AI for Early Disease Risk Detection and Clinical Decision Support:

A Practical Framework for Scalable and Responsible Healthcare AI

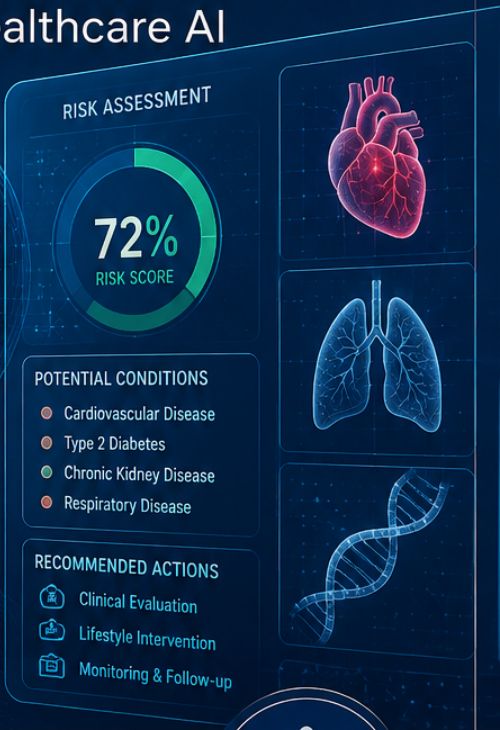
ARTIFICIAL INTELLIGENCE

PREDICTIVE ANALYTICS

EARLY RISK DETECTION

CLINICAL DECISION SUPPORT

RESPONSIBLE AI



Harnessing the power of AI to detect risks earlier, support clinicians, improve outcomes, and build a healthier future.

EMMANUEL AGBEKO ENYO

Published: 2026

AI for Early Disease Risk Detection and Clinical Decision Support

A Practical Framework for Scalable and Responsible Healthcare AI

EMMANUEL AGBEKO ENYO

Published 2026

Copyright

First published 2026.

Medical and Professional Disclaimer

This book is an educational and informational work. It is written for healthcare professionals, technologists, researchers, administrators, and students who want to understand how artificial intelligence can be applied to disease risk prediction and clinical decision support.

It is not medical advice. Nothing in this book should be used to diagnose, treat, cure, or prevent any disease in any individual patient. Readers who have questions about their own health should speak to a qualified clinician.

It is not legal or regulatory advice. Healthcare regulation differs by jurisdiction and changes frequently. Any organisation planning to develop or deploy a clinical AI system should obtain advice from qualified regulatory and legal professionals in the relevant territory.

Several further points deserve emphasis before you read on:

- Predictions produced by an AI system are statistical estimates. They are not diagnoses, and they do not establish that a patient has or will develop a disease.
- AI systems intended for clinical use require appropriate validation, regulatory clearance where applicable, and local evaluation before deployment. Performance reported in a published study is not a guarantee of performance in your setting.
- Clinical decisions remain the responsibility of appropriately qualified healthcare professionals working within their scope of practice.
- Technical standards, regulatory frameworks, and the published evidence base all evolve. Where this book describes the state of a regulation or a body of evidence, readers should confirm the current position against primary sources before acting on it.
- Case studies in Part VII are illustrative constructions used to demonstrate a method. They do not describe specific real patients, and they are not claims about the performance of any particular commercial product.

The author and publisher accept no liability for any loss or damage arising from the use or misuse of the information contained in this book.

Contents

Preface

I want to start by naming the problem this book exists to address.

Over the past decade, healthcare has accumulated an enormous quantity of published machine learning research and a comparatively small quantity of machine learning that actually changes what happens to patients. Systematic reviews of clinical prediction models routinely find hundreds of published models for a single condition, of which a handful have been externally validated and almost none have been evaluated prospectively in the setting where they would be used. The gap between "the model achieved an AUC of 0.87 on a held-out test set" and "the ward runs differently now" is where most healthcare AI projects quietly end.

That gap is not primarily a modelling problem. Teams that fail rarely fail because they chose gradient boosting when they should have chosen a neural network. They fail because the outcome they predicted was not the outcome anyone could act on. They fail because the data that was available at model-training time was not available at the moment of prediction. They fail because the alert fired at 3 a.m. into an inbox nobody reads, or fired so often that clinicians learned to dismiss it without looking. They fail because nobody agreed in advance who would be accountable when the model was wrong.

This book is organised around that observation. It takes the technical material seriously — you will find chapters on feature construction, model selection, discrimination and calibration, and validation design — but it treats those as necessary rather than sufficient. Roughly half the book is about everything that surrounds the model: the data foundation underneath it, the clinical workflow it enters, the equity and safety obligations it carries, and the monitoring that has to continue for as long as it runs.

A word on how to read it. The chapters are sequential and build on each other, but they are also written to be usable on their own. A clinical lead who wants to understand what to ask a vendor can read Chapters 8, 9, 13 and 24 and be substantially better armed. A data sci-

entist new to healthcare who already knows how to train a model will get the most from Chapters 4, 5, 9, 11 and 18, which cover the things that are different about this domain. An executive deciding whether to fund a programme should read Chapter 20 and Chapter 21.

I have tried throughout to be honest about what is established, what is emerging, and what is speculation. Where I describe a finding, I cite a source you can check. Where evidence is thin, I say so rather than dressing up a plausible-sounding claim as a fact. Healthcare AI has suffered from a surplus of confident prediction, and I would rather this book be useful than impressive.

One conviction runs through all of it. The purpose of a clinical prediction model is not to be accurate. It is to help a clinician and a patient make a better decision than they would have made otherwise. Accuracy is instrumental to that, and a model can be highly accurate and completely useless. Keeping the decision at the centre — rather than the algorithm — is the single habit that most reliably separates programmes that work from programmes that produce papers.

— *Emmanuel Agbeko Enyo, 2026*

PART I

Foundations of Healthcare AI

What the technology is, what it is not, and why healthcare is a harder domain than it first appears.

CHAPTER 1

The Evolution of AI in Healthcare

Rule-based expert systems to modern machine learning — and what actually changed.

Artificial intelligence entered medicine long before anyone called it that. In the early 1970s a team at Stanford built MYCIN, a program that recommended antibiotic regimens for bloodstream infections by working through roughly six hundred hand-written rules. In evaluation it performed comparably to infectious disease specialists. It was never used clinically. The reasons were mundane and are worth remembering: it ran on a mainframe, it took a physician half an hour to enter a case, and nobody could answer the question of who would be liable if it was wrong.

Fifty years later, the technology has changed almost completely and those three obstacles — access, workflow cost, and accountability — remain the ones that decide whether a system survives contact with a hospital.

From written rules to learned patterns

The distinction that matters most is where the logic comes from.

A **rule-based system** encodes knowledge that a human expert has already articulated. Someone writes "if the patient's systolic blood pressure is below 90 and lactate is above 2, raise a sepsis concern." The system's behaviour is completely inspectable and completely limited by what its authors thought to write down. Most of the clinical decision support running in hospitals today is still of this type, and much of it works fine. Drug–drug interaction checking, allergy alerts, and dose range warnings are rule-based, and they catch real errors.

A **machine learning system** derives its logic from data. Rather than being told which combination of vital signs indicates deterioration, it is shown many thousands of patient records labelled with what subsequently happened, and it finds the statistical patterns that separate

one group from another. Nobody writes the rule. The rule is inferred, and it may involve interactions between dozens of variables that no clinician would have specified in advance.

That is the source of both the power and the trouble. Machine learning can pick up signals that human experts have not codified. It can also pick up signals that reflect how a particular hospital happened to operate during the period the data was collected, which is not the same thing as clinical knowledge at all.

The terms you will meet

The vocabulary in this field is used loosely, including by people who should know better. Because precision here prevents confusion later, the following definitions are used consistently throughout this book.

Term	What it means in practice
Artificial intelligence	The broad field of building systems that perform tasks associated with human cognition. In healthcare it is now largely a synonym for machine learning, though it properly includes rule-based systems too.
Machine learning (ML)	Methods that learn a mapping from inputs to outputs by fitting parameters to data rather than by explicit programming.
Supervised learning	ML where each training example carries a known outcome label. Nearly all clinical risk prediction is supervised.
Deep learning	ML using neural networks with many layers. Dominant for images, waveforms, and free text; frequently no better than simpler methods for tabular clinical data.
Predictive analytics	The applied practice of using historical data to estimate the probability of a future event. Overlaps heavily with supervised ML.
Generative AI	Models that produce new content — text, images, structured summaries — rather than a score or a label. Large language models are the prominent example.

Term	What it means in practice
Clinical decision support (CDS)	Any system that presents clinicians with information intended to improve a specific decision. Predates AI by decades; AI is one possible engine behind it.

Note that the last row is a category of *purpose*, and the rows above it are categories of *method*. A CDS system may be powered by a rule, a regression equation, a gradient boosted tree, or a language model. The clinical governance obligations attach to the purpose, not the method.

> **IMPORTANT CONSIDERATION**
>
> "AI" on a vendor slide deck carries no technical information. It is compatible with a linear regression fitted to five variables and with a hundred-billion-parameter transformer. When evaluating a product, the useful questions are: what are the inputs, what is the output, what data was it fitted to, and how was it validated. The architecture is close to the last thing you need to know.

What actually changed

Three things converged to make the current wave different from the expert systems era.

The first is data. The HITECH Act of 2009 and equivalent policies elsewhere drove electronic health record adoption from a minority of hospitals to near-universal within a decade. That produced longitudinal, machine-readable records at population scale for the first time. Whatever else can be said about EHRs, they created the substrate that predictive modelling requires.

The second is method. Gradient boosting made tabular prediction reliably strong with modest tuning. Deep learning made perceptual tasks — reading a scan, segmenting an organ, transcribing dictation — tractable in a way they had never been. The 2016 demonstration that a convolutional network could grade diabetic retinopathy from fundus photographs at specialist-level sensitivity was a genuine inflection

point, not because retinopathy screening was the most important problem but because it proved the category was possible.

The third is deployment infrastructure. A model that cannot receive data from the EHR and return a result into the clinician's workflow is a research artefact. Interoperability standards, particularly HL7 FHIR, and the emergence of hospital-side model serving have made real-time integration ordinary rather than heroic.

Where clinical AI actually is today

It is worth grounding expectations in what has been authorised rather than what has been announced. An analysis of 1,016 US Food and Drug Administration authorisations covering 736 unique AI-enabled devices from 1995 to 2024 found that 84.4% were image-based, with radiology accounting for 88.2% of imaging reviews. Signal-based devices — largely cardiovascular — made up 14.5%. Devices drawing on electronic health record data accounted for **0.4%**. Applications the authors classified as "predictive" represented 1.5% of the total.

Read that carefully, because it cuts against the prevailing narrative. The regulated, cleared, commercially available body of clinical AI is overwhelmingly about interpreting pictures. The EHR-driven risk prediction that this book is about is, in regulatory terms, still early. Much of it operates as non-device clinical decision support, developed and deployed by health systems themselves, outside the premarket pathway.

That is not an argument against doing it. It is an argument for taking local validation and local governance seriously, because in most cases there is no regulator standing behind the model you are about to switch on.

Why healthcare is harder than it looks

Healthcare has the properties that make prediction attractive: high-dimensional longitudinal data, outcomes that matter, and decisions made under uncertainty. It also has a set of characteristics that defeat naive application.

Outcomes are recorded, not observed. A diagnosis code in a record means somebody billed for it. Absence of a code means the condition was absent, or undiagnosed, or diagnosed elsewhere, or simply not coded. Models trained on codes learn to predict coding behaviour, which correlates with disease but is not disease.

Data generation is not random. A patient has a troponin measured because a clinician was worried. The presence of the test is informative about clinical suspicion; treating missingness as random destroys that information and can invert conclusions.

The system reacts to its own predictions. If a sepsis model works and triggers earlier treatment, sepsis incidence in the deployed population falls, and the model's apparent accuracy degrades. Successful intervention corrupts the signal the model was trained on. This is not a bug; it is a structural feature of deploying prediction into a system that acts on it.

Errors are asymmetric and consequential. Missing an aortic dissection is not the same magnitude of error as flagging a benign chest pain. Metrics that weight these equally are not measuring anything a clinician cares about.

Populations differ. A model fitted at an academic referral centre encounters different case mix, different testing thresholds, and different documentation practice at a community hospital thirty miles away. Chapter 9 deals with this at length, because it is the single most common cause of unpleasant surprises.

> **KEY TAKEAWAY**
>
> AI in healthcare is not new; what is new is the combination of population-scale electronic data, methods that handle it, and infrastructure to deliver results into workflow. The cleared device landscape remains dominated by imaging. EHR-based risk prediction — the subject of this book — is largely governed locally rather than by regulators, which raises rather than lowers the burden on the organisations deploying it.

CHAPTER 2

From Reactive Healthcare to Predictive Healthcare

Four postures towards disease, and where prediction genuinely helps.

Health systems have historically been built to respond. A patient develops symptoms, presents, is investigated, and is treated. Every component — bed capacity, clinic scheduling, reimbursement, the structure of the medical record — is optimised for that sequence. Prediction proposes something different: acting on a probability before the presentation occurs.

That shift is worth being precise about, because "predictive healthcare" is used to describe four quite different postures.

Four postures towards disease

Reactive care responds to established disease. A patient arrives with crushing chest pain; the system diagnoses and treats a myocardial infarction. This is what hospitals are best at and it is not going away.

Preventive care applies population-level interventions to people who are currently well, based on membership in a broad category. Vaccination schedules, age-based screening programmes, and blood pressure targets are preventive. The logic is epidemiological rather than individual: we do not know which specific person will benefit.

Predictive care estimates an individual's probability of a future event and modulates action accordingly. Two 58-year-old patients with the same nominal risk factors may receive different follow-up intervals because their computed risk differs. The intervention itself is often the same as in preventive care; what changes is who receives it and how intensively.

Personalised care goes further, tailoring the specific intervention to individual characteristics — genomic, phenotypic, behavioural — rather than only tailoring intensity. Oncology, where tumour molecular profil-

ing routinely determines therapy selection, is the most developed example.

These are cumulative rather than sequential. No health system replaces reaction with prediction. It adds a predictive layer that determines where finite attention goes.

Posture	Trigger	Typical question	Example
Reactive	Symptoms present	What is wrong now?	Managing an acute stroke
Preventive	Population category	What should everyone in this group receive?	Age-based colorectal screening
Predictive	Estimated individual risk	Who in this group needs more attention?	Risk-stratified follow-up in CKD
Personalised	Individual biology or phenotype	Which intervention suits this person?	Tumour profiling directing therapy

Risk stratification: the workhorse

Most operational value in healthcare AI comes from risk stratification, which is a less glamorous idea than it sounds. Given a defined population, order it by estimated risk and allocate a scarce resource — a care manager's caseload, a clinic slot, a pharmacist's medication review — to the top of that order.

Stratification is useful precisely when the resource is limited. If a health system can call every patient with diabetes monthly, it does not need a model to decide whom to call. If it can call two hundred of the eight thousand, ranking matters enormously, and even a moderately discriminating model beats calling people alphabetically or waiting for them to deteriorate.

This framing has a practical implication that recurs throughout the book: **the correct performance question is usually not "how accurate is the model" but "among the top N patients this model selects, how**

many would benefit from the action we plan to take." Those are different questions with different answers.

Early warning systems

Inpatient early warning is the other established pattern. Vital signs, laboratory values, and nursing observations are continuously scored, and a threshold triggers escalation to a rapid response team or a senior clinician.

Manual versions have existed for two decades. The National Early Warning Score, used widely in the UK, is an aggregate of seven bedside observations that any nurse can compute on paper. Machine learning versions promise earlier detection by using more variables and their trajectories.

The evidence here is genuinely mixed and deserves to be stated plainly. Some ML early warning systems have improved outcomes in prospective evaluation. Others have performed far worse in the field than in development. The most instructive case is examined in Chapter 9: a widely deployed proprietary sepsis model that achieved an area under the curve of 0.63 in external validation, missed 67% of sepsis cases at its recommended threshold, and generated alerts on 18% of all hospitalisations.

Population health analytics

At the level above the individual, the same models support planning. Which neighbourhoods have concentrations of poorly controlled hypertension. How many patients will progress to needing dialysis in the next three years, and does the region have capacity. Which practices have unusually high rates of avoidable admission.

The analytic machinery is similar; the decision it supports is a resource allocation decision rather than a clinical one, and the governance is correspondingly different. Errors here do not directly harm an individual patient, but systematic underestimation of risk in a particular population will systematically underfund its care.

IMPORTANT CONSIDERATION

A prediction is a statement about a population of similar patients, not a statement about this patient. When a model reports 20% risk, it means that among patients whose recorded characteristics resemble this one's, roughly one in five experienced the outcome. It does not mean this patient has the disease, is developing the disease, or will develop it. Every piece of clinician-facing language in a deployed system should be written to make that distinction unmissable, because the natural reading of a number on a screen is far more definite than the number deserves.

What prediction cannot do on its own

A risk score changes nothing. The value is created entirely by the action taken in response, and that means an intervention must exist, be effective, be available, and be acceptable to the patient.

This is the most common unforced error in healthcare AI programmes. A team builds a model that identifies patients at high risk of readmission, deploys it, and discovers that the health system has no readmission-prevention capacity to offer. The list is accurate and useless. Worse, it consumes clinician attention, which is the scarcest resource in the building.

Before any model is built, three questions should have clear answers:

1. What specific action will be taken when the model flags a patient?
2. Who will take it, and do they have the capacity?
3. Is there evidence that the action improves the outcome in the flagged group?

If the third question cannot be answered affirmatively, the honest position is that the model is a hypothesis-generating tool, and it should be evaluated as one.

KEY TAKEAWAY

Predictive care adds a layer of individualised risk estimation on top of reactive and preventive care; it does not replace them. The value of a prediction is realised only through an available, effective action. Model quality is necessary; an intervention with capacity behind it is what makes the difference. And prediction is never diagnosis — a probability about a group of similar patients is a fundamentally different object from a clinical finding about this one.

PART II

Early Disease Risk Detection

What "early" means, where the data comes from, and how raw records become predictive signal.

CHAPTER 3

Understanding Early Disease Risk Detection

Why earlier is not automatically better, and what conditions are actually amenable.

"Early detection" sounds self-evidently good. It is not, and the reasons it is not are the foundation of everything that follows.

Detecting a condition earlier is valuable only when three things hold: there is a period during which the condition is present or developing but not yet clinically apparent; there is an intervention that works better when applied during that period than after; and the harms of acting on the earlier information are smaller than the benefits. Remove any one of those and earlier detection is neutral at best.

Screening medicine learned this the hard way. Programmes that detected more disease without improving mortality — because the additional cases found were indolent ones that would never have caused harm — taught the field to be suspicious of detection rates as an endpoint. The same discipline needs to be applied to predictive models.

The natural history window

Most chronic disease follows a broadly similar trajectory:

Susceptibility → Subclinical change → Detectable abnormality → Symptoms → Established disease → Complications

Prediction operates in the first two stages, screening in the third, diagnosis in the fourth, and everything downstream is management. The commercial and academic literature blurs these constantly, and readers should hold them apart.

The width of the subclinical window determines how much prediction can buy you. Chronic kidney disease progresses over years and offers a genuinely long runway. Sepsis develops over hours, which makes early warning valuable but leaves little room for error in timing. Acute

conditions with no meaningful prodrome are poor candidates for prediction regardless of how good the model is.

Risk factors, indicators, and trajectories

Three types of signal contribute to risk estimation, and they behave differently.

Risk factors are characteristics associated with elevated probability that are typically stable or slow-moving: age, family history, genotype, smoking status, socioeconomic circumstances. They set a baseline. Their predictive contribution is real but limited, because they are shared by very large numbers of people who will never develop the condition.

Clinical indicators are measurements reflecting current physiological state: blood pressure, HbA1c, estimated glomerular filtration rate, ejection fraction. They are more informative than risk factors because they reflect what is happening now rather than what is statistically likely.

Trajectories are how indicators change over time. This is where machine learning has the clearest advantage over traditional scoring. A single creatinine of 110 μmol/L means one thing in a patient whose value has been stable at 110 for five years and something quite different in a patient who was at 75 eighteen months ago. Classical risk equations, built around a single cross-sectional measurement, cannot see the difference. Models built on longitudinal data can.

> **IMPORTANT CONSIDERATION**
>
> Trajectory features are also where leakage most often creeps in. A rapidly rising creatinine in the three days before a nephrology referral may be predictive of kidney failure, but if the referral itself is what generated the extra measurements, the model has partly learned to detect that a clinician was already worried. Always ask whether a feature would have been available, in that form, at the moment the prediction is supposed to be made. Chapter 5 returns to this.

Where the evidence currently stands

The following is a candid assessment by condition. Readers should treat it as a snapshot and check current literature, because several of these areas are moving quickly.

Condition	Status of risk prediction	Notes
Cardiovascular disease	Established	Multivariable risk equations are embedded in guidelines internationally. The American Heart Association's PREVENT equations, published in 2024, estimate 10- and 30-year risk of cardiovascular disease including heart failure, and notably do not use race as an input. ML approaches show incremental gains over regression in some datasets.
Type 2 diabetes	Established	Risk scores and laboratory-based identification of prediabetes are routine. The main value of ML is in ranking large primary care populations for outreach rather than in improving individual accuracy.
Chronic kidney disease progression	Established	The Kidney Failure Risk Equation, derived by Tangri and colleagues and published in JAMA in 2011, has been validated across more than 30 countries and is used to time nephrology referral and dialysis planning. Among the strongest examples of a risk model changing clinical pathways.
Sepsis and acute deterioration	Mixed	Widely deployed, inconsistently validated. Performance in the field has frequently fallen well short of development-set performance. Genuine successes exist; so do prominent failures.

Condition	Status of risk prediction	Notes
Cancer	Mixed by site	Image-based detection (mammography, dermatology, colonoscopy) is the most mature area of clinical AI. Pre-symptomatic risk prediction from EHR data is far less mature. Multi-cancer early detection blood tests are an active research area with unresolved questions about overdiagnosis.
Alzheimer's disease and cognitive decline	Emerging	Blood-based biomarkers and imaging have advanced substantially, and disease-modifying therapies have changed the calculus of early identification. Prediction from routine clinical data alone remains research-grade.
Infectious disease outbreaks	Established at population level	Syndromic surveillance and forecasting are operationally used by public health agencies. Individual-level prediction of infection is much weaker.

Two general patterns are visible in that table. First, the conditions where prediction is established are the ones with a long subclinical window, a cheap and available measurement, and an intervention with an evidence base. Second, imaging-based tasks have matured faster than record-based tasks, which reflects both the technical tractability of images and the regulatory pathway available to them.

The overdiagnosis problem

Any system that identifies more people as "at risk" creates a population that will be investigated, monitored, and sometimes treated. Some of those people would never have developed the condition. They absorb harm — anxiety, procedural complications, medication side effects, cost, insurance consequences — with no possibility of benefit.

This is not a theoretical concern. It is the central lesson of decades of screening research and it applies with full force to predictive models, which typically flag far more people than they correctly identify. A

model with a positive predictive value of 12% is telling you that seven of every eight flagged patients do not have and will not develop the outcome in the prediction window.

Whether that is acceptable depends entirely on what happens next. If the flag triggers a phone call and a blood test, an 88% false positive rate may be a fine trade. If it triggers a biopsy, it is not.

KEY TAKEAWAY

Earlier detection creates value only when a subclinical window exists, an effective intervention is available within it, and the harms of acting on imperfect information stay below the benefits. Trajectory information — how a patient's measurements are changing — is where predictive models most clearly outperform traditional cross-sectional risk scores. Evidence maturity varies enormously by condition; treat cardiovascular, diabetes, and CKD progression as established and most else as emerging.

CHAPTER 4

Healthcare Data as the Foundation of AI

Sources, structure, and the quality problems that determine everything downstream.

Every serious practitioner arrives at the same conclusion eventually: the model is rarely the constraint. The data is. A team that spends 70% of its effort on data and 30% on modelling is allocating roughly correctly; the reverse ratio is a warning sign.

This chapter is a survey of what healthcare data actually looks like, followed by an honest account of what is wrong with it.

Where the data comes from

Electronic health records. The backbone. Demographics, encounters, diagnoses, procedures, medications, vital signs, laboratory results, and clinical notes. Strengths: longitudinal, broad, already collected. Weaknesses: recorded for care delivery and billing, not for research, and every field carries the fingerprints of the workflow that produced it.

Laboratory results. The most reliable structured data in the record. Numeric, timestamped, produced by calibrated instruments. Caveats that matter: reference ranges and assay methods differ between laboratories, units differ between countries, and a change of analyser can shift values enough to look like a change in the patient.

Medical imaging. Radiology, pathology, ophthalmology, dermatology, cardiology. High information density and the area where deep learning has been most successful. Requires substantial storage and compute, and DICOM metadata is often as important as the pixels.

Claims and administrative data. Excellent coverage of what was billed across all providers a patient visited — which is more than any single EHR sees. Poor clinical granularity, and a lag of months. Diagnos-

is codes on claims reflect reimbursement incentives, which is a form of bias with a direction you can usually predict.

Medication data. Prescriptions, dispensing records, administration records. Prescribing is not the same as dispensing, which is not the same as taking. Inpatient administration records are far more reliable than outpatient prescription lists.

Clinical notes. Where most of the clinical reasoning lives: symptoms, functional status, social circumstances, differential diagnoses, the reason a test was ordered. Unstructured, and requiring natural language processing to use at scale. Language models have made this dramatically more feasible than it was five years ago, though extraction errors and hallucination remain live concerns.

Wearables and remote monitoring. Continuous heart rate, activity, sleep, continuous glucose, home blood pressure, weight scales. Dense and increasingly accurate. Two structural problems: coverage is biased towards younger, wealthier, more health-engaged people, and consumer-grade device accuracy varies by device, by skin tone in the case of some optical sensors, and by activity state.

Genomic data. Increasingly available and clinically decisive in oncology and pharmacogenomics. Polygenic risk scores for common disease are an active research area whose incremental value over conventional risk factors is modest and whose transportability across ancestry groups is a known limitation.

Patient-reported outcomes. Symptom burden, functional status, quality of life. Capture what no laboratory value does. Collected inconsistently, and non-response is strongly associated with the outcomes you care about.

Public health and social determinants data. Deprivation indices, environmental exposure, food access, housing. Often the strongest available proxies for the drivers of health outcomes. Usually available at area level rather than individual level, which limits precision and raises ecological fallacy risks.

Structured versus unstructured

Structured data lives in defined fields with defined types. Unstructured data — notes, reports, images, waveforms — does not. A rough working estimate widely cited in health informatics is that most of the clinically meaningful content of a record is unstructured, and this matches the experience of anyone who has tried to reconstruct a clinical history from coded fields alone.

The practical consequence: a model built only on structured fields is working with a partial view, and the part it is missing is disproportionately the part containing clinical judgement.

The five problems that will consume your project

1. Missingness that is not random. This is the deepest issue. In a research dataset, a missing value usually means an oversight. In a clinical dataset, it usually means a clinician did not think the test was necessary — which is itself strong evidence about the patient's state. Imputing a missing troponin with the population mean discards that evidence and can produce a model that performs worse on sicker patients.

The practical approach is to treat the *fact* of measurement as a feature in its own right, and to be explicit about which mechanism you believe applies:

Mechanism	Meaning	Typical handling
Missing completely at random	Missingness unrelated to anything	Simple imputation is safe
Missing at random	Explained by observed variables	Multiple imputation using those variables
Missing not at random	Depends on the unobserved value itself	No purely statistical fix; model the mechanism, add indicator features, and state the assumption

Most clinical missingness is of the third kind.

2. Quality. Duplicate records, transposed digits, unit confusion, values from the wrong patient, height recorded in centimetres in one system and inches in another. Physiologically impossible values are the easy case because you can filter them; plausible-but-wrong values are the dangerous case.

3. Interoperability. Data arrives from systems that were never designed to talk to each other. HL7 v2 messaging, FHIR resources, DICOM, proprietary vendor schemas, and CSV extracts assembled by hand. Reconciling patient identity across them — record linkage — is a substantial engineering problem in its own right, and errors in it propagate silently.

4. Standardisation. The same concept expressed differently. SNOMED CT, ICD-10, ICD-11, LOINC, RxNorm, ATC, CPT, local dictionaries. Mapping between them is lossy. Common data models — OMOP CDM in particular — exist precisely to make multi-site work possible, and adopting one early costs less than retrofitting it.

5. Bias. Discussed fully in Chapter 13, but it originates here. If a population is under-represented in the training data, under-tested, or has its symptoms differentially documented, the model will inherit those patterns and reproduce them at scale. Data does not record reality; it records what the health system noticed and wrote down.

> **IMPORTANT CONSIDERATION**
>
> Before modelling begins, insist on a written **data availability specification** stating, for every intended feature: which source system it comes from, at what latency it becomes available, its completeness in the target population, and whether it will exist at prediction time in production. Projects that skip this routinely discover late that a key predictor is only populated after the event they are trying to predict, or arrives 48 hours too late to act on. That discovery is much cheaper on paper than after six months of modelling.

A minimum data readiness assessment

Run this before committing to a project.

- ☐ The target outcome can be reliably identified in the data, and someone has manually reviewed a sample of cases to confirm it
- ☐ Cohort inclusion and exclusion criteria are expressible in the available data
- ☐ The number of outcome events is sufficient for the intended model complexity
- ☐ Each candidate predictor's completeness has been measured, not assumed
- ☐ The missingness mechanism has been considered for each predictor
- ☐ Units and reference ranges have been harmonised across sources
- ☐ Patient identity resolution across systems has been validated
- ☐ Data latency in production has been measured end to end
- ☐ Demographic composition of the dataset has been compared to the deployment population
- ☐ Governance approval covers the intended use, including any future model updates

> **KEY TAKEAWAY**
>
> Healthcare data is a byproduct of care delivery and billing, not a research instrument, and every one of its defects has a clinical or administrative explanation. Missingness is usually informative rather than accidental. Build the data foundation, and the specification of what will be available at prediction time, before anyone opens a modelling notebook.

CHAPTER 5

Feature Engineering and Clinical Risk Factors

Turning a longitudinal record into variables a model can use — without cheating.

Feature engineering is where clinical knowledge enters the model. It is also where most subtle failures originate. This chapter covers both.

The pipeline is simple to state:

Raw data → Features → Model → Risk score → Clinical interpretation

Each arrow hides a decision that can invalidate everything downstream.

The prediction point

Before constructing a single feature, fix the moment at which the prediction is made. Everything follows from this.

Define three things precisely:

- **Index time (T_0).** The instant the prediction is generated. Admission? Midnight? Every hour? Completion of a clinic visit?
- **Lookback window.** How far back from T_0 data may be drawn. Twelve months of laboratory history? The current admission only?
- **Prediction horizon.** The period after T_0 during which the outcome counts. The next 6 hours? 12 months? Five years?

No feature may use information generated after T_0. This sounds obvious and is violated constantly, usually in indirect ways: a discharge summary that was written later but timestamped to admission, a laboratory result whose collection time is recorded but whose resulting time is not, a diagnosis code applied retrospectively to the whole encounter.

The consequence is target leakage — the model appears excellent in development and collapses in production, because the information it depended on does not exist yet at the moment of use.

> **IMPORTANT CONSIDERATION**
>
> The single most reliable test for leakage: for each feature, ask a clinician "at 2 p.m. on the day of prediction, could I look this up in the chart?" If the answer is no, or "only after the lab comes back," the feature is either invalid or must be lagged. A model that suddenly loses ten points of AUC when you enforce this rule was never as good as it appeared — you have not broken it, you have found the bug.

Categories of feature

Demographic and stable. Age at index, sex, ethnicity where clinically justified, deprivation index. Cheap, always available, weakly predictive alone. Chapter 13 discusses when including ethnicity is appropriate and when it encodes discrimination.

Vital signs. Rarely used raw. The informative constructions are summaries over the lookback window: most recent value, minimum, maximum, mean, standard deviation, slope, count of measurements, time since last measurement, and count of values outside a physiological range. For deterioration prediction, the derived features usually outperform the raw ones substantially.

Laboratory values. Same treatment, plus two clinical refinements. First, express values relative to the patient's own baseline where one exists — a 30% rise in creatinine from an individual's stable value carries more meaning than an absolute threshold. Second, encode recency, because a haemoglobin from fourteen months ago should not be treated as current.

Medication features. Current active medications by class, count of distinct medications (a reasonable proxy for overall burden), recent initiations and discontinuations, adherence proxies from refill gaps, and specific high-signal agents. New initiation of a loop diuretic is a different signal from long-term stable use.

Diagnosis and comorbidity. Individual conditions from coded history, plus validated composite indices such as Charlson or Elixhauser. Composites are compact and transportable; individual codes carry more information but many more parameters.

Utilisation patterns. Admissions in the prior 6 and 12 months, emergency attendances, outpatient non-attendance, length of stay history, primary care contact frequency. These are frequently among the strongest predictors in readmission and deterioration models. They are also, importantly, partly measures of access rather than illness — a patient who cannot get to clinic looks similar to a patient who does not need to.

Temporal and trajectory features. Slope of a value over the window, variability, area under a curve above a threshold, time since a defining event, and counts of threshold crossings. This category is where longitudinal modelling earns its keep.

Lifestyle and social factors. Smoking, alcohol, physical activity, BMI, housing stability, food insecurity. Clinically important and inconsistently recorded, with the inconsistency itself patterned by who gets asked.

A worked example

Suppose the task is predicting 12-month risk of progression to kidney failure in a primary care population with stage 3 CKD. Index time is a clinic visit; lookback is 24 months; horizon is 12 months.

Raw data	Engineered feature	Clinical rationale
Serial creatinine values	eGFR slope, mL/min/1.73m² per year	Rate of decline is the strongest single predictor of progression
Urine albumin-creatinine ratio	Most recent value, log-transformed; measured vs. not measured	Albuminuria is the key modifiable risk marker; failure to measure is itself informative about care intensity

Raw data	Engineered feature	Clinical rationale
Blood pressure readings	Mean systolic over 12 months; proportion above 140	Chronic exposure matters more than a single reading
Medication list	ACE inhibitor / ARB / SGLT2 inhibitor active at index	Reflects both treatment and, indirectly, clinician assessment
Encounter history	Nephrology contact in prior 24 months	Marks patients already under specialist care — must be handled deliberately, since it is a treatment marker, not a disease marker
Laboratory panel	Serum potassium, bicarbonate, phosphate, haemoglobin most recent	Metabolic complications track disease severity

The nephrology-contact row illustrates a recurring dilemma. It will improve discrimination. It also encodes a clinician's prior judgement, meaning the model partly learns to imitate referral decisions rather than to predict biology — and it will behave differently in a system with different referral thresholds. There is no universally right answer; there is only an obligation to make the choice consciously and document it.

Practical guidance

Prefer features a clinician can name. Not because interpretable models are always better, but because features with clinical meaning survive distribution shift more gracefully and can be debugged when something goes wrong.

Beware proxies for care access. Utilisation features, test frequency, and documentation completeness all measure engagement with the health system alongside illness. Models built heavily on them will systematically under-rank patients who have less access — precisely the patients most in need.

Fit transformations inside the cross-validation fold. Scaling parameters, imputation models, and category encodings computed on the full dataset before splitting leak test information into training. This inflates reported performance and is one of the most common methodological errors in published clinical ML.

Prune aggressively. A model with 40 well-chosen features is usually within noise of one with 400, is far easier to validate, degrades more gracefully, and is vastly easier to explain to a governance committee.

Version the feature definitions. Store the code that generates features alongside the model. A model whose feature definitions cannot be reproduced exactly cannot be revalidated, and revalidation is not optional.

> **KEY TAKEAWAY**
>
> Fix the prediction point first; every feature decision depends on it. Trajectory and variability features are where longitudinal data adds value over cross-sectional risk scores. Guard obsessively against leakage — the "could I look this up right now?" test catches most of it. And recognise that some strongly predictive features encode clinician behaviour or system access rather than patient biology, which is acceptable only when you know you are doing it.

PART III

Building Predictive Healthcare AI

Choosing methods, running the development process, and measuring performance in ways that mean something clinically.

CHAPTER 6

Machine Learning for Disease Risk Prediction

A practitioner's guide to method selection, without the hype.

There is a persistent belief that clinical prediction is limited by algorithm choice. The evidence does not support it. A frequently replicated finding in tabular clinical prediction is that well-specified logistic regression, modern gradient boosting, and a tuned neural network land within a narrow band of each other, and that the differences between them are smaller than the differences produced by better outcome definition, better cohort construction, or more data.

Choose the method that fits the data structure, the interpretability requirement, and the team's ability to maintain it. Then spend the time you saved on validation.

The methods that matter

Logistic regression. Models the log-odds of the outcome as a linear combination of predictors. Coefficients are directly interpretable as odds ratios. Naturally produces calibrated probabilities when correctly specified. Handles small datasets well and extends cleanly with regularisation (ridge, lasso, elastic net) when predictors are numerous or collinear.

Use it as the default and the benchmark. If a complex model cannot beat a properly built regression by a clinically meaningful margin, deploy the regression. Nearly every risk equation embedded in clinical guidelines — cardiovascular risk, kidney failure risk, obstetric risk — is a regression, and this is not an accident of history.

Decision trees. Recursive binary splits producing a flowchart. Trivially interpretable, handle non-linearity and interactions automatically, and unstable — small data changes produce different trees. Rarely deployed alone; important as a building block.

Random forests. Many trees fitted to bootstrapped samples with random feature subsets, averaged. Robust, hard to overfit badly, little tuning required. Probability estimates tend to be poorly calibrated out of the box and generally need post-hoc correction.

Gradient boosting. Trees fitted sequentially, each correcting the previous ensemble's errors. XGBoost, LightGBM and CatBoost are the standard implementations. Consistently among the strongest performers on tabular clinical data, handle missing values natively, and are the reasonable default when regression is insufficient. They will overfit if left untuned, and their probability outputs usually need recalibration.

Support vector machines. Find a maximum-margin separating boundary, optionally in a transformed space. Historically important, effective with many features and few observations. Do not produce probabilities natively and scale poorly to large datasets. Now uncommon in this application.

Neural networks and deep learning. Layered non-linear transformations learned end to end. Dominant where the input has structure the network can exploit — images, waveforms, sequences, free text. On ordinary tabular clinical data, they typically match rather than beat gradient boosting while costing more to build, tune, and explain.

Specific architectures earn their place for specific data:

- **Convolutional networks** for imaging.
- **Recurrent networks and transformers** for irregularly sampled time series and clinical text.
- **Transformer-based language models** for extracting structured facts from notes, summarising histories, and generating draft documentation.

Time-series and survival methods. Where the question is *when*, not just *whether*, survival models are the correct family. Cox proportional hazards for interpretable time-to-event modelling with censoring; random survival forests and gradient-boosted survival models where non-linearity matters. Censoring — patients who leave the dataset before

the outcome occurs — is pervasive in healthcare and is handled properly by survival methods and badly by naive classification.

Choosing

Situation	Reasonable first choice
Tabular data, need transparent coefficients, regulatory or guideline context	Regularised logistic regression
Tabular data, performance is the priority, interpretability via post-hoc methods acceptable	Gradient boosting
Outcome timing matters, censoring present	Cox model or survival forest
Images, waveforms, or free text as primary input	Deep learning with an appropriate architecture
Irregular longitudinal sequences with many time points	Recurrent or transformer sequence model
Fewer than a few hundred outcome events	Regression with strong regularisation, and consider whether to proceed at all
Very high stakes, need mechanistic auditability	Regression or a constrained model, accepting some performance loss

Sample size

Rules of thumb such as "ten events per variable" have been superseded by formal sample size calculations for prediction models, developed by Riley and colleagues, which account for the expected outcome prevalence, the number of candidate predictors, and the anticipated model performance. These typically demand considerably more data than the old heuristics, particularly for models intended to produce well-calibrated probabilities rather than only to rank patients.

The uncomfortable implication is that many published clinical prediction models are developed on datasets too small to support their complexity. If your event count is in the low hundreds and you are considering a model with dozens of parameters, the honest options are to simplify the model, obtain more data, or explicitly frame the work as exploratory.

Class imbalance

Clinical outcomes are usually rare. Sepsis in perhaps 5–7% of admissions; progression to kidney failure in a small fraction of a CKD cohort; 30-day readmission in 10–20%.

Aggressive rebalancing — heavy oversampling, SMOTE, or class weighting — improves apparent performance on threshold metrics and systematically destroys calibration, because it changes the outcome prevalence the model believes it is operating in. Since calibrated probabilities are exactly what clinical decision-making needs, this is a bad trade in most healthcare applications.

The better approach is to keep the natural prevalence, evaluate with precision-recall alongside ROC, and choose an operating threshold based on the clinical cost of each error type. If rebalancing is used for optimisation reasons, recalibrate afterwards on data with the true prevalence.

IMPORTANT CONSIDERATION

Be sceptical of any comparison claiming one algorithm is superior on a single dataset with a single split. Differences of one or two points of AUC are frequently within the variation produced by re-running the split with a different random seed. Meaningful comparisons report confidence intervals, use repeated cross-validation, and — most importantly — compare against a clinically sensible baseline, such as the existing risk score or the clinician's own judgement. A model that beats nothing has demonstrated nothing.

Where generative models fit

Large language models occupy a different position in this pipeline and it is worth being precise about it. They are strong at reading and writing, which in this context means extracting structured facts from unstructured notes, summarising a long record, drafting documentation, and rendering an explanation in readable prose.

They are not, in their general-purpose form, calibrated risk estimators. Asking a language model "what is this patient's probability of sepsis" produces a fluent number with no reliable statistical relationship to the truth. The productive architecture keeps them in the roles they are good at: as a feature extractor upstream of a risk model, or as an explanation layer downstream of one. Chapter 22 examines this further.

> **KEY TAKEAWAY**
>
> Method choice matters less than practitioners expect. Start with regularised logistic regression as a benchmark, move to gradient boosting for tabular data when it earns its place, use deep learning where the input has exploitable structure, and use survival methods when timing and censoring matter. Calculate sample size formally. Resist aggressive rebalancing, because calibration is what clinical decisions need.

CHAPTER 7

Developing a Healthcare Risk Prediction Model

The thirteen-step workflow, and the checklist that goes with it.

This chapter is the operational core of the book. It sets out the sequence a competent team follows, with the failure mode attached to each step, because knowing what goes wrong is more useful than knowing what should happen.

Step 1 — Define the clinical problem

Write a single sentence naming the decision the model will inform, the person who makes it, and what changes as a result.

> *"For adults with stage 3 CKD in primary care, identify those at high risk of progression to kidney failure within two years, so the practice pharmacist can prioritise medication optimisation and timely nephrology referral."*

If you cannot write this sentence, stop. The absence of it is the leading cause of technically successful, clinically pointless models.

Failure mode: starting from "we have this data, what can we predict?"

Step 2 — Define the target outcome

Translate the clinical concept into a computable definition: which codes, which laboratory thresholds, which time windows, which exclusions. Then have a clinician manually review a random sample of records the definition labels positive and negative, and measure how often it is right.

Failure mode: using a convenient proxy. Predicting healthcare cost as a stand-in for health need is the canonical example, and Chapter 13 explains why it produced one of the most consequential bias failures on record.

Step 3 — Identify the population

Specify inclusion and exclusion criteria, the index event, and the observation windows. Exclude patients for whom the prediction is meaningless or the intervention inapplicable — those already under specialist care, those receiving palliative care, those below an age threshold.

Failure mode: a training cohort that does not resemble the deployment population, most often because convenient exclusions removed the complex patients.

Step 4 — Collect data

Extract with reproducible, version-controlled code. Record extraction dates, source system versions, and every filter applied. Confirm governance approval covers the intended use.

Failure mode: a one-off manual extract that cannot be reproduced when the model needs updating.

Step 5 — Clean and preprocess

Harmonise units, resolve duplicates, remove physiologically impossible values, standardise codes, and resolve patient identity across sources. Document every rule. Quantify what each rule removed.

Failure mode: silent exclusion. A cleaning rule that drops 8% of records is a cohort decision, and if those records are concentrated in one demographic group it is a fairness decision too.

Step 6 — Engineer features

Apply Chapter 5. Fix the prediction point, build features from the lookback window only, and audit for leakage.

Failure mode: leakage, in all its varieties.

Step 7 — Split the data

Split before anything else touches the data. For healthcare, split by *patient*, never by record — the same patient appearing in training and test inflates performance dramatically.

Prefer a temporal split for the final held-out evaluation: train on earlier periods, test on a later one. This approximates the real deployment situation, in which a model fitted to the past is applied to the future, and it will reveal temporal drift that a random split hides. Where multiple sites are available, hold one out entirely.

Failure mode: random record-level splitting, which is the single most common source of implausibly good published results.

Step 8 — Train the model

Tune hyperparameters using cross-validation *within the training set only*. Log every experiment: data version, feature version, hyperparameters, random seed, metrics.

Failure mode: tuning against the test set through repeated peeking. After the fifth look, it is a validation set and no longer an unbiased estimate of anything.

Step 9 — Validate

Evaluate on the untouched held-out set, once. Report discrimination, calibration, and clinical utility together. Break results down by subgroup. If external data exists, use it.

Failure mode: reporting only AUC on internal data, which tells you almost nothing about deployment behaviour.

Step 10 — Evaluate performance

Apply the full metric set of Chapter 8, with confidence intervals, and against the incumbent — whether that is an existing score or unaided clinical judgement.

Failure mode: absence of a comparator.

Step 11 — Assess clinical utility

Choose the operating threshold from consequences, not from the F1-maximising point. At the chosen threshold, state plainly: how many patients are flagged per week, what proportion of them will experience the outcome, how many events are missed, and what the required workload is. Decision curve analysis is a useful formal tool here.

Failure mode: a threshold that generates an alert volume the service cannot absorb.

Step 12 — Prepare for deployment

Integration design, latency requirements, failure behaviour, user interface, escalation pathway, documentation, training, and the monitoring plan. Covered in Part VI.

Failure mode: treating deployment as an engineering afterthought rather than a design problem with clinical content.

Step 13 — Monitor

Performance, calibration, subgroup behaviour, input distributions, alert volume, and user response, continuously. Define the triggers for revalidation and retirement before go-live.

Failure mode: no monitoring. A model that is not monitored is an unmeasured clinical intervention.

Reporting standards

Do not invent your own reporting structure. Two international standards cover this work and using them makes evaluation by others — including your own governance committee — enormously easier.

TRIPOD+AI (Collins et al., *BMJ*, 2024) is the reporting guideline for clinical prediction models developed with regression or machine learn-

ing. It replaces the original 2015 TRIPOD statement and adds AI-specific items covering fairness, data provenance, and code availability.

PROBAST+AI (Moons et al., *BMJ*, 2025) is the companion tool for assessing quality, risk of bias, and applicability in prediction model studies. Use it as a self-assessment before publication or deployment, not just as a critique of others' work.

Where the model will be evaluated in a trial, **SPIRIT-AI** and **CONSORT-AI** (both *Nature Medicine*, 2020) cover protocols and reports respectively, and **DECIDE-AI** (*Nature Medicine*, 2022) covers the early-stage live clinical evaluation that sits between offline validation and a full trial.

Healthcare AI Development Checklist

Problem definition

- ☐ Clinical decision named, with the decision-maker identified
- ☐ Specific action defined for flagged patients
- ☐ Capacity to deliver that action confirmed
- ☐ Evidence that the action improves outcomes in the flagged group, or explicit acknowledgement that this is unproven
- ☐ Existing standard of care documented as the comparator

Data and cohort

- ☐ Outcome definition computable, and manually validated on a sample
- ☐ Inclusion and exclusion criteria specified before extraction
- ☐ Index time, lookback window, and prediction horizon fixed
- ☐ Extraction code version-controlled and reproducible
- ☐ Cleaning rules documented with counts of affected records
- ☐ Cohort demographics compared against the deployment population
- ☐ Sample size adequacy assessed formally

Modelling

- ☐ Split performed at patient level before any preprocessing
- ☐ Temporal or external holdout used for final evaluation

- ☐ All transformations fitted inside training folds only
- ☐ Leakage audit completed feature by feature
- ☐ Baseline model (regression or existing score) fitted for comparison
- ☐ Hyperparameter search confined to the training set
- ☐ Experiments logged with seeds and data versions

Evaluation

- ☐ Discrimination reported with confidence intervals
- ☐ Calibration plotted, not merely summarised by a test statistic
- ☐ Precision-recall reported alongside ROC
- ☐ Performance at the intended operating threshold stated in absolute counts
- ☐ Subgroup performance reported across relevant populations
- ☐ Clinical utility assessed, ideally with decision curve analysis
- ☐ External validation performed, or its absence explicitly declared

Readiness

- ☐ Model documentation produced in a standard format
- ☐ Reporting follows TRIPOD+AI
- ☐ Risk of bias assessed against PROBAST+AI
- ☐ Failure behaviour specified for missing or malformed input
- ☐ Monitoring plan written with thresholds and named owners
- ☐ Retirement criteria defined
- ☐ Governance and, where applicable, regulatory review completed

> **IMPORTANT CONSIDERATION**
>
> Steps 1 to 3 determine more about the outcome of the project than steps 8 to 10. Teams under time pressure compress the definitional work and expand the modelling work, which is exactly backwards. If a programme has three months, spending the first six weeks on problem definition, outcome validation, and cohort construction is not slow — it is the only way the remaining time produces anything usable.

KEY TAKEAWAY

The development workflow is thirteen steps, and the ones most often rushed are the first three and the last two. Split by patient and by time. Fit every transformation inside the fold. Report against TRIPOD+AI and self-assess against PROBAST+AI. Define monitoring and retirement criteria before go-live, not after.

CHAPTER 8

Evaluating AI Model Performance

Why accuracy is the least useful number you can report.

Start with an example that makes the point better than any argument.

A hospital builds a model to predict in-hospital cardiac arrest. Roughly 1 in 200 admissions experience one. The model is evaluated and reports **99.5% accuracy**.

The model predicts "no arrest" for every patient. It has never identified a single case. It is 99.5% accurate because the outcome is rare.

Accuracy — the proportion of predictions that are correct — is nearly meaningless when classes are imbalanced, which in clinical prediction is almost always. It should essentially never be the headline metric of a clinical model.

The confusion matrix

Everything else derives from four counts.

	Outcome occurred	Outcome did not occur
Model flagged	True positive (TP)	False positive (FP)
Model did not flag	False negative (FN)	True negative (TN)

In clinical terms, a **false positive** is a patient investigated, monitored, or treated unnecessarily — cost, anxiety, procedural risk, and clinician time. A **false negative** is a patient who was not flagged and who deteriorated — the failure the system existed to prevent.

These are not equivalent, and no metric that treats them as equivalent is measuring clinical value.

The metrics that matter

Metric	Definition	Question it answers
Sensitivity (recall)	TP / (TP + FN)	Of patients who had the outcome, what fraction did we catch?
Specificity	TN / (TN + FP)	Of patients who did not, what fraction did we correctly leave alone?
Precision (PPV)	TP / (TP + FP)	Of patients we flagged, what fraction actually had the outcome?
NPV	TN / (TN + FN)	Of patients we did not flag, what fraction were genuinely fine?
F1 score	Harmonic mean of precision and recall	A single balanced summary — convenient, and clinically arbitrary
ROC-AUC	Area under the sensitivity vs. 1–specificity curve	How well does the model rank patients, across all thresholds?
PR-AUC	Area under the precision-recall curve	Same, but sensitive to rare outcomes — usually more informative here

Two properties deserve emphasis.

Sensitivity and specificity are properties of the model. PPV and NPV are properties of the model *and the population*. Move a model from a high-prevalence intensive care population to a low-prevalence ward population, and sensitivity holds while PPV collapses. This is Bayes' theorem operating, not model failure, and it explains a large share of disappointing deployments.

ROC-AUC is optimistic for rare outcomes. Because the vast majority of patients are negatives, a large absolute number of false positives barely moves the false positive rate. Precision-recall makes the same errors visible. Report both.

Worked example

Ten thousand admissions. Sepsis develops in 500 (5%). At the chosen threshold:

- True positives: 350
- False negatives: 150
- False positives: 1,400
- True negatives: 8,100

Sensitivity = 350/500 = **70%** Specificity = 8,100/9,500 = **85%** Precision (PPV) = 350/1,750 = **20%** NPV = 8,100/8,250 = **98%** Accuracy = 8,450/10,000 = **84.5%**

Now translate. The system flags 1,750 patients. Four out of five of them do not have sepsis. It misses 150 patients who do. If review takes fifteen minutes, this generates about 440 hours of clinical work per ten thousand admissions.

Whether that is a good system depends on whether the 350 caught cases are caught *earlier than they would have been otherwise*, and whether the service can absorb 440 hours. Neither question is answerable from the metrics alone. This is the point.

Calibration

Discrimination asks whether the model ranks patients correctly. Calibration asks whether the numbers mean what they say — of patients assigned 20% risk, do approximately 20% experience the outcome?

A model can discriminate well and be badly calibrated. If it reports 60% when the true rate is 25%, ranking is preserved but every absolute number is wrong. Since clinicians and patients reason with absolute risk — "one in five" carries meaning that "in the 87th percentile" does not — calibration is arguably the more important property for shared decision-making. Van Calster and colleagues memorably called it the Achilles heel of predictive analytics, and the description holds.

Assess calibration visually with a calibration plot: predicted probability on the x-axis, observed frequency on the y-axis, ideally on the diagonal. Report calibration slope (1.0 is ideal; below 1.0 indicates overfitting) and calibration-in-the-large (mean predicted versus observed). Do not rely on the Hosmer-Lemeshow test alone; it is underpowered in small samples and rejects trivially in large ones.

Calibration is also the property most likely to degrade after deployment, because outcome prevalence shifts. Chapter 18 covers monitoring it.

Clinical utility

The final layer asks whether using the model produces better decisions than not using it.

Decision curve analysis compares the net benefit of the model against the strategies of treating everyone and treating nobody, across a range of threshold probabilities. The threshold at which a clinician would act encodes their implicit judgement about the relative harm of the two error types, and decision curves show whether the model adds value at those thresholds. A model can have excellent AUC and no net benefit at any plausible threshold.

Number needed to evaluate — how many flagged patients must be reviewed to find one true case — is a blunt but effective way to communicate burden to a clinical audience.

Reporting standards for evaluation

- Report discrimination, calibration, and clinical utility. All three.
- Give confidence intervals on everything.
- State absolute counts at the operating threshold, not only rates.
- Break performance down by relevant subgroups.
- Name the comparator explicitly.
- Report the outcome prevalence in the evaluation set, so PPV can be interpreted.

IMPORTANT CONSIDERATION

When a vendor presents a single AUC figure, the useful follow-up questions are: on what population, at what prevalence, with what operating threshold, generating how many alerts per thousand patients, at what positive predictive value, and validated where? A vendor unwilling or unable to answer those has not done the work, and the number in the brochure should be treated accordingly.

KEY TAKEAWAY

Accuracy is misleading for rare outcomes and should not headline a clinical evaluation. Sensitivity and specificity belong to the model; PPV and NPV depend on prevalence and will change when you move the model. Report ROC and precision-recall together, plot calibration rather than testing it, and always translate performance into absolute counts of patients flagged, caught, and missed.

CHAPTER 9

Clinical Validation and Real-World Performance

Why models that work in the lab fail on the ward.

There is a specific, well-documented pattern in healthcare AI: strong development performance, weaker external performance, weaker still prospective performance. It is consistent enough that any team reporting internal results should assume real-world performance will be worse and plan accordingly.

This chapter explains the levels of validation, why the degradation occurs, and what to do about it.

The validation hierarchy

Technical validation confirms the software works — inputs parsed, outputs produced, latency within specification, failures handled. Necessary, and it says nothing about clinical performance.

Internal validation estimates performance on data from the same source as development, using cross-validation, bootstrapping, or a random holdout. It corrects for optimism from model fitting. It cannot detect anything about transportability.

Temporal validation tests on a later time period from the same institution. This is the weakest form of external validation and the most under-used. It catches drift in coding practice, treatment patterns, and case mix — and it is nearly free, because the data already exists.

External validation tests on data from a different institution, region, or health system, with no involvement in development. This is where transportability is established, and where performance usually drops.

Prospective validation runs the model on live data going forward, generating real predictions on real patients, with outcomes ascertained afterwards. Predictions are made without knowledge of the future, and

every operational problem — missing feeds, latency, data arriving in unexpected formats — becomes visible.

Clinical validation evaluates whether *using* the model changes clinical outcomes. This requires a comparative design: randomised where feasible, stepped-wedge, or a well-controlled before-and-after. It is the only level that establishes clinical benefit, and it is the level at which the evidence base for clinical AI is thinnest.

Real-world evaluation is the continuing assessment after deployment: does performance hold, does the workflow persist, are clinicians still acting on it, has anyone been harmed.

> **IMPORTANT CONSIDERATION**
>
> These levels are not interchangeable. A model with excellent internal validation and no external validation has demonstrated that it fits its development data well. That is a weaker claim than it sounds, and it is not evidence that the model will work in your hospital. Ask specifically which levels have been completed, and by whom.

The instructive failure

The clearest published example of the gap between reported and real-world performance concerns a widely implemented proprietary sepsis prediction model, evaluated externally by Wong and colleagues at the University of Michigan and published in JAMA *Internal Medicine* in 2021.

Across 27,697 patients and 38,455 hospitalisations, the model achieved an area under the curve of **0.63** (95% CI 0.62–0.64) at the hospitalisation level. At the manufacturer's recommended threshold it identified **33%** of sepsis cases — missing 67% — with a positive predictive value of **12%**. It generated alerts on **18%** of all hospitalisations. The authors estimated the model identified only 183 of 2,552 sepsis patients who had not already received timely antibiotics, implying limited added value over existing clinical practice.

The model was, at the time, deployed at hundreds of hospitals.

Several lessons compound here. Proprietary models may be deployed at scale without independent external validation. Vendor-reported performance may not transfer. Local evaluation is not a nicety. And the alert burden of a low-PPV model deployed across a whole hospital is a patient safety issue in its own right, because it consumes the attention that other alerts need.

Why performance degrades

Dataset shift is the umbrella term, and it takes three forms:

- **Covariate shift.** The distribution of inputs changes; the relationship between inputs and outcome does not. A model trained on a younger population applied to an older one.
- **Label shift.** The outcome prevalence changes. Deploying an ICU model on a general ward. Calibration breaks first.
- **Concept shift.** The relationship between inputs and outcome changes. A new treatment alters the natural history; a diagnostic threshold is redefined; a coding standard is updated.

Population differences between sites are larger than people expect: case mix, referral patterns, socioeconomic composition, comorbidity burden, and access.

Practice differences are subtler and often more damaging. How often are vital signs recorded overnight. What is the threshold for ordering a lactate. Which of several equivalent codes does this organisation's coding team prefer. A model that learned "frequent measurement indicates concern" will misfire in a hospital that measures everyone hourly by protocol.

Feedback loops. A successful model changes the population it predicts on. This is not a flaw to be engineered away; it is a permanent feature that must be built into the monitoring plan.

Model drift. Even without any change to the model, its performance decays as the world moves. Every deployed clinical model has a shelf life.

What to do about it

Validate locally before deployment. Every time. Even for a regulator-cleared product, even for a model developed at a similar institution. Run it silently on your own historical data and measure discrimination, calibration, subgroup performance, and alert volume in your population.

Run a silent period. Before any output reaches a clinician, run the model live for weeks to months, generating and storing predictions without displaying them. This measures operational reliability and real-world performance simultaneously, at zero clinical risk. It is the highest-value and most frequently skipped step in the whole process.

Recalibrate rather than rebuild. Transporting a model to a new setting often requires only updating the intercept, or the intercept and slope, to match local prevalence. This is far cheaper than redevelopment and often sufficient.

Plan for periodic revalidation. Set an interval — annual is a common default — and define the performance thresholds that trigger earlier review.

Reproducibility

A model that cannot be reproduced cannot be validated, revalidated, or investigated after an incident. The minimum requirements: version-controlled code for extraction, features, training and inference; recorded data versions and extraction dates; fixed random seeds; a pinned environment specification; and stored model artefacts with their evaluation results.

This is not bureaucratic overhead. When a monitoring alert fires in eighteen months and someone asks whether the model changed or the world changed, this is the only material that can answer the question.

KEY TAKEAWAY

Validation is a hierarchy, not a binary, and each level answers a different question. External and prospective validation routinely reveal performance well below development figures — the Epic sepsis model's fall to an AUC of 0.63 with 12% PPV in external validation is the canonical illustration. Validate locally before deployment without exception, run a silent period before showing anything to a clinician, and treat recalibration as the standard tool for transporting a model between settings.

PART IV

Clinical Decision Support

Getting a prediction in front of a clinician in a form they can act on, without drowning them.

CHAPTER 10

What Is Clinical Decision Support?

The categories, the delivery modes, and where AI fits.

Clinical decision support is any system that presents a clinician with information intended to improve a specific decision at a specific moment. It is a category defined by function, not by technology. A laminated card on a resuscitation trolley is decision support. So is a drug interaction pop-up, so is a deep learning model scoring a chest radiograph.

The distinction matters because governance attaches to function. A system that influences a clinical decision carries obligations regardless of whether a neural network or a lookup table sits behind it.

What AI-driven CDS actually does

Risk alerts. A score crosses a threshold and someone is notified. The dominant pattern for deterioration, sepsis, and acute kidney injury. Highest potential impact and highest risk of alert fatigue.

Patient prioritisation. A ranked list, reviewed at a scheduled time rather than pushed as an interruption. Care management outreach, pharmacist review, screening outreach. Far kinder to workflow than alerting, and consistently underrated. Where the decision is not time-critical, this should usually be the default design.

Diagnostic support. Highlighting a finding on an image, suggesting differentials, flagging a result pattern consistent with an under-recognised condition. Mature in imaging, less so elsewhere.

Treatment support. Dosing calculations, guideline-concordant options, contraindication checking, therapy selection informed by molecular profiling.

Monitoring and surveillance. Continuous evaluation of a population against criteria — patients overdue for a test, patients on a drug requir-

ing monitoring that has not occurred, deteriorating trends in a chronic disease register.

Documentation and administrative support. Ambient scribing, draft letters, coding assistance, summarisation of long records. Not clinical decision support in the strict sense, but the area where generative AI has moved fastest into routine use, and where the workload benefit is currently most demonstrable.

Delivery modes

How information arrives determines whether it is used, and it is a more consequential design decision than most teams treat it as.

Mode	Description	Best for	Main risk
Interruptive alert	Modal pop-up requiring a response	Time-critical, high-stakes, high-PPV situations	Alert fatigue; workflow disruption
Passive display	Score shown in a chart section or column	Contextual awareness	Ignored entirely
Worklist / ranked queue	Prioritised list reviewed on a schedule	Non-urgent prioritisation	Never opened; needs an owner
Order set integration	Suggested orders appear where clinicians already work	Standardising response	Automation bias
Asynchronous notification	Message to an inbox or team channel	Population outreach	Delay; unclear ownership
Ambient / background	Operates without explicit clinician interaction	Documentation, coding	Invisible errors

A single principle governs the choice: **interrupt only when the situation is urgent and the prediction is likely to be right.** A model with

12% PPV should almost never generate a modal pop-up. It might reasonably generate a ranked list that a nurse reviews twice a day.

The five rights

A framework from the informatics literature that predates modern AI and remains the most useful checklist available. Effective decision support delivers:

1. The **right information** — evidence-based, specific, actionable
2. To the **right person** — whoever can actually act
3. In the **right format** — matched to the decision
4. Through the **right channel** — the system they already use
5. At the **right time** — when the decision is being made

Most failed AI deployments fail on rights 2, 4, or 5 rather than on right 1. The prediction was correct; it reached a clinician who could not act on it, in a system they had already closed, four hours after the decision was made.

Prediction and decision are different objects

The relationship between what a model produces and what a clinician does is worth stating explicitly:

Model output → *"18% probability of deterioration in 24 hours"* **Clinical interpretation** → *"Higher than I would have guessed; the trend on the observation chart supports it"* **Contextual integration** → *"But she is on a treatment escalation plan already, and the family has agreed a ceiling of care"* **Decision** → *"Increase observation frequency, no ICU referral"*

The model contributed one input to a decision that also involved things it could not see. This is the normal and desirable case. A system designed as though the prediction were the decision will be wrong in exactly the situations where context matters most — which are the situations that matter most.

IMPORTANT CONSIDERATION

Regulatory frameworks in several jurisdictions distinguish decision support that a clinician can independently review from support that they cannot. Where a clinician can see the basis of a recommendation and reach their own conclusion, the regulatory burden is typically lower. Where the system's reasoning is opaque, or where the recommendation is time-critical enough that independent review is impractical, the classification is usually stricter. This is one reason explainability has regulatory as well as ethical weight. Confirm the current position in your jurisdiction with qualified advisers.

Transparency obligations

Regulation has begun to require disclosure of what sits behind a prediction. In the United States, the HTI-1 final rule, published by the Office of the National Coordinator for Health Information Technology in January 2024, established requirements for certified health IT relating to "predictive decision support interventions." Certified developers must make specified source attributes available to users — covering the intervention's purpose, development, validation, and maintenance — so that clinicians and organisations can assess a tool before relying on it.

The direction of travel is clear and it is towards disclosure. Organisations building models internally should assume that the documentation expectations applied to vendors will increasingly be applied to them.

KEY TAKEAWAY

Clinical decision support is defined by function, and AI is one of several possible engines. Delivery mode is as consequential as model quality: interrupt only when the situation is urgent and the prediction is likely to be correct, and default to ranked worklists otherwise. The five rights framework catches most design failures. A prediction is one input to a decision, never the decision itself.

CHAPTER 11

Designing Human-Centered AI

Alert fatigue, automation bias, and the workflow realities that decide whether anyone uses your model.

If there is one number in this book worth memorising, it is this one. A systematic review and meta-analysis of eleven studies covering 570,776 prescriptions found that clinicians overrode approximately **90%** of drug–drug interaction alerts (95% CI 85–95%).

Those alerts are rule-based, well-established, and frequently correct. They are ignored nine times out of ten because there are too many of them, most are not relevant to the patient in front of the clinician, and the cost of dismissing one is a single click while the cost of reading one is thirty seconds the clinician does not have.

Any AI system that generates alerts is entering an environment already saturated with them. It is competing for a resource — clinical attention — that is fully consumed. This is the design constraint that dominates everything else.

Alert fatigue

Alert fatigue is the progressive desensitisation that follows repeated exposure to notifications, most of which do not require action. Its consequences are direct: clinicians dismiss alerts without reading them, including the ones that matter, and important signals are lost inside the noise.

The arithmetic is brutal for low-PPV models. A model flagging 18% of admissions at 12% PPV, in a 500-bed hospital with 60 admissions a day, produces roughly 11 alerts daily, of which perhaps 1 identifies a patient who needed the intervention. Within a fortnight the ward has learned what that alert means, and it does not mean "stop what you are doing."

Practical mitigations, in rough order of effectiveness:

Raise the threshold. Fewer, more precise alerts beat more, less precise ones, even at the cost of sensitivity. This is counter-intuitive to modellers and obvious to clinicians.

Change the channel. Move non-urgent output from interruptive alerts to a reviewed worklist. Most of the value survives; nearly all of the fatigue disappears.

Target the right recipient. An alert to a named person who can act beats a broadcast to everyone who might.

Suppress repeats. Do not re-alert on the same patient for the same reason within a defined window unless the situation has materially changed.

Give a reason and an action. "Sepsis risk elevated" prompts dismissal. "Sepsis risk 34%: lactate 3.2 rising, HR 118, temp 38.9, no antibiotics in 6h — consider sepsis screen" prompts a decision.

Measure and act on the response. Override rate, time to action, and clinician-reported usefulness are performance metrics. If override rate exceeds 90%, the system is not working, whatever its AUC says.

Automation bias

The opposite failure. Automation bias is the tendency to over-trust automated output — accepting a recommendation that is wrong, or failing to notice something the system did not flag.

It takes two forms. **Commission errors**: acting on an incorrect recommendation. **Omission errors**: failing to detect a problem because the system did not raise it. The second is more insidious, because a high negative predictive value can be quietly reassuring in a way that suppresses independent judgement.

Design responses:

- Display confidence or uncertainty, not a bare number
- Show the contributing factors, so the clinician can sanity-check the reasoning

- State the known limitations at the point of use — populations where performance is weaker, inputs that were missing
- Never imply that absence of an alert means absence of risk
- Preserve a genuine, low-friction path to disagree, and record when it is used

That last point is worth expanding. A system where disagreeing requires more effort than complying will produce compliance, and the compliance will be uninformative. Conversely, disagreement data is one of the most valuable monitoring signals available — a rising override rate is often the first indication that something has shifted.

Human-in-the-loop

Three configurations, and the choice is a governance decision as much as a technical one.

Human-in-the-loop. A person reviews and approves each output before it takes effect. Appropriate for consequential, irreversible decisions. Highest safety, highest cost, does not scale.

Human-on-the-loop. The system operates continuously; a person supervises and can intervene. Appropriate for monitoring and triage at volume.

Human-out-of-the-loop. Fully automated action. Appropriate only for low-stakes, high-volume, reversible tasks with strong evidence — routine flagging of overdue tests, for example. Rare and should stay rare in clinical settings.

The decisive question is not how accurate the model is but how bad the worst plausible error is and how quickly it would be caught.

Workflow integration

A model that requires a clinician to leave their workflow will not be used. This is not a matter of clinician attitude; it is a matter of arithmetic. A separate login costs perhaps ninety seconds. In a clinic running seven-minute appointments, that is not available.

Requirements that are non-negotiable in practice:

- Output appears inside the existing EHR interface
- No separate authentication
- Latency low enough to be invisible — sub-second where the clinician is waiting
- Automatic patient context; nothing re-entered by hand
- Graceful degradation when data is missing, with the degradation visible
- Documented behaviour when the model is unavailable

Trust

Clinician trust is earned, easily lost, and not simply a function of accuracy. The factors that build it:

Transparency about performance. Publish local validation results to the users. Clinicians are trained in diagnostic test characteristics and are entirely capable of interpreting sensitivity and PPV. Treating them as unable to handle the numbers is both wrong and corrosive.

Involvement in design. Clinicians who helped specify the tool defend it. Those presented with a finished product resist it, often correctly.

Honesty about limitations. A system that acknowledges what it cannot do is trusted more than one that presents uniform confidence.

Responsiveness to feedback. A visible mechanism for reporting problems, and evidence that reports lead to change. Nothing destroys credibility faster than a feedback button that goes nowhere.

Consistency. Unexplained behaviour changes after a silent model update will end adoption permanently.

How poor design increases workload

Concrete patterns, all observed repeatedly in practice:

- **Alerts without actions.** "This patient is high risk" with no recommended step. The clinician now has an obligation and no guidance.
- **Duplicated documentation.** The model's assessment must be recorded separately from the clinical note.
- **Unbounded review lists.** A queue with no capacity limit, growing faster than it can be worked, generating guilt rather than care.
- **Alerts to the wrong role.** Firing to a consultant who is in theatre rather than the ward nurse who is at the bedside.
- **No suppression.** Re-alerting on the same patient every four hours for a condition already assessed and documented.
- **Mandatory justification fields.** Requiring free-text explanation for every dismissal, which produces "n/a" typed 200 times a day and a dataset of nothing.

IMPORTANT CONSIDERATION

Measure clinician workload before and after deployment, and treat an increase as a defect rather than a cost of progress. Useful measures: time in the EHR per patient, number of alerts received per shift, override rate, and a short usability instrument administered at intervals. A model that improves outcomes while adding twenty minutes a day to every clinician's workload will be switched off within a year, and it will deserve to be.

KEY TAKEAWAY

Clinical attention is the binding constraint, and it is already fully consumed — roughly 90% of drug interaction alerts are overridden. Design for that reality: raise thresholds, prefer worklists to interruptions, target named recipients, and always pair a flag with a reason and a suggested action. Guard equally against automation bias by showing uncertainty and contributing factors, and by keeping disagreement easy and recorded.

CHAPTER 12

Explainable and Interpretable Healthcare AI

What clinicians actually need to see, and what the popular methods do and do not deliver.

Two statements about the same patient:

> *"The model predicts high risk."*
>
> *"The model predicts high risk. The largest contributors are an eGFR decline of 9 mL/min over 14 months, a urine albumin-creatinine ratio of 84 mg/mmol, and systolic blood pressure averaging 156 over the past year. No ACE inhibitor or ARB is currently prescribed."*

The first is an instruction. The second is an argument, and a clinician can engage with an argument — agree, disagree, notice that the eGFR decline followed a documented episode of dehydration, or observe that the ACE inhibitor was stopped for hyperkalaemia and the model does not know.

That is the entire case for explainability, and it is a strong one.

Why it matters here specifically

Clinical reasoning requires it. Medicine works by constructing and testing explanations. A number with no rationale cannot enter that process; it can only be obeyed or ignored.

Error detection depends on it. Explanations reveal when a model is right for the wrong reason. Published cases include models that keyed on scanner metadata rather than pathology, and on the presence of a chest drain rather than the pneumothorax the drain was treating. Only inspection of what the model attended to catches these.

Accountability requires it. A clinician who acts on a recommendation is answerable for the outcome. Answerability without visibility is an unreasonable position to place someone in.

Regulation increasingly requires it. Transparency obligations under the HTI-1 rule in the United States and under the EU AI Act for high-risk systems both move in this direction.

Patients are entitled to it. A patient told they are high risk may reasonably ask why. "The computer says so" is not an acceptable answer.

Interpretable versus explained

An important distinction that is often blurred.

An **interpretable model** is one whose mechanism is directly inspectable — a logistic regression's coefficients, a short decision tree's branches, a points-based score. The explanation *is* the model.

An **explained model** is a complex model with a separate method applied afterwards to approximate why it produced a given output. The explanation is a second model of the first, and it can be wrong.

This is a real trade-off and it is frequently resolved too quickly in favour of complexity. Where an interpretable model performs comparably — which in tabular clinical prediction is often — it is usually the better clinical choice, because there is nothing to approximate.

The methods

Feature importance (global). Ranks which variables matter most across the whole model. Useful for sanity-checking that the model is using clinically plausible signals. Says nothing about an individual patient.

SHAP (SHapley Additive exPlanations). Assigns each feature a contribution to the difference between this patient's prediction and the average prediction, using a game-theoretic allocation with desirable mathematical properties. It is the most widely used method for individual explanation of tabular clinical models, and it produces the additive, per-patient attributions that clinician-facing displays need.

Limitations that matter: computation can be expensive; the standard implementations assume feature independence, which is badly violated by correlated clinical variables (creatinine and eGFR are not independent); and it explains the model's behaviour, not the biology. A large SHAP value is not a causal claim and must never be presented as one.

LIME (Local Interpretable Model-agnostic Explanations). Fits a simple local model around the individual prediction using perturbed samples. Intuitive and model-agnostic. Less stable than SHAP — repeat runs can give different explanations — which is a serious drawback in a clinical setting where the same patient should not produce different reasons on different days.

Counterfactual explanations. "If systolic blood pressure were below 130, predicted risk would fall from 22% to 14%." Often the most clinically useful form, because it maps directly onto action. Requires care: counterfactuals must be constructed only over modifiable features. A counterfactual on age is absurd, and one on ethnicity is offensive.

Attention and saliency maps. For imaging and text, highlighting the regions that drove the output. Intuitive and known to be unreliable as faithful explanations — several published analyses have shown saliency maps that appear plausible while being insensitive to model parameters. Use them as a prompt for human review, not as evidence of reasoning.

Example-based explanation. "This patient's trajectory resembles these previous cases." Aligns with how clinicians actually reason and is under-used.

What to put on the screen

Most explanation output is designed for data scientists. A SHAP force plot is not a clinical artefact.

A workable clinician-facing explanation contains:

1. **The prediction, with its meaning made explicit.** "22% estimated probability of progression within 24 months" — not "risk score 0.22," and not "high risk" alone.

2. **Three to five contributing factors, in clinical language**, each with direction and magnitude. Not fourteen. Not raw feature names.
3. **What can be modified.** Separating actionable from fixed contributors turns an assessment into a plan.
4. **What the model did not see.** Missing inputs, unavailable data, exclusions.
5. **Population and limitations.** "Validated in adults aged 40–80 in this health system; performance in patients under 40 is not established."
6. **A route to more detail** for those who want it, without forcing it on those who do not.

> **IMPORTANT CONSIDERATION**
>
> A well-designed explanation can increase inappropriate trust as easily as appropriate trust. A fluent, confident-sounding rationale makes a wrong prediction more persuasive, not less. This is a real and documented risk with generative explanation layers in particular, where a language model can produce a compelling narrative for a prediction that is simply incorrect. Explanations should be tested empirically for their effect on clinician decisions, not assumed to be beneficial because they are informative.

Model documentation

Alongside per-prediction explanation, every deployed model needs a standing document describing what it is. The **Model Facts label**, proposed by Sendak and colleagues in *npj Digital Medicine* in 2020, adapted the format of a nutrition label for this purpose and remains a good template. The Coalition for Health AI has since developed a similar "model card" approach, with a public registry, aimed at giving health systems a consistent way to evaluate AI tools.

Whatever format is used, it should state: intended use and users; the population it was developed on; the outcome definition; inputs required; performance including calibration and subgroup results; validation history; known limitations and out-of-scope uses; date of last validation; and a named owner.

This document belongs where clinicians can reach it in one click from the tool itself, not in a governance folder on a shared drive.

KEY TAKEAWAY

Explanation converts an instruction into an argument a clinician can evaluate, and that is what makes prediction usable. Prefer interpretable models where performance is comparable. Use SHAP for individual attribution and counterfactuals for actionability, while remembering that both explain the model rather than the biology. Design the clinician-facing view for clinicians, and test whether explanations actually improve decisions rather than assuming they do.

PART V

Responsible Healthcare AI

Equity, privacy, and safety — the obligations that come with putting a model in front of a patient.

CHAPTER 13

Bias, Fairness, and Health Equity

How models inherit and amplify existing disparities, and what to do about it.

In 2019, Obermeyer, Powers, Vogeli and Mullainathan published an analysis in *Science* of a commercial algorithm used by health systems in the United States to identify patients for enrolment in high-risk care management programmes. The algorithm was applied to roughly 200 million people annually.

It was not designed to consider race. Race was not an input.

At any given algorithm score, Black patients were substantially sicker than White patients with the same score. The authors estimated that correcting the bias would raise the proportion of Black patients receiving additional help from **17.7% to 46.5%**.

The cause was a single design decision. The algorithm predicted *healthcare costs* as a proxy for *healthcare needs*. Because less money is spent on Black patients at equivalent levels of illness — a consequence of unequal access, not of lower need — the algorithm learned to rank them as healthier.

This is the most important case study in healthcare AI, and every element of it is instructive. The team was competent. The proxy was reasonable-sounding and widely used. The bias was undetectable from overall accuracy metrics. It was found only because someone stratified performance by race and looked.

Where bias comes from

Historical bias. The data faithfully records a health system in which access, treatment, and outcomes have differed systematically between groups. A model trained on it reproduces those patterns, and reproduces them at scale and with the authority of a number.

Representation bias. A group is under-represented in the training data, so the model has less information about it and performs worse. This is the mechanism behind well-documented performance disparities in dermatology models across skin tones, and in pulse oximetry, where reduced accuracy in patients with darker skin has been shown to lead to unrecognised hypoxaemia.

Measurement bias. A variable means different things in different groups, or is measured with different accuracy. Pain documentation is a well-studied example, with substantial literature on differential assessment and treatment.

Label bias. The outcome definition itself is biased — the Obermeyer case exactly. Diagnosis-based labels carry the same problem wherever diagnosis rates differ by access rather than by prevalence.

Aggregation bias. A single model is fitted across groups for whom the relationship between predictors and outcome genuinely differs, so it fits the majority well and everyone else badly.

Deployment bias. A model performing equally well across groups is deployed in a way that does not reach everyone equally — a patient portal intervention for a population with uneven digital access, for instance.

The variable-inclusion question

Should race or ethnicity be an input?

The debate has moved substantially in recent years, and towards a clearer answer. Race is a social construct, not a biological variable. Including it as a predictor typically encodes the effects of racism — differential access, differential treatment, chronic stress exposure — as though they were properties of the patient. Several prominent clinical algorithms have been revised to remove race coefficients, including the equations for estimated glomerular filtration rate, where a race adjustment had systematically classified Black patients as having better kidney function and delayed their referral for transplantation.

The American Heart Association's PREVENT equations, published in 2024, were developed without race as an input, using social deprivation index instead where area-level data is available.

The current mainstream position, expressed as a working rule:

- **Do not include race as a proxy for biology.** There is almost never a biological mechanism, and the variable is doing other work.
- **Do consider including measured social determinants** — deprivation, insurance status, housing stability — which are the actual causal factors and are potentially actionable.
- **Always evaluate performance by race and ethnicity**, whether or not they are model inputs. Excluding a variable does not prevent bias; the Obermeyer algorithm did not use race.

That last point is the one that gets missed. Removing a variable removes a lever, not a problem.

Measuring fairness

Fairness is not one quantity. Several reasonable definitions exist and they are mathematically incompatible when base rates differ between groups — which they usually do.

Definition	Requirement	Sensible use
Demographic parity	Equal flag rates across groups	Rarely appropriate clinically; ignores genuine prevalence differences
Equal opportunity	Equal sensitivity across groups	Strong default when missing a case is the primary harm
Equalised odds	Equal sensitivity *and* specificity	Stricter; appropriate when both error types carry real cost
Predictive parity	Equal PPV across groups	Relevant when the flag triggers a scarce, burdensome intervention
Calibration within groups	Predicted probabilities accurate in each group	Essential when absolute risk drives the decision

Because they cannot all hold simultaneously, the choice is a value judgement about which harm matters most in this specific application. Make it explicitly, document it, and involve clinicians and patient representatives in making it. A team that has not chosen has chosen by default.

A practical fairness evaluation

- ☐ Relevant subgroups defined in advance, with clinical and community input
- ☐ Sample sizes within each subgroup checked for adequacy
- ☐ Discrimination reported per subgroup with confidence intervals
- ☐ **Calibration reported per subgroup** — the most frequently omitted and most consequential analysis
- ☐ Sensitivity and PPV at the operating threshold reported per subgroup
- ☐ Missing data rates compared across subgroups
- ☐ Outcome definition examined for differential ascertainment
- ☐ Intersectional subgroups examined where sample size permits
- ☐ Fairness criterion chosen and justified in writing
- ☐ Mitigation attempted where disparities are found, and residual disparity disclosed
- ☐ Subgroup performance monitored continuously after deployment

IMPORTANT CONSIDERATION

Subgroup analysis requires demographic data, which many organisations collect incompletely and some are reluctant to use. The reluctance is understandable and the consequence is worse: without the data, disparities are undetectable and persist indefinitely. Collecting and using demographic data for fairness auditing is a different activity from using it as a model input, and the two should not be conflated in governance discussions.

Mitigation

Pre-processing. Fix the data. Improve representation, correct the outcome definition, address differential missingness. The Obermeyer team's own fix — switching the prediction target from cost to a direct measure of health — reduced the disparity by around 84%. Changing the label is usually more effective than any algorithmic correction.

In-processing. Add fairness constraints to the training objective, or use adversarial approaches. Technically appealing, generally involves an accuracy trade-off, and can be brittle.

Post-processing. Adjust thresholds per group, or recalibrate within groups. Effective and operationally simple; raises legal and ethical questions about differential treatment that must be worked through with appropriate advice.

Deployment mitigation. Ensure the intervention is accessible to everyone flagged. A model that fairly identifies need, feeding a service that only some patients can reach, produces an unfair outcome regardless of the model's properties.

Equity as a design goal

The most useful reframing available: rather than asking "is this model unbiased," ask "does deploying this system reduce or widen existing disparities in this outcome?"

That question forces attention onto the whole pathway — identification, access, intervention, follow-up — rather than onto the model alone. It also admits the possibility that a technically imperfect model, deployed in a way that reaches under-served patients, improves equity, while a technically excellent one deployed through a channel only the well-resourced use makes things worse.

KEY TAKEAWAY

Bias enters through the outcome definition more often than through the algorithm; a widely used commercial risk algorithm systematically under-identified Black patients because it predicted cost rather than need. Removing race as an input does not prevent bias and is not a substitute for stratified evaluation. Fairness definitions are mutually incompatible — choose one explicitly and justify it. Report calibration by subgroup, not only discrimination, and judge the system by whether it narrows or widens disparities in practice.

CHAPTER 14

Privacy, Security, and Responsible Data Use

Protecting patients while making data useful.

Healthcare data is among the most sensitive information that exists about a person. A medical record can disclose infectious disease status, mental health history, reproductive decisions, substance use, and genetic risk — categories with direct consequences for employment, insurance, relationships, immigration status, and personal safety in some contexts.

It is also, uniquely, data that cannot be reissued. A compromised password is replaced in a minute. A disclosed HIV diagnosis is permanent.

This chapter sets out the practical obligations. It is not legal advice, and requirements differ materially by jurisdiction; organisations should obtain qualified advice for their territory.

The regulatory landscape

United States. The Health Insurance Portability and Accountability Act (HIPAA) governs protected health information held by covered entities and their business associates. The Privacy Rule controls use and disclosure; the Security Rule mandates administrative, physical and technical safeguards; the Breach Notification Rule sets disclosure obligations. Research uses typically require authorisation, an institutional review board waiver, or a de-identification pathway.

European Union and United Kingdom. The General Data Protection Regulation and UK GDPR treat health data as a special category requiring a specific lawful basis. Core principles — purpose limitation, data minimisation, storage limitation, accuracy — bear directly on model development, as do rights of access, rectification, erasure, and the provisions on automated decision-making.

EU AI Act. Regulation (EU) 2024/1689 entered into force on 1 August 2024, with obligations phasing in over subsequent years. AI systems that are, or are safety components of, medical devices are generally classified as high risk, attracting requirements on risk management, data governance, technical documentation, logging, transparency, human oversight, accuracy, robustness and cybersecurity.

Elsewhere. Canada's PIPEDA and provincial health privacy statutes, Australia's Privacy Act, Brazil's LGPD, and a growing number of national frameworks. Cross-border data transfer restrictions are a recurring practical obstacle to multi-site collaboration.

Assume this landscape will change during the life of any system you build.

De-identification and its limits

De-identification reduces but does not eliminate re-identification risk, and the literature on re-identification from ostensibly anonymous datasets is extensive and sobering.

Removing direct identifiers — name, medical record number, address, contact details — is necessary and grossly insufficient on its own.

Quasi-identifiers in combination are the real exposure. Date of birth, sex and postcode together identify a large fraction of a population. Rare diagnoses, unusual laboratory values, and distinctive care trajectories are identifying in themselves.

Statistical approaches include k-anonymity (each record indistinguishable from at least k–1 others on quasi-identifiers), l-diversity and t-closeness, which address k-anonymity's known weaknesses. All trade information against protection.

Differential privacy provides a mathematically rigorous guarantee by adding calibrated noise, with a tunable privacy budget. It is increasingly practical and it does cost accuracy — a trade-off requiring explicit decision rather than default.

Synthetic data generates artificial records preserving statistical structure. Useful for development, testing, and teaching. Not a general

substitute for real data in validation, and the privacy guarantees of synthetic generators are weaker than commonly assumed unless generated under a formal privacy mechanism.

> **IMPORTANT CONSIDERATION**
>
> Models themselves can leak training data. Membership inference attacks can determine whether a specific individual was in a training set; model inversion can reconstruct approximations of training records; large language models have been shown to reproduce verbatim text from training corpora. Treat the trained model as a data asset requiring the same access controls as the underlying records — not as a harmless derived artefact.

Governance

Data minimisation. Collect and retain only what is needed for the stated purpose. This is a legal requirement in several jurisdictions and a security benefit everywhere: data you do not hold cannot be breached.

Purpose limitation. Data collected for care carries different expectations than data collected for research. Reuse requires a lawful basis and should be assessed against what patients would reasonably expect.

Access control. Role-based, least-privilege, with regular review. Development teams rarely need identified data; they need a de-identified extract in a controlled environment.

Audit trails. Log who accessed what, when, and why, and monitor the logs. Unmonitored logs are theatre.

Encryption. At rest and in transit, with managed keys and defined rotation.

Retention and deletion. Defined schedules, actually executed, including backups and derived datasets.

Vendor management. Third parties handling patient data require contractual obligations, security assessment, breach notification terms, and clarity on sub-processors and data location.

Incident response. A tested plan naming responsibilities, containment steps, notification thresholds and timelines.

Federated and privacy-preserving approaches

Multi-site data is what makes models generalisable, and moving patient data between institutions is exactly what privacy regulation restricts. Several approaches resolve this tension.

Federated learning trains a shared model across institutions without centralising records: each site trains locally, only model updates are exchanged, and a coordinator aggregates them. Rieke and colleagues set out the case for this in healthcare in *npj Digital Medicine* in 2020, and it has since been used in several large multi-institution collaborations. It is not automatically private — updates can leak information — and is usually combined with secure aggregation or differential privacy.

Secure multi-party computation allows joint computation over inputs that remain private. Strong guarantees, high computational cost.

Trusted research environments, also called data safe havens, bring the analyst to the data rather than the reverse: researchers work inside a controlled environment and only approved outputs leave. This is now the dominant model for national-scale health data research in several countries and is often the most practical option available.

Patient trust

Legal compliance is the floor, not the ceiling. Public support for health data use is real but conditional, and it is consistently strongest where the purpose is clearly for patient benefit, where governance is transparent, and where commercial arrangements are disclosed. It collapses when data sharing arrangements emerge through journalism rather than through disclosure.

Practical implications: be public about what data is used and for what; make the governance process visible; involve patient representatives in oversight with actual authority; provide meaningful choice where feasible; and disclose commercial relationships proactively.

KEY TAKEAWAY

Health data is uniquely sensitive and uniquely irrevocable once disclosed. De-identification reduces risk without eliminating it, and quasi-identifier combinations are the main exposure. Trained models are data assets that can leak their training set and need corresponding controls. Federated learning and trusted research environments make multi-site work possible without moving records. And public trust, once lost through undisclosed data sharing, is extremely difficult to recover.

CHAPTER 15

Clinical Safety and AI Risk Management

Treating a model as a clinical intervention, because that is what it is.

A deployed clinical AI system changes what happens to patients. That makes it a clinical intervention, and it should be governed like one: risks identified in advance, controls specified, residual risk accepted by a named person, and performance monitored for the whole of its operational life.

Most healthcare organisations already have this machinery for medicines and devices. The task is usually to extend it rather than to invent it.

The risks

False negatives. The patient who deteriorates and was never flagged. The most serious failure, and the least visible — nobody investigates an alert that did not fire. Systems need an explicit mechanism for detecting missed cases, typically by retrospectively reviewing adverse outcomes against what the model predicted.

False positives. Unnecessary investigation, treatment, cost, and anxiety, plus the alert fatigue that erodes response to every other signal. High-volume, low-PPV systems cause harm even when each individual false positive is benign.

Automation bias. Covered in Chapter 11. Over-reliance producing both commission and omission errors.

Confabulation in generative systems. Language models produce fluent, plausible, and sometimes entirely fabricated content — invented laboratory values in a summary, a citation to a paper that does not exist, a medication the patient never took. In a documentation context this can enter the permanent record and be relied on downstream by clinicians who have no way of knowing it was generated. Any generat-

ive output entering a clinical record requires human verification against source data, and the system should make that verification easy rather than nominal.

Data drift and model drift. Slow degradation, invisible without monitoring. A model can be substantially worse than at go-live while appearing to work normally.

Miscalibration. Numbers that no longer mean what they say. Particularly damaging where clinicians have learned to trust the scale.

Deployment outside intended use. A model validated in adults applied to adolescents; a model built for one care setting used in another; a model for one condition repurposed for a related one. Extremely common and usually well-intentioned.

Technical failure. Feeds stopping, stale data being scored as current, silent unit changes, latency spikes. The dangerous version is failure that produces plausible output rather than an obvious error — a model scoring on yesterday's vitals gives a number, not an error message.

Unclear accountability. When nobody has defined who is responsible for the model's outputs, responsibility defaults to the clinician who acted on them, who had no role in building or validating the system. This is both unfair and a governance failure.

Hazard analysis

Before deployment, work through the plausible failure modes systematically. A workable structure adapted from clinical risk management:

Element	Question
Hazard	What could go wrong?
Cause	What would cause it?
Effect	What happens to the patient?
Severity	How bad, on the organisation's standard scale?
Likelihood	How often, realistically?

Element	Question
Existing controls	What already reduces this?
Additional controls	What else is needed?
Residual risk	What remains, and who accepts it?

Illustrative rows for a sepsis early warning system:

Hazard	Cause	Effect	Control
Missed sepsis	Model false negative	Delayed antibiotics, avoidable deterioration	Model supplements rather than replaces existing observation escalation; monitor missed cases against outcomes
Alert fatigue	Excessive alert volume	Genuine alerts ignored	Threshold set for PPV not sensitivity; suppression rules; weekly volume monitoring
Stale data scored	Interface failure	Prediction on outdated values	Timestamp validation; suppress prediction if inputs older than defined limit; display data recency
Use outside scope	Applied to paediatric ward	Unvalidated performance	Enforce cohort rules in code, not policy; display eligible population in interface
Silent degradation	Case mix shift	Systematically wrong scores	Monthly calibration monitoring with defined thresholds and a named owner

Accountability

The question "who is responsible when the model is wrong?" needs a written answer before go-live, not after an incident. The answer is normally distributed rather than singular:

- **The clinician** is responsible for the clinical decision, which the model informed.
- **The organisation** is responsible for deciding to deploy, for validating locally, and for monitoring.
- **The developer or vendor** is responsible for the model performing as documented and for disclosing limitations.
- **The clinical safety officer** — or equivalent named role — is responsible for the safety case.

What is not acceptable is a system where a clinician carries full responsibility for acting on a tool they could not inspect, was not consulted about, and cannot switch off.

Reporting and learning

Clinical AI incidents should flow through the same reporting route as other patient safety incidents, with additional categories for AI-specific events: incorrect prediction contributing to harm, technical failure affecting output, systematic degradation, and inappropriate use outside intended scope.

Reports need to reach the people who can act — the model owner and the safety officer, not only the general incident inbox — and there must be a feedback loop demonstrating that reports produce change.

Healthcare AI Safety Checklist

Before deployment

- ☐ Intended use, population and clinical context documented and constrained in code
- ☐ Hazard analysis completed and signed off

- ☐ Named clinical safety officer with defined authority
- ☐ Local validation completed on the deployment population
- ☐ Subgroup performance assessed
- ☐ Operating threshold justified by clinical consequence and alert capacity
- ☐ Failure behaviour defined for missing, stale, and malformed input
- ☐ Manual override available, low-friction, and logged
- ☐ Fallback process defined for model unavailability
- ☐ Clinical users trained on capabilities *and* limitations
- ☐ Documentation accessible from the point of use
- ☐ Monitoring plan with thresholds, frequencies and named owners
- ☐ Incident reporting route established and tested
- ☐ Rollback procedure documented and rehearsed
- ☐ Retirement criteria defined

After deployment

- ☐ Discrimination and calibration monitored on the defined schedule
- ☐ Subgroup performance monitored
- ☐ Input distributions monitored for drift
- ☐ Alert volume and override rate tracked
- ☐ Missed cases reviewed against adverse outcomes
- ☐ Clinician feedback collected and acted on
- ☐ Incidents investigated with documented findings
- ☐ Periodic revalidation performed on schedule
- ☐ Every model change subject to review, testing and version control
- ☐ Governance committee receives regular performance reporting

IMPORTANT CONSIDERATION

Define the retirement criteria before go-live. Deployed models accumulate institutional attachment: someone built it, someone funded it, someone presented it at a conference. Without pre-agreed thresholds for withdrawal, an underperforming model tends to persist through inertia while the evidence against it accumulates. Write down, in advance, the performance level at which the system will be switched off, and who has the authority to do it.

KEY TAKEAWAY

A deployed model is a clinical intervention and needs the governance that implies: hazard analysis, named safety ownership, defined failure behaviour, incident reporting, and continuous monitoring. False negatives are the most serious and least visible risk, requiring deliberate retrospective detection. Generative outputs entering the record need verification against source data. And retirement criteria must be agreed before go-live, because they will not be agreed afterwards.

PART VI

Scalable Implementation

Moving from a notebook to a system that runs every day and keeps working.

CHAPTER 16

From Prototype to Clinical Deployment

The journey, and where most projects stop.

There is a well-known observation in health informatics that a large majority of clinical AI projects never reach routine use. The precise figure varies by who is counting and what they count, but the direction is not in dispute, and the reasons are consistent.

Projects die between the notebook and the ward. The transition looks like this:

Research prototype → Retrospective validation → Silent prospective deployment → Pilot with clinical users → Workflow integration → Full deployment → Continuous monitoring

Each stage has a distinct purpose and a distinct exit criterion. Skipping one does not save time; it defers a discovery to a more expensive moment.

The stages

Research prototype. A model trained on historical data, evaluated offline. Exit criterion: performance that would be clinically useful if it held up, and a documented use case.

Retrospective validation. Evaluation on held-out temporal and, where possible, external data. Exit criterion: performance holds, calibration is acceptable, subgroup analysis is clean.

Silent prospective deployment. The model runs live on real data, generating and storing predictions that nobody sees. This is the most valuable stage and the one most often skipped. It answers questions no retrospective analysis can: does the data feed actually deliver what the model needs, at the latency required, in the format expected? What is the real alert volume? Does prospective performance match retrospective?

Run it for long enough to accumulate a meaningful number of outcome events — typically weeks to months. Exit criterion: prospective performance consistent with retrospective, alert volume within the service's capacity, and technical reliability demonstrated.

Pilot. Output becomes visible to a limited group of clinicians in a limited setting. Now the human factors questions become answerable: do people look at it, do they act on it, does it help, does it get in the way. Exit criterion: clinicians report it useful, workload has not increased unacceptably, no safety concerns, and there is at least a signal on the intended process measures.

Workflow integration. Refining based on pilot findings, adding EHR integration, documentation, training, and support. Exit criterion: sign-off from governance and clinical leadership.

Full deployment. Scaled rollout, staged where possible.

Continuous monitoring. Permanent. Chapter 18.

> **IMPORTANT CONSIDERATION**
>
> The silent deployment period is where most integration defects surface, and it costs nothing in clinical risk because no clinician sees the output. Teams under pressure to demonstrate value skip it and then discover, in front of users, that a laboratory feed reports results in different units on weekends, or that the model receives data ninety minutes after the clinical decision has been made. Recovering clinician confidence after that kind of launch is much harder than the delay would have been.

The infrastructure

Data pipelines. Reliable, monitored, versioned movement of data from source systems to the model. Requirements: defined latency, validation at ingestion, handling of late-arriving and corrected data, and alerting on failure. Silent pipeline failure is one of the most dangerous states a deployed model can be in, because it produces plausible output from stale input.

Feature computation. The same code must generate features in training and in production. Divergence between the two — training/serving skew — is a leading cause of unexplained performance loss. A feature store, or at minimum a single shared library, prevents it.

Model serving. Real-time inference via an API for interactive use; batch scoring for population lists. Requirements: latency budget, throughput, health checks, graceful degradation, and version pinning.

Integration. HL7 FHIR is the standard mechanism for reading data from and writing back to modern EHRs. SMART on FHIR allows an application to launch inside the EHR with patient context and authentication carried through. CDS Hooks is the standard for triggering decision support at defined workflow moments — opening a patient record, signing an order — which is precisely the integration pattern that risk prediction needs.

Version control and model registry. Everything versioned: extraction code, feature code, training code, model artefacts, configuration. The registry records lineage — which data, which code, which parameters produced this artefact — with promotion gates between development, staging and production.

Logging. Every prediction stored with its inputs, output, model version, and timestamp. Without this, no post-hoc investigation is possible and no monitoring is meaningful. It is also, in several regulatory frameworks, a requirement rather than a nicety.

Monitoring and alerting. Technical monitoring (uptime, latency, error rates, feed health) and clinical monitoring (performance, calibration, drift, alert volume) are separate systems with separate owners and both are required.

Documentation

At minimum, and kept current:

- Model documentation in a standard format, accessible from the point of use
- Technical architecture and data flow

- Validation report covering all levels performed
- Hazard analysis and safety case
- Standard operating procedures for the clinical workflow
- Monitoring plan with thresholds and owners
- Change control and version history
- Training materials for each user role

Training

Clinical training for an AI tool is different from training for a new form. It must cover:

- What the model does and, importantly, does not do
- What population it was validated in
- Its performance in *this* organisation, in plain terms: of ten flagged patients, roughly how many will have the condition
- What action is expected when it fires
- How to disagree, and that disagreeing is expected and recorded
- Known limitations and situations where it is unreliable
- How to report a problem and what happens to reports

Presenting local performance numbers honestly, including the disappointing ones, does more for adoption than any amount of enthusiasm. Clinicians who understand that a tool has a 20% positive predictive value will use it appropriately. Clinicians told it is "highly accurate" will either over-trust it or, on discovering the truth, stop using it entirely.

> **KEY TAKEAWAY**
>
> The path from prototype to deployment has seven stages with distinct exit criteria, and the silent prospective period is the one that most reliably prevents expensive surprises. Training/serving skew, silent pipeline failure, and integration latency are the technical failures that most often kill otherwise-good models. Train clinicians on honest local performance figures rather than on enthusiasm.

CHAPTER 17

Building a Scalable Healthcare AI Architecture

A reference architecture, component by component.

The architecture below is a conceptual reference. Specific technology choices vary; the layers and their responsibilities do not, and a system missing one of them will have that gap filled informally by someone's spreadsheet.

The loop at the bottom is the part most architecture diagrams omit and the part that makes the system improvable rather than static.

Layer by layer

Data sources. EHR, laboratory systems, imaging archives, pharmacy, monitoring devices, wearables, claims, registries, and external refer-

ence data. Each has its own format, latency, reliability and ownership. Document all four for every source.

Data integration. Ingestion, transformation and consolidation. Streaming interfaces (HL7 v2 messaging, FHIR subscriptions) for real-time needs; batch extraction for population analytics. This layer owns patient identity resolution, terminology mapping, unit harmonisation, and de-duplication. Mapping to a common data model such as OMOP here — rather than downstream — pays for itself the first time a second site is added.

Data quality and governance. Automated validation at ingestion: completeness, range checks, referential integrity, timeliness. Records lineage, enforces access control, and maintains the audit trail. Crucially, it must be able to *stop* bad data reaching the model, not merely note that it arrived.

Feature engineering. Transforms validated data into model inputs. The single most important architectural requirement here is that training and inference use the same code path. A feature store providing point-in-time correct lookups is the mature solution; a shared, versioned library is the minimum acceptable one.

Model layer. Training infrastructure, experiment tracking, the model registry, and the artefacts themselves. Development, staging and production environments with controlled promotion between them.

Risk prediction engine. Where a score becomes a decision. It applies operating thresholds, suppression rules, eligibility filters, and routing logic; determines whether this patient at this moment warrants an output; and decides which channel it goes to. Keeping this logic explicit and configurable — rather than buried in the model or in the interface — means thresholds can be tuned without retraining and without a software release.

Clinical decision support interface. What the clinician sees, and how it appears in their workflow. Chapters 10 to 12 cover the design requirements.

Clinician review and action. Human evaluation and the resulting clinical decision. The architecture must capture what happened here: was the output seen, was it acted on, was it overridden, what was done. Without this, the loop cannot close.

Outcome monitoring. Links predictions to subsequent outcomes and actions. Feeds performance monitoring, drift detection, safety review, and — eventually — model updating.

Non-functional requirements

Requirement	Typical expectation
Latency (interactive)	Under 1 second from trigger to display
Latency (batch)	Complete within the operational window (e.g. overnight)
Availability	Matching the host clinical system, typically 99.9%+
Failure mode	Fail visibly, never silently; suppress output rather than emit stale predictions
Scalability	Peak load with headroom; measured, not assumed
Auditability	Every prediction reconstructible from stored inputs and model version
Recoverability	Defined RPO and RTO consistent with clinical criticality

The failure mode row deserves emphasis. A model that returns a plausible number when its inputs are stale is more dangerous than one that returns nothing, because nothing is obviously wrong and a plausible number is not.

Build, buy, or partner

Approach	Advantages	Disadvantages
Build in house	Fits local population and workflow; retains data and expertise; full control of updating	Requires sustained multidisciplinary capability; you carry the entire validation and safety burden
Buy commercial	Faster; vendor holds regulatory clearance where applicable; support included	Performance may not transfer; limited configurability; opacity; dependency risk
Partner / co-develop	Shared cost and expertise; models fitted to local data	Governance complexity; unclear IP and liability; sustainability depends on the partner

Whichever route, the local validation obligation is identical. A cleared commercial product with published performance still requires evaluation in your population before it is trusted there. The Epic sepsis model was, at the time of the Michigan evaluation, deployed at hundreds of hospitals on the strength of vendor-reported performance.

Scaling patterns

Multi-site. Central development with local recalibration is usually the right default: one model, site-specific intercepts and thresholds. Fully separate per-site models are appropriate only where populations genuinely differ in the underlying relationships, and they multiply the validation and monitoring burden by the number of sites.

Multi-model. As a portfolio grows past a handful, shared infrastructure becomes essential — a common feature store, a single registry, unified monitoring, one governance process. Organisations that build each model as a bespoke project reach an unmaintainable state at around five to ten models.

Cost. Inference cost is usually trivial for tabular models and material for imaging and large language models. Data storage and movement are

frequently the dominant line. Model the running cost before committing, including the cost of the clinical time the system will consume.

KEY TAKEAWAY

The reference architecture runs from data sources through integration, quality, features, model, prediction engine and interface to clinician action and outcome monitoring — with a feedback loop that most diagrams omit. Keep threshold and routing logic in a configurable prediction engine rather than inside the model. Ensure training and serving share one feature code path. Design the system to fail visibly, and build shared infrastructure before the portfolio outgrows bespoke projects.

CHAPTER 18

Continuous Monitoring and Model Governance

Deployment is the beginning of the model's life, not the end.

A clinical prediction model is not a piece of software that is finished. It is a measurement instrument operating in a changing environment, and like any instrument it drifts and requires recalibration.

The framing that works best in practice: treat the model as a member of clinical staff whose competence is periodically assessed, rather than as a device that is installed. It has a scope of practice, its performance is reviewed, it can be retrained, and it can be retired.

What changes after deployment

Population. Case mix shifts. A service expands to a new catchment. Referral criteria change. Demographics move.

Practice. A new guideline alters treatment thresholds. A drug becomes first-line. A diagnostic pathway is redesigned. Documentation and coding practice evolve.

Technology. A laboratory changes analyser. The EHR is upgraded and a field is repurposed. A device vendor changes sampling frequency.

The outcome itself. Diagnostic criteria are redefined. Sepsis-3, published in JAMA in 2016, materially changed how sepsis is identified; any model trained on earlier definitions was predicting a different construct afterwards.

The model's own effect. Successful intervention reduces the outcome rate in the flagged population, degrading apparent performance. Chapter 2 introduced this; it is worth restating because it confounds monitoring in a way teams rarely anticipate. Improvement can look like failure.

What to monitor

Input monitoring is the earliest signal and does not require outcomes.

- Distribution of each feature versus the training distribution
- Missingness rates per feature
- Volume of records scored
- Data freshness and pipeline latency
- New or unmapped codes appearing

Prediction monitoring is also outcome-independent.

- Distribution of predicted probabilities
- Alert volume, absolute and per thousand patients
- Proportion of the population above threshold
- Breakdown by unit, service and subgroup

Performance monitoring requires outcomes and therefore lags by the prediction horizon.

- Discrimination (ROC-AUC and PR-AUC) with confidence intervals
- **Calibration**, plotted, with slope and intercept — the first thing to degrade and the most consequential
- Sensitivity and PPV at the operating threshold
- Subgroup performance across all monitored groups

Usage monitoring captures whether the system is functioning as a clinical intervention.

- Proportion of outputs viewed
- Override and dismissal rates
- Time from alert to action
- Documented actions following alerts
- Clinician-reported usefulness

Outcome monitoring is the point of the exercise.

- The clinical outcome the system exists to improve

- Process measures on the intervention pathway
- Balancing measures: workload, unnecessary investigation, cost
- Safety incidents connected to the system

> **IMPORTANT CONSIDERATION**
>
> Monitoring without pre-defined thresholds and named owners is data collection, not monitoring. Before go-live, write down for each metric: the acceptable range, the value that triggers review, the value that triggers suspension, who receives the report, how often, and who has authority to act. A dashboard nobody is accountable for will be watched attentively for six weeks and then ignored.

Detecting drift

Statistical distance measures — population stability index, Kolmogorov–Smirnov, Jensen–Shannon divergence — compare current input distributions to a reference. Population stability index is common in practice with conventional interpretation bands (below 0.1 stable, 0.1–0.25 moderate shift, above 0.25 significant shift), though these are heuristics rather than laws.

Sequential monitoring methods such as CUSUM detect gradual change more sensitively than repeated independent tests, which is the right property here since real drift is usually slow.

Calibration drift is often the first detectable performance change and the most clinically important. Track calibration-in-the-large and slope on a rolling window.

Beware multiple testing. A dashboard tracking forty metrics weekly will produce alarms by chance. Define which signals matter and require confirmation before acting.

A monitoring schedule

Frequency	Activity
Continuous	Pipeline health, latency, error rates, prediction volume

Frequency	Activity
Daily	Alert volume; input completeness; automated anomaly checks
Weekly	Override rates; unit-level alert distribution; feedback triage
Monthly	Calibration; input drift; subgroup alert rates; incident review
Quarterly	Full performance evaluation with confidence intervals; subgroup performance; usage and workflow review; report to governance
Annually	Formal revalidation; documentation refresh; fairness audit; decision on continuation, update or retirement
Event-driven	Any incident; any material change to source systems, guidelines or population

Updating

Recalibration adjusts intercept and slope to restore agreement between predicted and observed rates, leaving the model's ranking untouched. The lowest-risk and most frequently appropriate intervention.

Refitting retrains the same specification on newer data. Moderate risk; requires revalidation.

Redevelopment changes features or architecture. Effectively a new model, requiring the full development and validation pathway.

Continuous learning — models that update automatically from incoming data — is technically feasible and governance-hostile. Automatic updating means the deployed model is not the validated model, performance can degrade without anyone changing anything, and feedback loops can be reinforced silently. Regulators have developed mechanisms such as predetermined change control plans to allow controlled adaptation within pre-specified bounds. Outside such a framework, treat automatic updating in clinical settings as requiring exceptional justification.

Governance

An organisation running clinical AI needs a standing body with real authority. Its remit:

- Approving new models for development and for deployment
- Reviewing validation and safety documentation
- Setting and approving operating thresholds
- Receiving performance and fairness reports on a schedule
- Overseeing incidents
- Approving updates
- Deciding retirement

Composition should include clinical leadership, informatics, data science, information governance, clinical safety, patient representation, and operational management. The patient representative seat should carry the same standing as the others; a token appointment produces token oversight.

Retirement. Models should be withdrawn when performance falls below the agreed threshold and cannot be restored, when the clinical pathway changes such that the prediction no longer supports a decision, when a better alternative exists, when the workload cost exceeds the benefit, or when nobody is willing to own it. That last criterion is a real and useful test: a model without an accountable owner should not be running.

> **KEY TAKEAWAY**
>
> Models degrade because the world changes, and calibration degrades first. Monitor inputs, predictions, performance, usage and outcomes — with pre-agreed thresholds and named owners, or it is not monitoring. Prefer recalibration to refitting and refitting to redevelopment. Treat continuous automatic learning as requiring exceptional justification, and be willing to retire models, including successful ones whose moment has passed.

PART VII

Practical Applications

Five worked case studies, each ending at the same place: this is risk, not diagnosis.

CHAPTER 19

Practical Healthcare AI Use Cases

Applying the method to five conditions.

The five case studies in this chapter are **illustrative constructions**. They demonstrate how the framework in Parts II through VI applies to a specific clinical problem. They are not descriptions of particular real deployments, they do not report the performance of any commercial product, and the numbers used to illustrate operating characteristics are worked examples rather than empirical findings.

Where a real published result is cited, it is identified as such.

Each case follows the same structure, and the repetition is deliberate — the structure *is* the method.

Case Study 1 — Cardiovascular Risk in Primary Care

Clinical problem. Cardiovascular disease remains a leading cause of death globally. Effective preventive treatments exist — lipid lowering, blood pressure control, smoking cessation support — and benefit is concentrated in higher-risk patients. The decision is who to offer intensified prevention to, and how strongly to recommend it.

Existing standard. This is a mature area. Multivariable risk equations are embedded in guidelines internationally. The American Heart Association's PREVENT equations, developed and validated in a large multicohort dataset and published in *Circulation* in 2024, estimate 10-year and 30-year risk of total cardiovascular disease including heart failure, and were deliberately developed without race as an input, incorporating kidney function and, optionally, a social deprivation index.

Where AI might add value. Any new model must beat this benchmark, which is a high bar. The plausible sources of improvement are longitudinal trajectory (blood pressure and lipid trends over years rather than a single reading), medication adherence patterns from dis-

pensing data, and variables not in the standard equations such as inflammatory markers or imaging-derived measures where available.

Element	Specification
Population	Adults 40–75, no established cardiovascular disease, registered ≥24 months
Index time	Routine primary care contact or annual review
Lookback	5 years of structured data
Horizon	10-year risk of first major cardiovascular event
Outcome	Myocardial infarction, stroke, or cardiovascular death from linked records
Features	Age, sex, deprivation index, smoking, BP trajectory, lipid trajectory, HbA1c, eGFR, BMI trend, family history, antihypertensive and statin history

Decision support design. Not an alert. A field in the annual review template showing estimated risk with contributing factors, and a counterfactual: what risk would be with blood pressure at target and on a statin. The clinician uses it in a shared decision-making conversation.

Equity considerations. Deprivation is a strong predictor and is also a marker for reduced access to the intervention. Monitor whether treatment initiation following a high-risk flag differs by deprivation quintile — if the model identifies deprived patients as high risk but they are less likely to receive treatment, the system has documented an inequity without fixing it.

What this is not. An estimate of population-level probability. It does not tell this patient whether they will have a heart attack, and the interface should say so in words.

Case Study 2 — Type 2 Diabetes Risk

Clinical problem. A large proportion of people with type 2 diabetes are undiagnosed, and a larger population has prediabetes. Structured lifestyle intervention has a substantial evidence base for reducing progression. The constraint is programme capacity, so the decision is who to invite.

Existing standard. Established risk scores and opportunistic HbA1c testing. Straightforward to apply and dependent on the patient attending.

Where AI might add value. Not primarily in individual accuracy — the existing scores are decent — but in ranking an entire registered population automatically, without requiring a consultation, and in incorporating trajectory: an HbA1c moving from 38 to 43 mmol/mol over three years is a different signal from a stable 43.

Element	Specification
Population	Adults 25–79 without diagnosed diabetes
Index time	Monthly batch scoring of the whole registered list
Lookback	5 years
Horizon	3-year incident type 2 diabetes
Outcome	Diagnosis code, HbA1c ≥48 mmol/mol on two occasions, or initiation of glucose-lowering therapy
Features	Age, sex, ethnicity where clinically justified, deprivation, BMI and BMI trajectory, HbA1c and fasting glucose trajectories, blood pressure, lipids, family history, gestational diabetes history, relevant medications, prior test frequency

Illustrative operating characteristics. Suppose the practice can invite 200 patients a quarter from a list of 12,000. The model ranks all 12,000; the top 200 are invited. If the observed 3-year incidence in that top group is 22% against a population rate of 3%, the programme is reach-

ing a substantially enriched group — which is the entire purpose. Note that the metric that matters here is enrichment in the top N, not AUC.

Design. No alert. A quarterly worklist owned by the practice nurse, integrated with the invitation system, with clinician review before invitations are sent.

Equity considerations. Prior test frequency is predictive and is also a marker of engagement. A model leaning heavily on it will under-rank patients who rarely attend — who are frequently those at highest risk. Test this explicitly by comparing model rank against attendance frequency, and consider constraining or removing the feature if the association is strong.

What this is not. Diagnosis. Every invited patient requires a diagnostic test.

Case Study 3 — Chronic Kidney Disease Progression

Clinical problem. Most patients with CKD never progress to kidney failure; a minority do, and for them, timely nephrology referral, dialysis planning, and transplant work-up materially affect outcomes. Late presentation to renal services is associated with worse outcomes and higher costs. The decision is who to refer and when.

Existing standard. This is the strongest example in the book of a risk model changing practice. The Kidney Failure Risk Equation, developed by Tangri and colleagues and published in JAMA in 2011, uses age, sex, eGFR and urine albumin-creatinine ratio in its four-variable form. It has been externally validated in more than 30 countries and is embedded in referral guidance internationally.

Where AI might add value. Trajectory again — eGFR slope over years is powerfully informative and the four-variable KFRE uses a single cross-sectional value. Additional laboratory markers, medication history, and comorbidity may add incremental discrimination. Any new model should be benchmarked directly against KFRE, and the honest

possibility is that it will not beat it by enough to justify the added complexity. That is a legitimate result.

Element	Specification
Population	Adults with stage 3–4 CKD, not already under nephrology care
Index time	Any eGFR result meeting criteria
Lookback	3 years
Horizon	2-year and 5-year risk of kidney failure requiring replacement therapy
Outcome	Dialysis initiation, transplant, or sustained eGFR <10
Features	Age, sex, eGFR most recent and slope, uACR, systolic BP mean, diabetes status, calcium, phosphate, bicarbonate, albumin, haemoglobin, ACEi/ARB/SGLT2 inhibitor use, hospitalisation history

Design. A ranked list for the primary care–nephrology interface, reviewed at a monthly multidisciplinary meeting. Risk expressed as absolute two-year and five-year percentages, because referral thresholds in guidance are expressed that way.

Monitoring subtlety. If the system works, referral and treatment intensification increase in the flagged group, and observed progression falls below prediction. The model will appear to be over-predicting. This is success, and it must be anticipated in the monitoring plan or it will be misread as drift.

What this is not. A referral decision. It is one input to a clinical judgement that includes patient preference, frailty, comorbidity and goals of care.

Case Study 4 — Sepsis and Early Deterioration

Clinical problem. Sepsis is common, time-critical, and frequently recognised late. Each hour of delay in appropriate antibiotics in septic

shock is associated with worse outcomes. The decision is which deteriorating patient needs urgent senior review now.

Existing standard. Aggregate early warning scores based on bedside observations, computed manually or automatically, with defined escalation. Sepsis-3, published in JAMA in 2016, provides the current consensus definition and the qSOFA criteria.

The cautionary evidence. This is the domain with the most prominent published failure. A widely deployed proprietary sepsis prediction model, evaluated externally by Wong and colleagues in JAMA *Internal Medicine* in 2021 across 38,455 hospitalisations, achieved an AUC of 0.63, identified 33% of sepsis cases at the recommended threshold with a PPV of 12%, and generated alerts on 18% of hospitalisations.

Anyone building or buying in this space should read that paper before writing a specification.

Element	Specification
Population	Adult inpatients outside critical care
Index time	Hourly, or on each new observation set
Lookback	Current admission
Horizon	Sepsis onset within 6–12 hours
Outcome	Sepsis-3 criteria applied to structured data, manually validated on a sample
Features	Vital sign values, trends and variability; laboratory results and trends; medications; observation frequency; time since admission; surgical status; comorbidity

Design requirements, given the above. Threshold set for positive predictive value rather than sensitivity, accepting missed cases in exchange for a usable alert rate. Alerts routed to the nurse in charge, not broadcast. Suppression for 12 hours after an assessed and documented alert. Every alert carries the specific contributing observations. The existing observation-based escalation continues unchanged — the

model supplements it and never replaces it. Missed cases reviewed monthly against mortality and ICU transfer.

What this is not. A diagnosis of sepsis, and not a substitute for clinical assessment at the bedside. The absence of an alert is not evidence that a patient is well, and training must say so directly.

Case Study 5 — Cognitive Decline

Clinical problem. Dementia is commonly recognised late. Earlier identification allows investigation of reversible contributors, advance care planning, support for carers, and — for a subset with confirmed Alzheimer's pathology — consideration of disease-modifying therapy where available and appropriate. The decision is who to assess further.

Status of the evidence. This is the least mature of the five cases and should be treated as research-grade. Blood-based biomarkers and imaging have advanced substantially, but prediction of cognitive decline from routine clinical data alone remains an emerging area rather than an established one.

Element	Specification
Population	Adults 65+ without a dementia diagnosis
Index time	Annual batch scoring
Lookback	5 years
Horizon	3-year incident dementia diagnosis
Outcome	Diagnosis code, dementia-specific medication, or specialist assessment — with the ascertainment limitations documented
Features	Age, education proxy, deprivation, vascular risk factors, hearing impairment, medication changes suggesting cognitive difficulty, missed appointments, accident and injury history, weight loss, and structured cognitive test results where recorded

Design. A very cautious one. Output goes to a clinician for consideration, never to a patient, and never as an automated communication. The wording matters enormously: "factors present that are associated with increased likelihood of cognitive decline; consider assessment if clinically appropriate" rather than any statement resembling a prediction of dementia.

Ethical weight. This is the case where the harms of a false positive are least about resource use and most about the person. Being told, or coming to believe, that an algorithm expects you to develop dementia is a serious harm in itself. Patient and carer representatives should be involved in designing the communication approach from the beginning, not consulted on a finished product.

Equity considerations. Diagnosis rates for dementia vary substantially by ethnicity, language, and deprivation, largely reflecting access and assessment practice rather than incidence. A model trained on recorded diagnoses inherits that pattern directly.

What this is not. Any kind of diagnostic statement. The gap between "associated risk factors are present" and "this person has a neurodegenerative disease" is enormous, and every element of the interface must preserve it.

The common thread

Read across the five and the same points recur.

The existing standard of care is usually a decent risk score, and beating it is harder than it looks. Trajectory features are consistently where machine learning contributes. Delivery mode should be a worklist unless the decision is genuinely time-critical. Features that encode access or engagement will systematically disadvantage the patients who most need identifying. Successful intervention will make the model look miscalibrated. And in every single case, the output is a probability about a population of similar patients — never a diagnosis of the person in front of you.

KEY TAKEAWAY

The same method applies across conditions: define the decision, benchmark against the existing standard, fix the prediction point, choose a delivery mode proportionate to urgency and precision, audit for access-related features, and anticipate that success will look like miscalibration. Cardiovascular, diabetes and CKD prediction are mature; sepsis is deployed but inconsistent; cognitive decline is emerging and demands the greatest caution in how anything is communicated.

PART VIII

A Practical Framework for Responsible Healthcare AI

Ten principles and nine phases, distilled from everything preceding.

CHAPTER 20

The Responsible Healthcare AI Framework

Ten questions that decide whether a system should be built, deployed, and kept running.

Everything in this book condenses to ten questions. They are ordered deliberately: each one is a gate, and a project that fails an early gate should not proceed to a later one regardless of how well it would perform there.

The framework is intended to be used three times — at initiation, before deployment, and at each annual review — with the answers written down and dated.

1. Clinical Need

Does this solve a meaningful healthcare problem?

- Is the clinical problem clearly stated, with a decision and a decision-maker?
- Does a specific, available action follow from a positive prediction?
- Is there evidence that the action improves outcomes in the flagged group?
- What is the current standard of care, and is AI plausibly better than improving that?
- Have clinicians and patients confirmed the problem is worth solving?

Fails when: the project began with available data rather than a clinical question, or when the intervention that would follow a flag does not exist.

2. Data Quality

Is the underlying data reliable and representative?

- Are sources documented for provenance, completeness and latency?
- Has the outcome definition been manually validated against records?
- Is the missingness mechanism understood for each key predictor?
- Does the training population resemble the deployment population, demographically and clinically?
- Are all required inputs actually available at prediction time in production?
- Is the extraction reproducible?

Fails when: a key predictor turns out to be unavailable in production, or the training cohort excluded the complex patients the system will encounter.

---

3. Model Performance

Does the model demonstrate appropriate predictive performance?

- Discrimination reported with confidence intervals, on a proper hold-out?
- Calibration assessed and plotted, not just tested?
- Precision-recall reported alongside ROC for a rare outcome?
- Performance stated in absolute terms at the operating threshold?
- A meaningful baseline outperformed?
- Sample size formally adequate for the model's complexity?

Fails when: the only evidence is an AUC from a random split, or no comparator exists.

---

4. Clinical Validation

Has the system been appropriately validated?

- Temporal validation performed?
- External validation performed, or its absence explicitly declared?
- Local validation performed on the deployment population?
- A silent prospective period run before any clinical exposure?
- Clinical utility assessed — decision curve analysis or equivalent?
- Reporting follows TRIPOD+AI; risk of bias assessed against PROBAST+AI?

Fails when: the model goes live on the strength of vendor-reported or development-set performance.

---

5. Explainability

Can users understand the basis and limitations of predictions?

- Was an interpretable model considered, and is the added complexity justified by measured gain?
- Does each prediction carry clinically meaningful contributing factors?
- Are modifiable factors distinguished from fixed ones?
- Are missing inputs disclosed at the point of use?
- Is model documentation reachable in one click from the tool?
- Have explanations been tested for their effect on decisions rather than assumed helpful?

Fails when: the clinician-facing output is a number with no rationale, or the explanation is a data-science artefact nobody clinical can read.

---

6. Human Oversight

Are qualified professionals involved in consequential decisions?

- Is the human-in-the-loop configuration appropriate to the stakes?
- Can a clinician override easily, and is the override recorded?
- Is the delivery mode proportionate to urgency and to PPV?
- Has alert burden been quantified against the service's capacity?
- Are automation bias mitigations in place?
- Is accountability documented — who is responsible for what?

Fails when: a low-PPV model is wired to an interruptive alert, or nobody has written down who owns the output.

---

7. Equity

Does the system perform appropriately across relevant populations?

- Were subgroups defined in advance with clinical and community input?
- Is discrimination *and calibration* reported per subgroup?
- Has the outcome definition been examined for differential ascertainment?
- Has a fairness criterion been chosen and justified in writing?
- Is the intervention equally accessible to everyone flagged?
- Will subgroup performance be monitored after deployment?
- Does deployment narrow or widen the existing disparity?

Fails when: subgroup analysis was not done, or was done and the disparity was noted and not addressed.

---

8. Privacy and Security

Is patient information appropriately protected?

- Is there a lawful basis for the intended use, confirmed with qualified advice?
- Are minimisation, access control, encryption, audit logging and retention in place?
- Have re-identification risks been assessed for any shared data?
- Is the trained model itself protected as a data asset?
- Are third-party arrangements contractually sound?
- Is there a tested incident response plan?
- Would patients recognise this use as consistent with what they were told?

Fails when: compliance was treated as a document rather than a set of implemented controls.

---

9. Scalability

Can the system operate reliably in real-world healthcare environments?

- Does it meet latency, availability and throughput requirements under peak load?
- Does it fail visibly rather than silently?
- Do training and serving share one feature code path?
- Are versioning, logging and a model registry in place?
- Is it integrated into existing workflow rather than bolted alongside it?
- Are running costs, including clinical time, understood?
- Can it extend to additional sites without redevelopment?

Fails when: the pilot depended on manual data preparation that cannot scale.

10. Continuous Monitoring

Can performance and safety be monitored after deployment?

- Are inputs, predictions, performance, usage and outcomes all monitored?
- Does every metric have a threshold, an owner and a reporting cadence?
- Is subgroup performance monitored, not only overall?
- Is there an incident reporting route with a feedback loop?
- Is revalidation scheduled?
- Are update and change control processes defined?
- Are retirement criteria written down, with named authority to act on them?

Fails when: monitoring is a dashboard nobody owns.

Using the framework

Score each principle as **Met**, **Partially met**, or **Not met**, with evidence and a date. Record the gaps and who is accountable for closing them.

A reasonable governance posture:

- Any principle **Not met** blocks deployment until resolved or formally risk-accepted by a named individual with authority.
- Any principle **Partially met** requires a documented plan with an owner and a date.
- The full assessment is repeated annually and on any material change.

The value of the framework is not in the scoring. It is in the conversations the questions force between people who would otherwise not have them — the data scientist who has not asked what happens to a flagged patient, the clinician who has not asked what the positive pre-

dictive value is, and the executive who has not asked who switches it off.

KEY TAKEAWAY

Ten principles, applied as sequential gates: clinical need, data quality, model performance, clinical validation, explainability, human oversight, equity, privacy and security, scalability, and continuous monitoring. Assess at initiation, before deployment, and annually. Record evidence and dates, and require a named individual to accept any residual risk.

CHAPTER 21

The Healthcare AI Implementation Roadmap

Nine phases, with entry and exit criteria.

The framework in Chapter 20 asks whether a system should exist. This chapter sets out how to get there.

```
Phase 1 — Identify the Problem
Phase 2 — Assess Data Availability
Phase 3 — Develop the Model
Phase 4 — Validate
Phase 5 — Clinical Evaluation
Phase 6 — Pilot Deployment
Phase 7 — Integrate Into Workflow
Phase 8 — Monitor and Govern
Phase 9 — Improve and Revalidate
```

Phases 8 and 9 loop indefinitely. The roadmap has no end state.

Phase 1 — Identify the Problem

Typical duration: 4–8 weeks

Engage clinicians, operational leads and patient representatives. Define the decision, the decision-maker, and the action that follows. Confirm the intervention exists and has capacity. Document the current standard of care. Establish that data plausibly exists.

Exit criteria: a written problem statement; a named clinical sponsor; confirmed intervention capacity; agreement that AI is the right tool rather than a process change.

Most common mistake: compressing this phase because it produces no code.

Phase 2 — Assess Data Availability

Typical duration: 6–12 weeks

Inventory sources. Assess quality, completeness and latency. Validate the outcome definition against manually reviewed records. Construct and characterise the cohort. Check demographic representativeness. Confirm production availability of every candidate input. Secure governance approval.

Exit criteria: documented data availability specification; validated outcome definition; adequate event count; approvals in place.

Most common mistake: assuming a field is populated because it exists in the schema.

Phase 3 — Develop the Model

Typical duration: 8–16 weeks

Split first, by patient and by time. Engineer features. Fit a baseline. Fit candidate models. Tune within the training set. Audit for leakage. Select on the validation set.

Exit criteria: a selected model with a locked specification; a full experiment log; a documented leakage audit; a baseline comparison.

Most common mistake: iterating against the test set until it stops being a test set.

Phase 4 — Validate

Typical duration: 6–12 weeks

Evaluate once on the untouched holdout. Report discrimination, calibration and clinical utility. Perform subgroup analysis. Obtain external data if it exists. Choose the operating threshold from clinical consequence and alert capacity. Write the validation report to TRIPOD+AI and self-assess against PROBAST+AI.

Exit criteria: validation report complete; performance clinically adequate; subgroup performance acceptable or disparities documented with a mitigation plan; threshold agreed with clinical leadership.

Most common mistake: choosing the threshold to optimise a statistic rather than to fit the service's capacity.

Phase 5 — Clinical Evaluation

Typical duration: 3–6 months

Run silently in production. Compare prospective to retrospective performance. Measure real alert volume. Test the integration end to end. Complete the hazard analysis. Design the interface with clinical users. Prepare training and documentation.

Exit criteria: prospective performance consistent with retrospective; alert volume within capacity; technical reliability demonstrated; safety case signed off; governance approval to expose output to clinicians.

Most common mistake: skipping this phase entirely.

Phase 6 — Pilot Deployment

Typical duration: 3–6 months

Expose output to a limited group in a limited setting. Train them properly, on honest local numbers. Collect quantitative usage data and qualitative feedback. Track process and balancing measures. Iterate on the interface. Review incidents.

Exit criteria: clinicians report the tool useful; workload impact acceptable; no unresolved safety concerns; a signal on process measures; documented iteration.

Most common mistake: piloting with enthusiasts only, then being surprised by general adoption.

Phase 7 — Integrate Into Workflow

Typical duration: 3–6 months

Full EHR integration. Update procedures, role descriptions and training. Roll out in stages. Establish support routes. Complete documentation.

Exit criteria: integrated and stable; all users trained; procedures updated; support in place; monitoring operational.

Most common mistake: rolling out to everyone at once and having no capacity to respond when something breaks.

Phase 8 — Monitor and Govern

Continuous

Execute the monitoring schedule from Chapter 18. Report to governance. Investigate incidents. Track outcomes.

Ongoing criteria: performance within thresholds; subgroup performance stable; usage sustained; incidents managed; reporting current.

Most common mistake: monitoring attentively for a quarter and then quietly stopping.

Phase 9 — Improve and Revalidate

Annual, and event-driven

Formal revalidation. Fairness audit. Decide: continue, recalibrate, refit, redevelop or retire. Refresh documentation. Reassess against the ten principles.

Ongoing criteria: revalidation current; documentation accurate; a documented decision on continuation.

Most common mistake: allowing a model to persist because retiring it feels like admitting failure.

Organisational Readiness Checklist

Before a healthcare organisation begins an AI programme at all:

Strategy and sponsorship

- ☐ Executive sponsor with budget authority
- ☐ Named clinical sponsor with credibility among users
- ☐ Programme aligned to a stated organisational priority
- ☐ Multi-year funding, not a single project grant

Capability

- ☐ Data engineering capacity
- ☐ Data science capacity with healthcare-specific experience
- ☐ Clinical informatics capacity
- ☐ Clinical safety expertise
- ☐ Information governance expertise
- ☐ Change management and training capacity

Infrastructure

- ☐ Reliable access to source data
- ☐ Environment for development with appropriate controls
- ☐ Model serving and integration capability
- ☐ Monitoring infrastructure
- ☐ Version control and registry

Governance

- ☐ AI governance committee with real authority
- ☐ Patient representation with standing equal to other members
- ☐ Defined approval pathway from concept to deployment
- ☐ Incident reporting route covering AI-specific events

- ☐ Documented accountability model
- ☐ Regulatory and legal advice available

Culture

- ☐ Clinicians involved in design, not consulted on finished products
- ☐ Realistic expectations, including about timelines
- ☐ Willingness to stop projects that are not working
- ☐ Transparency with patients about data use
- ☐ Learning response to failure rather than a blaming one

IMPORTANT CONSIDERATION

Realistic timelines matter. From problem identification to full deployment is commonly eighteen months to three years for a first model in an organisation without existing infrastructure. Programmes promised in six months either skip validation and clinical evaluation, or do not deliver. The second failure is far preferable to the first, and the honest conversation about duration is best had at the beginning.

KEY TAKEAWAY

Nine phases, each with explicit entry and exit criteria, and the last two loop forever. The phases most often compressed — problem identification, data assessment, and silent clinical evaluation — are the ones that determine whether the rest succeeds. Expect eighteen months to three years for a first deployment, and be more suspicious of a fast timeline than of a slow one.

PART IX

The Future of Healthcare AI

What is arriving, what is still research, and what will still be true in ten years.

CHAPTER 22

The Future of Predictive and Clinical AI

Emerging capabilities, honestly labelled.

Predicting the future of a fast-moving field is a good way to look foolish in retrospect. What follows is organised by how established each area is, and readers should treat the classification as more useful than any specific forecast.

Arriving now

Multimodal models. Systems combining structured data, imaging, waveforms and text in a single model. The clinical logic is compelling — clinicians integrate exactly these sources — and technical results are promising. The practical obstacles are unglamorous: aligning modalities in time, handling patients missing entire modalities, and the fact that adding an expensive input must be justified by the incremental benefit over the cheap ones. A model requiring an MRI to add two points of AUC over a model requiring a blood test is not an improvement in most settings.

Large language models for documentation. This is the area that has moved fastest into routine use, and largely for a reason unrelated to prediction: it addresses clinician workload, which is the problem clinicians actually have. Ambient scribing, draft correspondence, coding support and record summarisation are being adopted at scale. The governance requirements are real — verification against source data, handling of confabulation, audit trails — but the value proposition is clearer than for most predictive applications.

Language models as feature extractors. Rather than asking a language model to predict risk, use it to convert unstructured notes into structured variables that feed a conventional risk model. This plays to the technology's strengths and keeps calibrated probability estimation in the hands of methods designed for it. It is, in my view, the most underrated architecture in clinical AI at present.

Remote monitoring at scale. Continuous glucose monitoring, home blood pressure, implantable device telemetry and consumer wearables generate dense longitudinal data. Integration into clinical workflow lags the data availability substantially, and the unsolved problem is not analysis but responsibility: who is monitoring the stream, during which hours, and what happens at 2 a.m.

Emerging

Digital biomarkers. Measurable indicators derived from digital devices — gait from an accelerometer, typing dynamics, speech characteristics, sleep architecture. Genuinely promising for conditions where subtle functional change precedes clinical presentation, including neurodegenerative disease. Validation is at an early stage and standardisation across devices is minimal.

Federated learning in practice. The technique is established; routine operational use across health systems is not. The barriers are more organisational than technical — data harmonisation, governance agreements, and the question of who owns the resulting model.

Synthetic health data. Useful now for development, testing and education. Its role in validation remains contested, and privacy guarantees are weaker than often claimed unless generated under a formal mechanism such as differential privacy.

AI agents. Systems that plan and execute multi-step tasks rather than answering single queries. Plausible near-term applications are administrative — prior authorisation, referral coordination, appointment logistics. Autonomous clinical action is a substantially different proposition and the accountability questions are not close to resolved.

Polygenic risk scores in routine care. Incremental value over conventional risk factors for common disease is modest, and transportability across ancestry groups is a known and serious limitation given the composition of existing genomic datasets.

Research-stage

Causal inference for treatment effect estimation. Arguably more important than better prediction. Clinicians need to know what will happen *if they act*, not only what will happen. Estimating individualised treatment effects from observational data is methodologically hard and progressing.

Foundation models for health records. Models pre-trained on large volumes of clinical data and adapted to many downstream tasks. Early results are interesting; questions about validation, drift and governance for a general-purpose clinical model are largely unanswered.

Real-time closed-loop systems. Automated insulin delivery is the established example and demonstrates the pattern: a narrow, well-characterised physiological system with a fast feedback signal. Extension to broader physiological control remains distant.

> **IMPORTANT CONSIDERATION**
>
> A recurring pattern in this field is worth naming: the technical capability arrives several years before the organisational capacity to use it safely. Deep learning could read retinal photographs in 2016; screening programmes using it took considerably longer to establish, and the delay was about pathways, liability, reimbursement and workforce rather than about the model. Assume the same lag for everything in the "emerging" list above. The constraint is rarely the algorithm.

What will not change

Three things are worth betting on.

Data quality will remain the limiting factor. No architecture compensates for a badly defined outcome or a biased cohort.

Validation and monitoring will remain mandatory. More capable models do not need less oversight. Arguably they need more, because their failure modes are harder to anticipate.

Human judgement will remain necessary. Clinical decisions involve values, preferences, context and responsibility. Better prediction improves one input to that process without replacing it.

> **KEY TAKEAWAY**
>
> Language models for documentation and as feature extractors are arriving now; multimodal prediction, digital biomarkers, federated learning and agents are emerging; causal treatment effect estimation and clinical foundation models are research-stage. Organisational capacity, not technical capability, is the usual bottleneck. Data quality, validation and human judgement will remain necessary regardless of what the models can do.

CHAPTER 23

The Future of Responsible AI in Healthcare

Governance, accountability, and the things that have to hold.

Capability and responsibility are advancing at different speeds. This chapter is about closing that gap.

Human–AI collaboration

The framing of "AI versus clinicians" has always been a poor description of the actual question, which is how to combine two systems with different and partly complementary failure modes.

Models are consistent, tireless, and able to integrate more variables than a person can hold. They fail on unusual cases, on situations unlike their training data, and on anything requiring context outside their inputs.

Clinicians are variable, fatigable, and subject to well-documented cognitive biases. They are also capable of recognising that something is wrong in a way they cannot articulate, of integrating the patient's circumstances and preferences, and of taking responsibility.

The productive design question is which decisions to route where, and how to make the handover work. The evidence on combined human–AI performance is more mixed than the optimistic framing suggests: combinations sometimes outperform both components and sometimes underperform the better one, depending heavily on interface design and on whether the human can tell when the model is likely to be wrong. That last capability — appropriate reliance rather than uniform trust — is the design target.

The governance landscape

Several strands are consolidating.

Regulatory frameworks are extending from devices to systems. The EU AI Act, in force since August 2024, imposes obligations on high-risk AI including risk management, data governance, documentation, logging, transparency, human oversight and robustness. Device regulators including the FDA have been developing approaches for AI that changes after authorisation, notably predetermined change control plans that permit specified modifications within pre-agreed limits.

Transparency requirements are arriving through health IT certification as well as through AI-specific law. The HTI-1 rule's source attribute requirements for predictive decision support are an early example of a broader direction.

Standards are maturing. NIST's AI Risk Management Framework (AI RMF 1.0, 2023) provides a voluntary structure organised around governing, mapping, measuring and managing AI risk. ISO/IEC 42001 specifies requirements for an AI management system. Neither is healthcare-specific, and both are being adopted in healthcare settings.

Professional consensus is filling the gaps faster than regulation. TRIPOD+AI for reporting, PROBAST+AI for bias assessment, SPIRIT-AI and CONSORT-AI for trials, DECIDE-AI for early clinical evaluation, and FUTURE-AI as an international consensus guideline for trustworthy and deployable healthcare AI, published in the *BMJ* in 2025. These are the practical instruments available today.

Assurance infrastructure is being built. The Coalition for Health AI has developed a model card format and a public registry, alongside proposals for certified assurance laboratories that would independently evaluate health AI tools. Whether independent assurance becomes standard practice is one of the more consequential open questions in the field.

International guidance. The World Health Organization published *Ethics and governance of artificial intelligence for health* in 2021, setting out six core principles, and followed it in 2024 with specific guidance on large multi-modal models. These matter particularly for health systems without well-developed national frameworks.

Accountability

The hardest unresolved problem. When an AI-informed decision causes harm, responsibility is distributed across a developer who built the model, an organisation that chose to deploy it, a governance committee that approved it, and a clinician who acted on it. Existing liability frameworks were not designed for this and are being applied to it awkwardly.

Some things can be established now without waiting for law to settle: written accountability for each component, documented decisions with dates and named individuals, logs sufficient to reconstruct any prediction, incident processes that produce change, and — critically — not placing clinicians in the position of carrying full responsibility for tools they cannot inspect and did not choose.

Equity as infrastructure

The equity risk in healthcare AI operates at two levels.

Within systems, models can encode and scale existing disparities, as Chapter 13 documented. Between systems, the benefits accrue to organisations with the data, capital and expertise to build and buy — which are not the organisations serving the most under-resourced populations. Left alone, the technology widens the gap between well-resourced and poorly-resourced settings.

Counterweights exist: models developed and validated in the populations where they will be used; open publication of validated models; federated approaches allowing smaller institutions to participate in development; and deliberate investment in low-resource deployment, where the potential benefit is arguably greatest given workforce shortages.

Patient trust

Trust is the enabling condition for everything else, and it is neither automatic nor permanent.

What sustains it, based on the available evidence about public attitudes to health data and AI: transparency about what is used and why; clear benefit to patients rather than only to institutions; visible human oversight; meaningful patient involvement in governance; honesty when things go wrong; and proactive disclosure of commercial arrangements.

What destroys it: data sharing discovered through journalism; systems that appear to prioritise cost over care; harms that are minimised or concealed; and the sense that decisions are being made about people rather than with them.

The obligations that persist

Whatever the technology becomes, six commitments hold:

Clinical usefulness. The system must improve decisions or outcomes. Technical performance is a means.

Safety. Anticipate harms, control them, monitor for them, respond when they occur.

Equity. Work appropriately across populations and narrow rather than widen disparities.

Explainability. Users must understand enough to use the system appropriately and to know when not to.

Privacy. Protect patient information and honour the expectations under which it was given.

Sustainability. Systems must be maintainable, monitorable and affordable over their life, and retirable when their time has passed.

KEY TAKEAWAY

Governance is consolidating through regulation, standards and professional consensus, but accountability remains genuinely unresolved and should not be left to resolve itself. Design for appropriate reliance rather than uniform trust. Treat equity as infrastructure, since the technology's default trajectory widens the gap between well-resourced and under-resourced settings. And recognise that patient trust is the enabling condition for all of it.

CHAPTER 24

Final Practical Checklist

Everything, in one place.

This chapter consolidates the checklists from throughout the book into a single reference. It is intended to be used directly — printed, adapted to local terminology, and worked through with the relevant people in the room.

Each item should have an owner and a date.

Clinical Need

- ☐ Clearly defined clinical problem, stated in one sentence
- ☐ Meaningful clinical outcome identified
- ☐ Intended users identified by role
- ☐ Specific action defined for flagged patients
- ☐ Capacity to deliver that action confirmed
- ☐ Evidence that the action improves outcomes, or explicit acknowledgement that it is unproven
- ☐ Current standard of care documented as comparator
- ☐ Clinical sponsor named
- ☐ Patient perspective sought on whether the problem is worth solving

Data

- ☐ Data sources identified with provenance, completeness and latency documented
- ☐ Outcome definition computable and manually validated against records
- ☐ Cohort inclusion and exclusion criteria specified before extraction
- ☐ Index time, lookback window and prediction horizon fixed
- ☐ Data quality assessed with documented cleaning rules and counts
- ☐ Missing data patterns evaluated and mechanism considered per predictor

- ☐ Units, reference ranges and terminologies harmonised
- ☐ Patient identity resolution validated across sources
- ☐ Demographic composition compared to deployment population
- ☐ Bias in data collection, documentation and outcome ascertainment assessed
- ☐ Production availability of every input confirmed at required latency
- ☐ Extraction code version-controlled and reproducible
- ☐ Governance approval covers intended use and future updates

Model

- ☐ Sample size adequacy assessed formally
- ☐ Data split by patient and by time, before any preprocessing
- ☐ All transformations fitted inside training folds only
- ☐ Leakage audit completed feature by feature
- ☐ Baseline model or existing score fitted for comparison
- ☐ Appropriate algorithm selected and the choice justified
- ☐ Hyperparameter search confined to training data
- ☐ Experiments logged with data versions, seeds and metrics
- ☐ Training completed and specification locked
- ☐ Discrimination reported with confidence intervals
- ☐ Calibration plotted, with slope and intercept
- ☐ Precision-recall reported alongside ROC
- ☐ Performance stated in absolute counts at the operating threshold
- ☐ Subgroup performance evaluated, including calibration
- ☐ Clinical utility assessed, ideally by decision curve analysis
- ☐ External or temporal validation completed, or its absence declared
- ☐ Reporting follows TRIPOD+AI
- ☐ Risk of bias self-assessed against PROBAST+AI

Clinical Integration

- ☐ Workflow defined end to end, including who does what

- ☐ Delivery mode chosen proportionate to urgency and positive predictive value
- ☐ Operating threshold justified by clinical consequence and alert capacity
- ☐ Alert volume projected and confirmed within service capacity
- ☐ Alert suppression and de-duplication rules defined
- ☐ Human oversight configuration established and documented
- ☐ Override available, low-friction and logged
- ☐ Output integrated into the existing EHR interface, no separate login
- ☐ Latency requirements specified and met
- ☐ Behaviour on missing, stale or malformed input defined
- ☐ Fallback process defined for model unavailability
- ☐ Clinician feedback mechanism established with a visible response loop
- ☐ Users trained on capabilities, limitations and local performance figures
- ☐ Documentation accessible from the point of use

Responsible AI

- ☐ Lawful basis for intended use confirmed with qualified advice
- ☐ Data minimisation, access control, encryption and retention implemented
- ☐ Audit logging in place and actively monitored
- ☐ Re-identification risk assessed for any shared data
- ☐ Trained model protected as a data asset
- ☐ Third-party arrangements contractually sound
- ☐ Incident response plan documented and tested
- ☐ Subgroups defined in advance with clinical and community input
- ☐ Fairness criterion chosen and justified in writing
- ☐ Disparities mitigated where found; residual disparity disclosed
- ☐ Intervention accessible to all flagged patients
- ☐ Explainability approach defined and clinician-facing view designed for clinicians
- ☐ Model documentation produced in a standard format
- ☐ Hazard analysis completed with controls and residual risk accepted by a named person

- ☐ Clinical safety officer named with defined authority
- ☐ Accountability model documented across developer, organisation and clinician

Deployment

- ☐ Silent prospective period completed and evaluated
- ☐ Pilot completed with usage data and qualitative feedback
- ☐ Staged rollout planned with support capacity
- ☐ Performance monitoring established with thresholds and owners
- ☐ Calibration monitoring established
- ☐ Subgroup performance monitoring established
- ☐ Input drift monitoring established
- ☐ Alert volume and override rate tracked
- ☐ Missed cases reviewed against adverse outcomes
- ☐ Clinical outcome and balancing measures tracked
- ☐ Incident reporting route established, covering AI-specific events
- ☐ Change control and version management in place
- ☐ Governance committee receiving regular reporting
- ☐ Revalidation scheduled
- ☐ Rollback procedure documented and rehearsed
- ☐ Retirement criteria defined, with named authority to act

> **IMPORTANT CONSIDERATION**
>
> A checklist is a tool for structuring a conversation, not a substitute for one. The value comes from a group of people with different expertise working through it together and disagreeing where they see things differently. A checklist completed by one person to satisfy a governance requirement provides documentation and no safety.

Conclusion

What this all comes down to.

The most sophisticated model in this book is worth less than a correctly defined outcome.

That is the summary, and if a reader takes one thing away it should probably be that. The algorithm at the centre of a clinical AI system is the part that has received the most attention, generated the most publication, and consumed the most engineering effort — and it is rarely the part that determines whether anything improves for a patient.

What determines that is a chain of unglamorous decisions. Whether the problem was one clinicians actually have. Whether the outcome was defined in a way that means what everyone assumes it means. Whether the data available at prediction time is the data the model was fitted to. Whether the flag reaches someone with the authority and the capacity to act. Whether the person who receives it understands what an 18% probability does and does not tell them. Whether anybody is still checking, two years later, that the numbers still mean what they meant at launch.

None of that is about artificial intelligence. It is about building a clinical service that happens to have a statistical model inside it.

The case for doing this work remains strong. Health systems everywhere are managing rising chronic disease burden with constrained workforces. Attention is the scarcest resource in medicine, and directing it towards the patients most likely to benefit is a genuinely valuable thing to be able to do. Where risk prediction has been done well — kidney failure risk influencing referral timing, cardiovascular risk equations shaping preventive treatment — it has changed practice for the better and it has done so quietly.

The case for caution is equally strong, and the evidence for it is in the published record rather than in speculation. A commercial algorithm applied to 200 million people annually systematically underidentified Black patients because it predicted cost instead of need. A

sepsis model deployed at hundreds of hospitals achieved an AUC of 0.63 on independent evaluation and missed two-thirds of cases. Roughly nine in ten drug interaction alerts are dismissed unread. These are not exotic edge cases. They are what happens when competent people build systems without asking the questions in Chapter 20.

The distance between those two sets of outcomes is not technical sophistication. Both the successes and the failures involved capable teams. The difference is in whether the surrounding work was done: the outcome validated, the model tested in the population where it would run, the alert designed for the ward it would fire on, the performance monitored after the launch party.

So the future of healthcare AI is not primarily a story about better algorithms, though the algorithms will get better. It is a story about whether health systems develop the organisational capability to translate data into timely, useful, responsible insight — and to keep doing it, unglamorously, for years, while the population shifts and the guidelines change and the original team moves on.

That capability is buildable. It requires clinicians who engage rather than dismiss, data scientists who understand that a model is a clinical intervention, executives who fund monitoring as readily as they fund development, and governance that is willing to switch things off. None of that is exotic. All of it is hard, mainly because it is slow and the incentives reward announcements.

Keep the patient at the centre. Keep the clinician in the loop. Keep measuring after the launch.

And hold on to the principle that everything else in this book is downstream of:

AI should augment clinical intelligence — not replace clinical responsibility.

References and Further Reading

The sources below were verified at the time of writing. Readers should consult the current version of each, particularly for regulatory and standards documents, which are revised frequently.

Reporting standards and methodological guidance

Collins GS, Moons KGM, Dhiman P, et al. **TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods.** *BMJ.* 2024;385:e078378. doi: 10.1136/bmj-2023-078378. Checklist available at https://www.tripod-statement.org

Moons KGM, Damen JAA, Kaul T, et al. **PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods.** *BMJ.* 2025;388:e082505. doi:10.1136/bmj-2024-082505

Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. **Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension.** *Nature Medicine.* 2020. doi:10.1038/s41591-020-1034-x

Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. **Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension.** *Nature Medicine.* 2020. doi: 10.1038/s41591-020-1037-7

Vasey B, Nagendran M, Campbell B, et al. **Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI.** *Nature Medicine.* 2022;28(5):924–933. doi:10.1038/s41591-022-01772-9

Lekadir K, Frangi AF, Porras AR, et al., for the FUTURE-AI Consortium. **FUTURE-AI: international consensus guideline for trustworthy and**

deployable artificial intelligence in healthcare. *BMJ.* 2025;388:e081554. doi:10.1136/bmj-2024-081554

Riley RD, Ensor J, Snell KIE, et al. **Calculating the sample size required for developing a clinical prediction model.** *BMJ.* 2020. PMID: 32188600.

Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. **Calibration: the Achilles heel of predictive analytics.** *BMC Medicine.* 2019. doi:10.1186/s12916-019-1466-7

Validation, bias and real-world performance

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. **Dissecting racial bias in an algorithm used to manage the health of populations.** *Science.* 2019;366:447–453. doi:10.1126/science.aax2342

Wong A, Otles E, Donnelly JP, et al. **External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients.** *JAMA Internal Medicine.* 2021;181(8):1065–1070. doi:10.1001/jamainternmed.2021.2626

Finlayson SG, Subbaswamy A, Singh K, et al. **The clinician and dataset shift in artificial intelligence.** *New England Journal of Medicine.* 2021. doi:10.1056/NEJMc2104626

Singh R, Bapna M, Diab AR, Ruiz ES, Lotter W. **How AI is used in FDA-authorized medical devices: a taxonomy across 1,016 authorizations.** *npj Digital Medicine.* 2025;8:388. doi:10.1038/s41746-025-01800-1

Felisberto M, dos Santos Lima G, Celuppi IC, et al. **Override rate of drug-drug interaction alerts in clinical decision support systems: a brief systematic review and meta-analysis.** *Health Informatics Journal.* 2024;30(2). doi:10.1177/14604582241263242

Clinical risk models

Khan SS, Coresh J, Pencina MJ, et al., for the Chronic Kidney Disease Prognosis Consortium and the American Heart Association Cardiovascular-Kidney-Metabolic Science Advisory Group. **Development and validation of the American Heart Association's PREVENT equations.**

Circulation. 2024;149(6):430–449. doi:10.1161/CIRCULATIONAHA.123.067626

Tangri N, Stevens LA, Griffith J, et al. **A predictive model for progression of chronic kidney disease to kidney failure.** JAMA. 2011. PMID: 21482743.

Singer M, Deutschman CS, Seymour CW, et al. **The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3).** JAMA. 2016;315(8):801–810. doi:10.1001/jama.2016.0287

Gulshan V, Peng L, Coram M, et al. **Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs.** JAMA. 2016;316:2402–2410. doi:10.1001/jama.2016.17216

Implementation, transparency and privacy

Sendak MP, Gao M, Brajer N, Balu S. **Presenting machine learning model information to clinical end users with model facts labels.** *npj Digital Medicine.* 2020. doi:10.1038/s41746-020-0253-3

Rieke N, Hancox J, Li W, et al. **The future of digital health with federated learning.** *npj Digital Medicine.* 2020;3:119. doi:10.1038/s41746-020-00323-1

Topol EJ. **High-performance medicine: the convergence of human and artificial intelligence.** *Nature Medicine.* 2019. doi:10.1038/s41591-018-0300-7

Policy, regulation and standards

World Health Organization. **Ethics and governance of artificial intelligence for health: WHO guidance.** Geneva: WHO; 2021.

World Health Organization. **Ethics and governance of artificial intelligence for health: guidance on large multi-modal models.** Geneva: WHO; 2024. Available at: https://iris.who.int/handle/10665/375579

National Institute of Standards and Technology. **Artificial Intelligence Risk Management Framework (AI RMF 1.0).** NIST AI 100-1. Gaithersburg, MD: NIST; 2023.

European Parliament and Council. **Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).** In force 1 August 2024.

Office of the National Coordinator for Health Information Technology. **Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing (HTI-1) final rule** — Decision Support Interventions certification criterion. Washington, DC: ONC; 2024.

US Food and Drug Administration. **Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations.** Draft guidance. Silver Spring, MD: FDA; 2025.

US Food and Drug Administration. **Artificial Intelligence-Enabled Medical Devices** (searchable list of authorised devices). Updated periodically.

Coalition for Health AI. **Applied model card and assurance lab certification frameworks.** Available at: https://www.chai.org

International Organization for Standardization. **ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system.** Geneva: ISO; 2023.

Organisations worth following

- World Health Organization — Digital Health and Innovation
- US Food and Drug Administration — Digital Health Center of Excellence
- European Medicines Agency and the European Commission's AI Act implementation resources
- UK Medicines and Healthcare products Regulatory Agency — Software and AI as a Medical Device programme

- National Institute for Health and Care Excellence — Evidence Standards Framework for Digital Health Technologies
- National Institute of Standards and Technology — AI Risk Management
- Coalition for Health AI
- The EQUATOR Network — reporting guideline repository
- Observational Health Data Sciences and Informatics (OHDSI) — OMOP common data model
- HL7 International — FHIR, SMART on FHIR and CDS Hooks specifications

About the Author

Emmanuel Agbeko Enyo writes on artificial intelligence, predictive analytics and digital health, with a particular focus on how healthcare organisations can adopt AI in ways that are clinically useful, equitable and safe.

[Author biography to be completed. This section is reserved for professional background, qualifications, current role, affiliations, research interests, and publication history. Placeholder text has been used rather than unverified detail.]

Contact and further information: *[to be added]*

Publication Information

Title: AI for Early Disease Risk Detection and Clinical Decision Support

Subtitle: A Practical Framework for Scalable and Responsible Healthcare AI

Author: Emmanuel Agbeko Enyo

Published: 2026

Subject categories: Healthcare Technology · Artificial Intelligence · Digital Health · Clinical Decision Support · Health Informatics · Academic and Professional Reference

Intended readership: Physicians and clinical researchers; nurses and allied health professionals; health data analysts and data scientists; AI and machine learning practitioners; healthcare administrators and technology leaders; digital health and health informatics professionals; policymakers and health-system decision-makers; graduate and postgraduate students in healthcare AI, health informatics and public health.

Edition: First edition

This work is educational and informational. It does not constitute medical, legal or regulatory advice. See the full disclaimer at the front of this book.