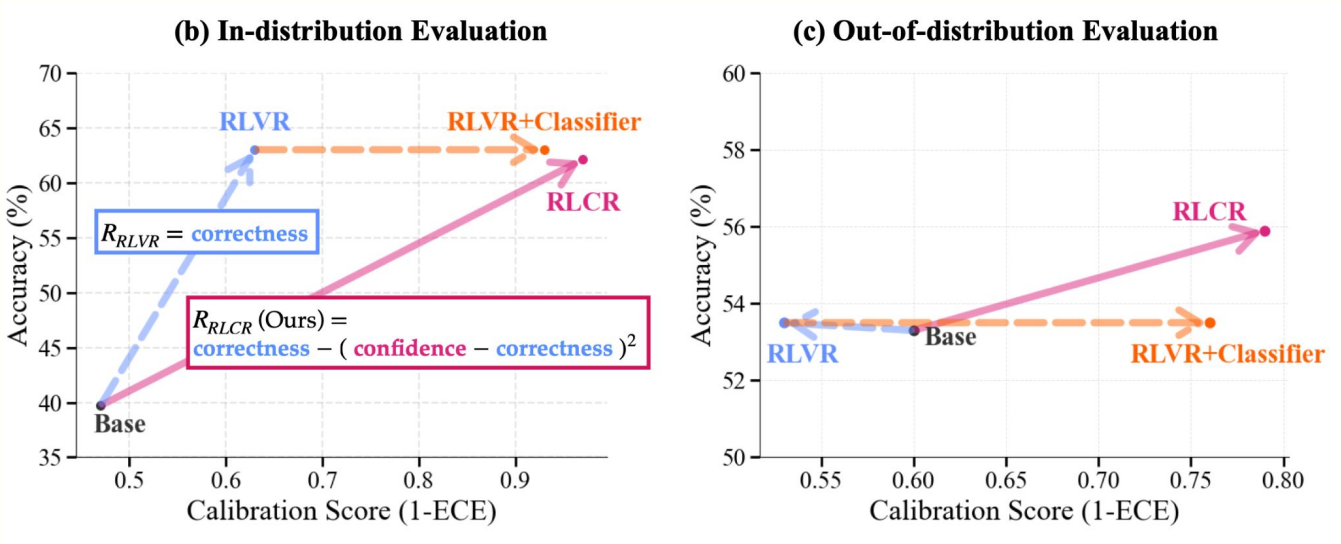
Questions?

Paper

BEYOND BINARY REWARDS: TRAINING LMS TO REASON ABOUT THEIR UNCERTAINTY

Mehul Damani* **Isha Puri*** **Stewart Slocum** **Idan Shenfeld** **Leshem Choshen**
Yoon Kim **Jacob Andreas**
Massachusetts Institute of Technology

Reinforcement Learning with Calibrated Rewards (RLCR[1])



[1] and Figures from Damani&Puri et al., Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty, ICLR 2026

Paper

UMPIRE: Uncertainty Quantification for Multimodal Large Language Models with Incoherence-adjusted Semantic Volume

Gregory Kang Ruey Lau[1,2]* **Hieu Dao**[1]* **Nicole Kan Hui Lin**[1] **Bryan Kian Hsiang Low**[1]
[1] Department of Computer Science, National University of Singapore, Singapore
[2] CNRS@CREATE, 1 Create Way, #08-01 Create Tower, Singapore 138602
{gregorylau,daohieu}@u.nus.edu, dcslkh@nus.edu.sg

Uncertainty Quantification for MLLMs

		Image						Audio		Video	
		VQAv2	OKVQA	AdVQA	MathVista	VQA-RAD	Avg.	SLUE	SpokenSQ.	VidMME	Avg. [All]
AUROC ↑	NC	$75.1_{\pm 0.1}$	$52.8_{\pm 0.7}$	$64.8_{\pm 0.6}$	$74.6_{\pm 0.2}$	$68.6_{\pm 1.5}$	67.2	-	-	-	-
	LN-Ent	$77.0_{\pm 0.2}$	$69.5_{\pm 0.2}$	$64.0_{\pm 0.3}$	$65.8_{\pm 0.3}$	$62.0_{\pm 1.1}$	67.7	$88.2_{\pm 0.5}$	$76.6_{\pm 0.2}$	$69.4_{\pm 1.5}$	71.6
	Sem.Ent	$83.5_{\pm 0.3}$	$70.9_{\pm 0.1}$	$75.9_{\pm 0.2}$	$80.7_{\pm 0.7}$	$75.8_{\pm 0.5}$	77.3	$89.1_{\pm 1.2}$	$81.0_{\pm 0.3}$	$73.1_{\pm 1.5}$	78.7
	Eigen	$85.3_{\pm 0.2}$	$73.6_{\pm 0.1}$	$76.6_{\pm 0.3}$	$81.1_{\pm 0.6}$	$79.2_{\pm 0.6}$	79.2	$89.0_{\pm 1.0}$	$80.7_{\pm 0.2}$	$80.0_{\pm 1.0}$	80.7
	Ours	**87.2**$_{\pm 0.2}$	**75.0**$_{\pm 0.1}$	**78.0**$_{\pm 0.2}$	**82.0**$_{\pm 0.6}$	**80.0**$_{\pm 0.3}$	**80.4**	**89.5**$_{\pm 0.3}$	**81.2**$_{\pm 0.2}$	**81.2**$_{\pm 0.4}$	**81.8**
ECE ↓	NC	$.343_{\pm .008}$	$.484_{\pm .017}$	$.335_{\pm .013}$	**.093**$_{\pm .007}$	$.157_{\pm .007}$	.282	-	-	-	-
	LN-Ent	$.055_{\pm .007}$	**.044**$_{\pm .002}$	$.079_{\pm .007}$	$.136_{\pm .013}$	$.135_{\pm .010}$	.090	$.143_{\pm .020}$	$.072_{\pm .002}$	**.096**$_{\pm .005}$	.095
	Sem.Ent	$.044_{\pm .003}$	$.193_{\pm .043}$	$.157_{\pm .015}$	$.195_{\pm .044}$	$.379_{\pm .018}$	.194	$.236_{\pm .012}$	$.292_{\pm .019}$	$.213_{\pm .014}$	.214
	Eigen	$.056_{\pm .004}$	$.153_{\pm .010}$	$.217_{\pm .005}$	$.325_{\pm .024}$	$.361_{\pm .025}$	.222	$.171_{\pm .007}$	$.244_{\pm .009}$	$.330_{\pm .005}$	.232
	Ours	**.043**$_{\pm .004}$	$.053_{\pm .012}$	**.042**$_{\pm .001}$	$.109_{\pm .003}$	**.104**$_{\pm .012}$	**.070**	**.080**$_{\pm .007}$	**.041**$_{\pm .001}$	$.127_{\pm .025}$	**.075**

[1] Lao&Dao et al., UMPIRE: Uncertainty Quantification for Multimodal Large Language Models with Incoherence-adjusted Semantic Volume, NeurIps 2026

Paper

Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models

Vallijah Subasri, PhD[1,2,3]; Amrit Krishnan, MSc[2]; Ali Kore, BEng[2]; et al

Author Affiliations ⌄ Article Information

[1]Peter Munk Cardiac Centre, University Health Network, Toronto, Ontario, Canada

[2]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada

[3]Genetics and Genome Biology Program, The Hospital for Sick Children, Toronto, Ontario, Canada

[4]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.

Paper

nature > npj digital medicine > articles > article

Article | Open access | Published: 09 May 2024

Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study

Lisa M. Koch, Christian F. Baumgartner & Philipp Berens

npj Digital Medicine **7**, Article number: 120 (2024) | Cite this article

Save article

9867 Accesses | **76** Citations | **11** Altmetric | Metrics

Detecting and Remediating Harn for the Responsible Deployment Models

Vallijah Subasri, PhD[1,2,3]; Amrit Krishnan, MSc[2]; Ali Kore, BEng[2]; et al

Author Affiliations | Article Information

[1]Peter Munk Cardiac Centre, University Health Network, Toronto, Ontario,

[2]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada

[3]Genetics and Genome Biology Program, The Hospital for Sick Children, Toronto, Ontario, Canada

[4]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada

Main papers – TODO:

- Read and understand both medical papers
- Read two blog posts on calibration

- Uncertainty Calibration:
  - Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty
  - UMPIRE: Uncertainty Quantification for Multimodal Large Language Models with Incoherence-adjusted Semantic Volume

- Distribution shifts:
  - Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models
  - Distribution shift detection for the postmarket surveillance of medical AI algorithms: a retrospective simulation study

Building intelligent Global Health and Humanitarian Response Technologies

Thank You

Except where otherwise indicated, original content in this presentation is © 2026 Hartley, Keitel, Klein, Sasu, Suttels, Arni, Kesmen, Nemo and licensed under CC BY 4.0: https://creativecommons.org/licenses/by/4.0/.

Third-party materials are excluded from this license and remain subject to their indicated rights.

LiGHT Laboratory for intelligent Global Health & Humanitarian Response Technologies.