FULCRUM COURSE NOTES · 08

Artificial Intelligence in Medicine

Every specialty, what has actually been built, what works in a clinic rather than a paper, and where it fails. Written for clinicians, students, and researchers entering the field.

FULCRUM

Utsav Poudel
Founder, Fulcrum

ABOUT FULCRUM

Who made this, and why it is free

Fulcrum is a volunteer-run non-profit. We help people from under-resourced regions learn artificial intelligence through one-on-one mentorship, research and publication supervision, project support, workshops, and a guest speaker series.

The premise is simple. Ability is spread evenly across the world; opportunity is not. What separates a capable student in Pokhara or rural Bihar from one at a well-funded lab is rarely intelligence. It is access to supervision, to reviewers, to equipment, and to the unwritten rules nobody writes down.

This deck is part of how we try to close that. It is free, it will stay free, and you may reuse and adapt it for your own teaching under CC BY 4.0. If it is useful to you, come and find us.

Find us

Website
fulcrumgo.github.io

More material
fulcrumgo.github.io/resources

Discord
discord.gg/gbQCGkupdJ

LinkedIn
linkedin.com/company/gofulcrum

Every deck we publish is free to download at fulcrumgo.github.io/resources

PART ONE

Medicine is the hardest place to deploy a model, and the most worth doing

A recommender system that fails shows you a bad film. A clinical model that fails delays a diagnosis. Everything in this deck follows from that difference.

How medicine has always made decisions

Long before machine learning, medicine was already a prediction discipline. It just made its rules explicit.

- **Clinical reasoning.** Pattern recognition by a human who can explain the pattern, then test it.
- **Evidence based medicine.** From the early 1990s, decisions tied to trial evidence rather than authority, with the randomised controlled trial at the top.
- **Statistical risk scores.** Framingham for cardiovascular risk, APACHE II for intensive care mortality. Regression models fitted to cohort data, in use for forty years.
- **Rule based expert systems.** MYCIN, Stanford in the 1970s, recommended antibiotics from roughly six hundred hand written rules and matched specialists. Never used clinically.

Worth remembering when someone calls clinical AI new. What changed is where the rules come from.

Why MYCIN matters

It worked and still failed to reach patients, for reasons of workflow, liability and integration rather than accuracy. Fifty years later those are still the reasons most clinical models never get used.

What current AI does differently

One change, and everything else follows from it.

- **The rules are learned, not written.** Nobody tells the model what a tumour looks like. It infers that from labelled examples, which is why it scales to problems too subtle to write down.
- **It finds signal humans cannot articulate.** A cardiologist cannot read ejection fraction off a standard ECG. A trained network can. That is a genuinely new capability, not a faster version of an old one.
- **It works on raw inputs.** Traditional risk scores need a human to decide which variables matter. Deep models take the pixels or the waveform directly.
- **It updates in days, not years.** A guideline takes a decade from trial to practice. A model can be retrained over a weekend, which is a strength and a governance problem at the same time.

The trade is real. You gain the ability to use signals nobody has named, and you lose the ability to read the rule off the page and argue with it.

Where the two ways of thinking clash

Most friction between clinicians and model builders traces back to one of these four.

- **Cause against correlation.** Medicine asks why a treatment works and demands a mechanism. A model reports what predicts an outcome, which is not the same claim and does not support the same decisions.
- **Explanation.** A clinician can justify a decision to a patient, a colleague and a court. Most models cannot, and post hoc explanation methods describe the model rather than the biology.
- **Evidence standards.** Medicine trusts the prospective randomised trial. Almost all clinical AI evidence is retrospective on data collected for other purposes.
- **Where the error lands.** A guideline that is wrong is wrong visibly and gets revised. A model that is quietly miscalibrated for one group keeps being wrong, in the same direction, without anyone noticing.

None of this argues against using models. It argues for holding them to the standard medicine already holds everything else to, which is the subject of the rest of this deck.

The useful framing

Traditional methods are transparent and limited. Learned models are capable and opaque. Good clinical AI is mostly the work of getting the capability without giving up the accountability.

Why clinical AI is not ordinary machine learning

- **The cost of errors is asymmetric.** A missed cancer and a false alarm are not equally bad, and which is worse depends on the disease, the test that follows, and the patient.
- **The data was never collected for you.** Records exist for billing and for care, not for training. What is missing is rarely missing at random.
- **Distribution shift is the rule.** A model trained on one hospital's scanner, population and protocol degrades at the next hospital. This is expected behaviour, not a bug.
- **Deployment is regulated.** Anything that informs a clinical decision is a medical device in most jurisdictions, with the evidence burden that implies.

A model that reads beautifully in a paper and never touches a patient is the normal outcome. Plan for the gap between the two from the first day.

The one question

Ask early: what clinical decision changes because of this output? If nobody would act differently, the accuracy number does not matter.

What clinical data actually looks like

Six broad kinds, each with its own failure modes.

- **Imaging.** Radiographs, CT, MRI, ultrasound, retinal photographs, whole slide pathology. Large, standardised as DICOM, and the most mature area by far.
- **Waveforms.** ECG, EEG, photoplethysmography, ventilator traces. High frequency, cheap to collect, badly under-used.
- **Structured records.** Labs, vitals, medications, diagnostic codes. Cheap to obtain and full of traps, because codes reflect billing as much as biology.
- **Free text.** Discharge summaries, referral letters, radiology reports. The richest clinical signal and the hardest to handle safely.
- **Omics.** Genomic, transcriptomic, proteomic. Very wide, very few samples, so overfitting is the default state.
- **Sensor and wearable streams.** Continuous glucose, accelerometry, home monitors. Growing quickly, and the data quality is variable.

PART TWO

Specialty by specialty

What exists, with named examples you can go and read. Where a system is in clinical use rather than a research result, it says so.

Radiology

The most commercially advanced area of clinical AI, and the one with the clearest workflow fit.

- **Triage and worklist reordering.** Flagging intracranial haemorrhage or large vessel occlusion so the urgent scan is read first. This is where most deployed value sits.
- **Detection assistance.** Lung nodules on CT, fractures on radiographs, breast lesions on mammography.
- **Quantification.** Volumes, densities and measurements that a human would do slowly and inconsistently.
- **Reporting support.** Draft findings that a radiologist edits and signs.
- **CheXNet** (Rajpurkar et al., 2017) trained on ChestX-ray14 was an early demonstration that a convolutional network could match radiologist performance on specific findings.

Reality check: the great majority of AI medical devices authorised by regulators to date are radiology products, and most of them assist rather than replace a reader.

Why radiology first

Images are already digital, already standardised, already stored centrally, and already have a written report attached as a label. Almost no other specialty starts from there.

Ophthalmology

- **Diabetic retinopathy screening** is the single clearest success story in clinical AI. Gulshan et al. (JAMA, 2016) showed deep learning grading retinal photographs at specialist level.
- **IDx-DR** received FDA authorisation in 2018 as the first autonomous diagnostic AI, meaning it returns a result without a clinician interpreting the image.
- **Why it worked:** a common disease, a cheap standardised image, a shortage of graders, and a clear action when positive (refer to an ophthalmologist).
- **Also active:** glaucoma from optic disc images, age related macular degeneration progression, retinopathy of prematurity.

The lesson generalises. Screening tasks with abundant images, scarce specialists and a simple downstream action are where AI helps most, and those conditions hold across much of South Asia and sub-Saharan Africa.

Pathology

- **Whole slide imaging** turned glass slides into gigapixel files, which made computational pathology possible at all.
- **CAMELYON16 and CAMELYON17** benchmarked detection of breast cancer metastasis in lymph nodes, and top systems matched or exceeded pathologists working under time pressure.
- **Mitosis counting, Gleason grading and tumour proportion scoring** are tasks where human agreement is genuinely poor, so consistency is worth a lot.
- **Prognosis from morphology.** Models can predict outcomes and molecular subtypes from stained tissue alone, which is a claim worth scrutinising carefully.

Slides are enormous, often 100,000 pixels across, so almost every method works on tiles and then aggregates. How you aggregate matters more than the backbone.

Dermatology

- **Esteva et al. (Nature, 2017)** classified skin lesions at dermatologist level using a network pretrained on ordinary photographs, which made the field take transfer learning seriously.
- **Teledermatology triage** is the realistic deployment: sorting which lesions need urgent review when a specialist is weeks away.
- **The known failure:** training sets are dominated by lighter skin. Performance on darker skin is frequently worse and frequently unreported.
- **Image quality varies wildly.** Phone cameras, lighting and focus differ far more than in radiology, where acquisition is controlled.

If you build in this area, report performance by Fitzpatrick skin type. If your dataset cannot support that, say so in the limitations rather than reporting a single number.

A real gap

Dermatology datasets from South Asian and African populations barely exist. Building one carefully is a genuine contribution, and needs a camera rather than a cluster.

Cardiology

- **ECG interpretation** is the oldest form of clinical AI. Rule based ECG software has shipped since the 1970s; deep learning has since raised accuracy substantially.
- **Attia et al. (Nature Medicine, 2019)** trained a network to detect low ejection fraction from a standard 12 lead ECG, finding something a cardiologist cannot see by eye.
- **Atrial fibrillation detection** on consumer wearables moved screening out of the clinic entirely, with all the false positive problems that implies in a low prevalence population.
- **Echocardiography** automation covers view classification, chamber segmentation and ejection fraction estimation.

The ejection fraction result is worth dwelling on. The model is not replicating a human reading; it is extracting a signal humans did not know was in the trace. That is where machine learning earns its place.

Oncology

- **Radiotherapy planning.** Automatic contouring of tumours and organs at risk cuts hours of manual work per patient, and is widely deployed.
- **Treatment response prediction** from serial imaging, and radiomics features extracted from scans.
- **Pathology and genomics integration** for subtyping and prognosis.
- **Trial matching.** Reading a patient record against eligibility criteria, which is tedious, error prone and well suited to language models.

Radiomics has a reproducibility problem worth knowing about. Feature values shift with scanner, reconstruction kernel and slice thickness, so many published signatures do not survive external validation.

Neurology and psychiatry

- **Stroke imaging** is among the highest value deployments anywhere: automated detection of large vessel occlusion and haemorrhage, with direct effect on time to treatment.
- **EEG** for seizure detection and classification, where continuous monitoring produces more data than any human can read.
- **Neurodegeneration.** Volumetric MRI analysis, and speech or language markers of cognitive decline.
- **Psychiatry** has no imaging biomarker, so work concentrates on language, speech and behavioural signals, with correspondingly weaker evidence and much larger ethical exposure.

Mental health prediction from social media and phone data attracts a lot of publication and very little clinical adoption. Ask who is acting on the output and under what consent before building in this space.

Surgery, critical care and anaesthesia

- **Intraoperative video** for phase recognition, instrument tracking and skill assessment.
- **Robotic assistance** remains teleoperated rather than autonomous. Nothing in routine practice cuts on its own.
- **ICU deterioration prediction** from continuous vitals, which is the most studied and most disappointing area in the field.
- **Anaesthesia depth and haemodynamic instability** prediction from waveform monitoring.

The Epic Sepsis Model is the cautionary tale everyone in clinical AI should know. Externally validated by Wong et al. (JAMA Internal Medicine, 2021), it performed far worse than its marketed accuracy while already deployed across hundreds of hospitals.

Why alerts fail

A model with excellent AUC can still be useless if it fires constantly on a low prevalence outcome. Alert fatigue is a deployment failure, not a modelling one, and it is invisible in the metrics.

Primary care, triage and public health

- **Symptom checkers and triage tools** reach far more people than any hospital system, and are held to a much lower evidence standard.
- **Risk scores** for cardiovascular disease, diabetes and readmission, embedded directly in the record.
- **Screening programme support.** Tuberculosis detection on chest radiographs is endorsed by WHO as a computer aided detection option where radiologists are scarce, which matters enormously in Nepal and across South Asia.
- **Outbreak detection and forecasting**, where the modelling is often the easy part and data reporting delays are the real constraint.

Tuberculosis screening deserves attention from anyone working in a high burden country. The disease is common, the image is cheap, the readers are scarce and the treatment exists. Those four facts are what make a task tractable.

Obstetrics, paediatrics and neonatology

- **Fetal ultrasound.** Standard plane detection, biometry and gestational age estimation, which reduces dependence on operator skill.
- **Cardiotocography interpretation**, where human agreement is famously poor.
- **Neonatal monitoring** for sepsis and apnoea from continuous vitals.
- **Growth and developmental screening** from anthropometric data.

Obstetric ultrasound is one of the clearest opportunities in low resource settings, because a handheld probe with automated guidance can put a usable scan in the hands of a midwife rather than a sonographer.

Infectious disease, microbiology and pharmacy

- **Antimicrobial resistance prediction** from genomic or phenotypic data, guiding treatment before culture results return.
- **Microscopy automation** for malaria, tuberculosis and parasites, replacing slide reading that is slow and depends on a skilled microscopist.
- **Drug interaction and dosing support**, including renal dose adjustment, which is a common and preventable source of harm.
- **Pharmacovigilance.** Mining adverse event reports and clinical notes for signals that a trial was too small to catch.

Drug discovery and genomics

- **AlphaFold2** (Jumper et al., Nature, 2021) predicted protein structure to near experimental accuracy and released structures for most known proteins, changing the starting point for structural biology.
- **Halicin** (Stokes et al., Cell, 2020) was identified as an antibiotic by screening chemical libraries with a trained model, then confirmed in the laboratory. A genuine machine led discovery.
- **Variant calling.** DeepVariant reframed genome variant identification as an image classification problem and improved accuracy over hand engineered callers.
- **Target identification, molecular property prediction and generative chemistry** are all active, and all still bounded by wet laboratory validation.

Keep the timeline honest. A candidate identified computationally still needs synthesis, animal work and three phases of trials. Nothing yet compresses the decade that follows the discovery.

Nursing, allied health and the parts nobody writes papers about

- **Nurse staffing and acuity prediction**, which affects patient outcomes more directly than most diagnostic models.
- **Pressure injury and fall risk prediction**, both preventable harms with clear interventions.
- **Rehabilitation.** Movement analysis from video or wearables for physiotherapy adherence and progress.
- **Speech and language therapy**, using automatic speech analysis to track articulation and fluency.
- **Scheduling, bed management and discharge planning.** Unglamorous operational work where small gains compound across a whole hospital.

Nurses generate and act on more clinical data than any other group, and are consulted least often when these systems are designed. If you want an unexploited research area, start there.

PART THREE

What cuts across every specialty

Four applications that are not tied to one organ system.

Clinical documentation and language models

- **Ambient documentation** listens to a consultation and drafts the note. This is the fastest spreading application of language models in medicine, because the pain it solves is universal.
- **Coding and billing support**, which is where hospitals find the money to fund everything else.
- **Information extraction** from unstructured notes to build usable research datasets.
- **Patient facing summaries**, rewriting a discharge letter at a readable level, in the patient's language.

Documentation burden is a leading contributor to clinician burnout. A tool that returns an hour a day to a doctor may improve care more than a diagnostic model with a better AUC.

The safety catch

A model that fabricates a plausible finding in a note creates a permanent error in the record that later clinicians will trust. Ambient tools need a human signing step, every time.

Large language models in clinical practice

- **Benchmark performance is strong.** Med-PaLM 2 and GPT-4 class models score above the passing threshold on USMLE style question sets.
- **Benchmark performance is not clinical safety.** Multiple choice questions have one correct answer, a stated stem and no consequences. Consultations have none of those properties.
- **Where they genuinely help:** summarising, translating, drafting, explaining a diagnosis in plain language, and answering the questions patients are embarrassed to ask.
- **Where they fail:** confident fabrication of citations, drug doses and guidelines, and silent degradation on populations underrepresented in training text.
- **Retrieval grounding** against an approved formulary or guideline set is currently the most reliable way to use them for factual clinical content.

The right framing for a clinician: it is an extremely capable writer with an unreliable memory and no accountability. Useful for transformation, dangerous as a source of fact.

How clinical AI is evaluated, and how it should be

- **AUC is not enough.** Report sensitivity and specificity at the operating point you would actually deploy, and say why you chose it.
- **Calibration matters more than discrimination** when the output is a probability a clinician will act on. A well ranked but badly calibrated model misleads.
- **External validation is the minimum bar.** Internal cross validation on one hospital's data tells you almost nothing about the next hospital.
- **Report by subgroup.** Age, sex, ethnicity, skin tone, device and site. Aggregate performance hides exactly the failures that matter.
- **Prospective and randomised evidence** is rare and is what actually establishes clinical benefit.

Use the reporting standards. CONSORT-AI and SPIRIT-AI for trials, TRIPOD-AI for prediction models, CLAIM for imaging. Reviewers increasingly expect them and they make a paper much harder to dismiss.

Regulation, liability and the route to a clinic

- **Software that informs a clinical decision is a medical device** in the United States, the European Union, the United Kingdom, India and most other jurisdictions.
- **Risk classification drives the evidence burden.** A triage flag and an autonomous diagnosis sit in very different regulatory categories.
- **Locked versus adaptive models.** Regulators have historically required a fixed model, because a system that keeps learning after approval is a system whose behaviour was never approved.
- **Liability is unsettled.** When a model contributes to a missed diagnosis, responsibility is currently distributed between clinician, hospital and manufacturer in ways no court has firmly settled.

Practical consequence for researchers: build the audit trail from the start. Version your data, your weights and your preprocessing. Retrofitting provenance to a two year old project is close to impossible.

PART FOUR

Where it goes wrong

These are not edge cases. Every one of them has happened at scale in a deployed system.

Bias, and how it enters without anyone intending it

- **Obermeyer et al. (Science, 2019)** found a widely used US health algorithm systematically under-referred Black patients, because it used healthcare spending as a proxy for healthcare need. Less had historically been spent on those patients, so the model learned they were less sick.
- **Device bias is real.** Sjoding et al. (NEJM, 2020) documented pulse oximeters overestimating oxygen saturation in patients with darker skin, so any model trained on those readings inherits the error.
- **Missing data is informative.** A test that was never ordered often signals something about the patient and the clinician, and models exploit that in ways that do not transfer.
- **Label noise.** Diagnostic codes reflect what was billed. Radiology reports reflect what was worth writing down. Neither is ground truth.

The pattern is consistent. The model is doing exactly what it was asked to do; the harm comes from the target variable being a bad stand in for what anyone actually cared about.

Shortcuts, drift and other quiet failures

- **Shortcut learning.** Models have learned to detect pneumonia from the hospital that produced the scan, and to detect malignancy from the presence of a surgical ruler in dermoscopy images.
- **Dataset shift over time.** Scanners are replaced, protocols change, coding systems change, and a model silently degrades with no error message.
- **Automation bias.** Clinicians defer to a confident output more than they should, so a wrong model can be worse than no model.
- **Feedback loops.** A model that changes clinical behaviour changes the data it is later retrained on, and can entrench its own errors.

None of this is detectable from a test set. It requires monitoring after deployment, which is the part almost nobody budgets for.

Privacy, consent and data governance

- **Health data is special category data** under GDPR and equivalent regimes, with a higher bar for processing than ordinary personal data.
- **De-identification is weaker than it looks.** Records have been re-identified from combinations of dates, postcodes and rare diagnoses, and faces can be reconstructed from head CT.
- **Consent for care is not consent for research**, and certainly not consent for commercial model training.
- **Federated learning and synthetic data** are useful and partial answers. Neither removes the governance question.

If you are a student with access to clinical data through a placement, get ethics approval before you touch it for a project. Retrospective approval is not a thing, and this ends careers.

PART FIVE

Doing this work without a hospital or a cluster

Which is the position most of our mentees are in.

Open datasets you can start with today

- **MIMIC-IV** and **eICU** on PhysioNet. Deidentified critical care records, free after a credentialing and ethics training step.
- **CheXpert**, **NIH ChestX-ray14** and **PadChest** for chest radiography.
- **ISIC** for dermoscopy, **CAMELYON** for pathology, **MedMNIST** for fast prototyping across many modalities.
- **UK Biobank** and **TCGA** for imaging with genomics and outcomes, subject to application.
- **PhysioNet challenges** for ECG, EEG and physiological waveforms, with published baselines to compare against.

The credentialing step on PhysioNet takes a few hours and is the single highest return administrative task in this field. Do it before you need it.

Compute is not the blocker

Most published clinical AI results are reproducible on a single consumer GPU or a free hosted notebook. Access to data and to a clinician who will tell you the question is worth asking is much harder to get.

Choosing a question that will survive review

- **Work with a clinician from day one.** The most common failure in student medical AI is a technically sound model answering a question no clinician has.
- **Local data is a real contribution.** A carefully built, well documented dataset from an under-represented population is more valuable than another incremental architecture.
- **External validation of a published model** is publishable, useful, and teaches more than building from scratch.
- **Negative results count here.** Showing that a widely used model fails on your population is a genuine finding and an ethical duty.
- **Deployment and workflow studies** are undersupplied because they are unglamorous and slow. That is exactly why there is room.

Open problems worth a career

- **Generalisation across sites** without retraining at every hospital.
- **Calibrated uncertainty**, so a model can say it does not know and a clinician can trust that silence.
- **Evaluation beyond high income settings.** Most benchmarks encode the populations, devices and disease prevalence of a handful of countries.
- **Multimodal integration.** Combining imaging, text, waveforms and genomics is where clinicians already work and where models mostly do not.
- **Monitoring after deployment**, which is close to unsolved and is where patient harm actually accumulates.

Where the gaps are

Well funded groups chase what is measurable on Western benchmarks. Tuberculosis, obstetric ultrasound, snakebite, rheumatic heart disease and malnutrition are under-researched for reasons that have nothing to do with difficulty.

A short reading list

Start with these, and follow their citations rather than reading survey papers.

- Gulshan et al., JAMA 2016. Diabetic retinopathy. The field's clearest early success.
- Esteva et al., Nature 2017. Skin cancer classification, and the case for transfer learning.
- Attia et al., Nature Medicine 2019. Finding what a human reader cannot see in an ECG.
- Obermeyer et al., Science 2019. How a sensible proxy produced racial bias at scale.
- Wong et al., JAMA Internal Medicine 2021. External validation of a deployed sepsis model.
- Jumper et al., Nature 2021. AlphaFold2.
- CONSORT-AI and SPIRIT-AI, Nature Medicine 2020. How to report a clinical AI trial properly.

Read the papers, not the press releases about the papers. The gap between the two in this field is unusually wide.

ABOUT FULCRUM

Bring us a clinical question

We supervise research from question to submission, including work in medicine and health. Free, and open to anyone from an under-resourced region.

Fulcrum is a volunteer-run non-profit helping people from under-resourced regions learn AI through free mentorship, research supervision, and open material like this deck. Founded 2025 in Kathmandu, Nepal. Nothing we do costs anything.

Website
fulcrumgo.github.io

Every deck, free
fulcrumgo.github.io/resources

Apply for mentorship
discord.gg/gbQCGkupdJ

Partnerships
linkedin.com/company/gofulcrum