

Chapter 11 / Capítulo 11



Primum non nocere: Software engineering
for medicine
ISBN: 978-9915-9876-8-2
DOI: 10.62486/978-9915-9876-8-2.Ch11
Pages: 183-197

© 2026 The authors. This is an open access
book chapter distributed under the terms of
the Creative Commons Attribution (CC BY)
4.0 License.

Chapter 11. Model Development and Validation

Annier Jesús Fajardo Quesada  

Eduardo Antonio Hernández González 

René Herrero Pacheco 

Rolando Javier Álvarez Pérez 

1 Granma University of Medical Sciences. Granma, Cuba.

2 University of Medical Sciences of Pinar del Río. Pinar del Río, Cuba.

“A model that works on the data it was trained on is not a model that works; it is a model that has proven it can *remember*. The true test of its worth lies not in the past, but in the future: in its ability to correctly predict cases it has never seen before, in patients who were not in its development sample, in hospitals where it has never been tested.”

The history of predictive model development is full of examples of algorithms that reported excellent metrics in papers but failed spectacularly in real-world practice. This is not because the researchers were dishonest but because they confused two radically different concepts: performance on training data and performance in the real world. A model can achieve 99% accuracy on the dataset where it was developed and is completely useless when faced with new patients. This gap between the lab and reality is among the greatest challenges of artificial intelligence in medicine, and it can be addressed only through rigorous validation.

Model validation is not a formality carried out at the end of the project but a continuous process that must accompany the entire development. It begins by dividing the data into training and test sets, continues with cross-validation to optimize hyperparameters, and culminates in external validation on independent populations. Each of these steps answers a different question: Can the model learn patterns from the data? Is its performance stable? Does it generalize to new contexts? Ignoring any of these questions is similar to building on sand.

In this chapter, we explore the different levels of validation for the predictive models. We will begin with the fundamental distinction between internal and external validation and why the former cannot replace the latter. We will continue with the concept of dataset shift, one of the main causes of failure in model deployment. We address the problem of generalization with real-world examples of algorithms that worked in one hospital and failed in another. Finally, we introduce clinical MLOps, the set of practices that enable the maintenance of models in production, monitoring their performance, and updating them when necessary. This is not an exhaustive technical guide but rather a reflection on what it truly means for a model to be reliable.

11.1 INTERNAL VS. EXTERNAL VALIDATION

A model that shines in your lab but fades in the real world is not artificial intelligence; it is a statistical mirage.

There is a silent hierarchy in the validation of predictive models that separates promise from reality, academic research from truly useful medicine. At its base lies internal validation, which is seemingly robust but potentially misleading. At its peak, almost inaccessible to many projects, stands for external validation, the only guarantee that an algorithm deserves clinical trust. Between the two, intermediate steps that few manage to complete. This pyramid is not a methodological whim; it is the accumulation of decades of bitter experience where brilliant algorithms failed spectacularly upon hitting the ground in real hospitals. For the developer who aspires to build tools that truly help patients, understanding this hierarchy is as fundamental as knowing neural network architectures. For clinicians evaluating whether to trust a new digital tool, demanding adherence is an exercise in professional responsibility.

The distinction between internal and external validation transcends statistical techniques to delve into the epistemology of computational medicine. What does it truly mean for an algorithm to “work”? That it correctly predicts on data from the same hospital, same equipment, same period, and same physicians who generated its training data? Or that it maintains its utility when it crosses institutional, temporal, and population boundaries? The answer determines whether we are building toy systems for academic publications or trustworthy medical infrastructure. This chapter explores each level of the validation pyramid, its promises, its limitations, and why skipping steps mean taking risks that can result in harm to patients.

11.1.1 INTERNAL VALIDATION: WHEN DATA LIES KINDLY

Internal validation constitutes the first mandatory filter in the development of any predictive model. Its logic is seemingly simple: we divide the available data into two portions, one to train the model and the other to evaluate it. If the model predicts well on data it did not see during training, we have demonstrated that it has not simply memorized the examples but has extracted generalizable patterns. However, this “generalization” is profoundly limited: it only guarantees that it works on data drawn from the same distribution as the training data. It is like proving that

a student has learned the lesson, not that they will be able to solve new problems in different situations.

Internal validation techniques have evolved to become more sophisticated and less naive. The simple training–test split, where we randomly set aside 20–30% of the data for evaluation, is the bare minimum. *k*-fold cross-validation increases rigor: we divide the data into *k* subsets, train them on *k*–1, and evaluate them on the remaining subset; this process is repeated *k* times so that each subset serves as a test set once. This provides more stable performance estimates and allows us to use all the data for training without sacrificing evaluation. Variants such as stratified cross-validation preserve the class proportions in each fold, which is crucial when outcomes are rare. Nested cross-validation adds an additional level for hyperparameter selection, preventing us from optimizing the model by indirectly using the test data.

However, all these sophisticated techniques share a fatal limitation: all the data come from the same universe. If we train on X-rays from the Central University Hospital and evaluate them using X-rays from the same hospital, which are taken with the same equipment and interpreted by the same radiologists, during the same period, we can measure the model's ability to adapt to the peculiarities of that specific context. It may have learned that a certain image contrast indicates pneumonia, when in reality it indicates that the technician at that hospital adjusts the equipment in a certain way. Internal validation, however technically rigorous it may be, is blind to these contextual biases.

Temporal validation represents an intermediate step that introduces an additional dimension of reality. It consists of training the model with data from a historical period and evaluating it with data from a later period. For example, patients were trained from 2015–2019, and patients were tested from 2020–2021. This approach is more demanding because it introduces the possibility of temporal changes, such as the evolution of clinical practice, new diagnostic protocols, changes in the patient population, or even disruptive events such as a pandemic. If the model maintains its performance over time, we gain confidence in its robustness. However, it remains valid within the same center, with the same teams and the same institutional biases. A model that survives temporal validation may still completely fail when transferred to another hospital.

11.1.2 EXTERNAL VALIDATION: CROSSING BOUNDARIES THAT MATTER

External validation is the gold standard for separating research toys from serious clinical tools. It involves evaluating the model on data from completely different centers, different institutions, different patient populations, different imaging equipment, different clinical protocols, and different medical practices. If the model maintains its performance in this foreign territory, we have genuine evidence that it has learned universal patterns of the disease, not local artifacts of the training context. External validation is the only way to prove that an algorithm generalizes in the medically relevant sense: that it can travel and remain useful.

The difference between performance in internal and external validation can be dramatic and devastating. The medical literature documents cases in which the areas under the curve of 0.90 or higher during development decreased to 0.60 or lower in the external validation. This drop is not an anomaly; it is the norm when models capture patterns specific to the training center. A COVID-19 detection algorithm for chest X-rays, trained at a hospital where critically ill patients were positioned a certain way on the stretcher, learned to detect the patient's orientation rather than viral pneumonia. It worked perfectly at that hospital but failed completely at others. Only external validation would have revealed this fragility before a potentially dangerous deployment.

Planning for external validation must begin before the first data point is collected. Ideally, a project sets aside a portion of the data for external validation—sourced from different centers—from the very start and keeps that “sanctuary” untouched until the end of development. This discipline avoids the temptation to “tap into” that data when performance in internal validation is mediocre or to adjust the model so that it performs better on the test dataset. It also requires establishing formal collaborations with other centers from the project's inception, not as an afterthought when reviewers demand external validation to accept the publication.

Geographically external validation, where the model is tested in different regions, adds another layer of rigor. An algorithm trained on data from Europe or North America may encounter populations with different disease prevalence rates, different age distributions, different ethnic groups with relevant genetic variations, and different lifestyles that alter clinical presentation. If the model survives, external validation across multiple continents provides evidence of robust generalizability. If it fails, it reveals important limitations that must be documented: the algorithm

may be useful in certain contexts but not in others—valuable information for its selective implementation.

Prospective external validation, in which the model is applied to new patients in real time and its predictions are compared with actual outcomes, represents the highest level of evidence. Unlike retrospective validation using historical data from other centers, prospective validation eliminates the risk that the model may have inadvertently been biased by characteristics of the validation data. It is the equivalent of a clinical trial in evidence-based medicine: the definitive test of utility. However, this approach is also the most expensive and time-consuming and requires clinical integration infrastructure, specific ethical approval, and patience to accumulate enough cases. Not all projects can reach this level, but all should aim for some form of external validation before considering their model ready for clinical use.

11.1.3 FROM THE PYRAMID TO PRACTICE: IMPLEMENTING RIGOROUS VALIDATION

For the developer of clinical algorithms, understanding this hierarchy transforms project planning from its very foundations. It is not enough to construct a brilliant model on local data; one must deliberately design the validation architecture. This includes decisions about which portion of the data to set aside for which level of validation, which centers to collaborate with, and how to ensure that the external data are truly independent (not just formally from another institution but collected by the same researchers using the same protocols). It also includes comprehensive documentation of the training characteristics: age, sex, comorbidities, equipment used, time period, and disease prevalence. This documentation allows others to assess whether their population is comparable and to expect similar performance or whether there are differences that suggest caution.

Transparency in validation reporting is an ethical and scientific imperative. Too many papers publish impressive internal validation metrics without clearly warning of their limitations, or mention “validation in an independent cohort” when it comes from the same hospital just six months later. The TRIPOD (Transparent Reporting of a Multivariate Prediction Model for Individual Prognosis or Diagnosis) guidelines provide frameworks for properly reporting the development and validation of predictive models, and their TRIPOD-AI extension adapts these recommendations to artificial intelligence. Following these guidelines is not bureaucracy; it is a service to the scientific community and to future patients who will be affected by decisions based on these models.

For evaluating clinicians, this hierarchy provides a framework for systematic scrutiny. When faced with any algorithm, the following questions should be answered: What level of validation has it achieved? Is it only internal? Temporal? External—across how many centers and how similar are those centers to my practice? The performance metrics must be interpreted in light of this validation: an AUC of 0.85 with multicenter external validation is more impressive than an AUC of 0.95 with only internal validation. Even with robust external validation, clinicians must remain vigilant: they should implement performance monitoring in their own practice, comparing model predictions with actual outcomes, because there can always be unanticipated differences between their context and the validation settings.

The relationship between validation levels is not one of substitution but of necessary complementarity. Internal validation is indispensable for iterative development: fine-tuning architectures, selecting features, and optimizing hyperparameters. Without it, we cannot construct models. External validation is indispensable for implementation: to demonstrate that the model serves a purpose beyond publishing articles. Without it, we should not use models. Both are part of a continuum that should be standard in serious clinical artificial intelligence. The fact that this is unfortunately not yet the case reflects the immaturity of the field, where the pressure to publish quickly and the lack of rigorous regulation allow insufficiently validated models to influence clinical decisions.

The future of validation lies in standardization, transparency, and continuity. Public model repositories, such as ModelHub or specific medical image repositories, allow researchers worldwide to evaluate published models on their own data and accumulate evidence of postpublication external validation. Continuous monitoring platforms in production, which track the performance of deployed models and alert when deviations (dataset shift) are detected, extend validation to real time. Collaborative validation networks, where multiple centers agree on common protocols to evaluate candidate models, distribute the workload and increase the generalizability of the evidence generated.

However, no infrastructure, guidelines, or platform can replace critical human judgment. A model validated externally in ten hospitals may fail in the eleventh if there is some unaccounted-for distinguishing feature: an unusual coding practice, an immigrant population with different genetics, and a recent change in equipment. Validation is not a one-time event but a continuous process. Every new patient, every new prediction, is an opportunity to assess whether the model remains reliable. In addition, when it ceases to be so, we must be prepared to update or withdraw it. Medicine based on artificial intelligence is no different from traditional evidence-based medicine: both require intellectual humility, healthy skepticism, and a willingness to change when the facts demand it. The difference is that algorithms, unlike drugs, can degrade silently as the world around them changes, without anyone noticing until the damage has already been done.

11.2 DATASET SHIFT: WHY YOUR MODEL FAILS AT THE NEXT HOSPITAL

“A model trained in the past is a guest who has not been invited to the present; sooner or later, the present catches on.”

A scene that has become painfully familiar in clinical informatics departments around the world. A team of researchers published a sepsis prediction algorithm with an AUC of 0.94 in a high-impact journal, which was validated at three teaching hospitals. The results generate excitement in the medical community. A community hospital, impressed by the evidence, decides to implement the system. In the first few months, the technical integration runs smoothly: alerts appear on the monitors, the doctors review them, and the workflow adapts. However, after six months, a nurse in the intensive care unit commented at a patient safety meeting that “that the sepsis algorithm is acting strange.” Retrospective records confirm her suspicions: patients labeled as high-risk who progress without complications and low-risk patients who develop septic shock. The performance, which is now measured against real external validation criteria, has decreased from the published 0.94 to a barely acceptable 0.71. The model has not changed; the world around it has. This phenomenon, which accounts for more failures in the implementation of medical artificial intelligence than any other, has a technical name: dataset shift.

Data shift, or distribution shift, is the statistical manifestation of an uncomfortable truth that we developers prefer to ignore: predictive models are not universal artifacts such as a physical equation but historical creatures anchored in the time and place of their creation. When the statistical properties of the data change between training and deployment, the model becomes a stranger speaking an obsolete language, interpreting signals that no longer mean what they once did, and ignoring new signals it never learned to recognize. Understanding the mechanisms of dataset shift—its types, causes, and above all, warning signs—is an essential skill for any team aspiring to build medical systems that work not only in the laboratory but also in patients’ real lives.

11.2.1 THE THREE FACES OF CHANGE: COVARIATE, PRIOR, AND CONCEPT SHIFT

Covariate shift is perhaps the most intuitive type of dataset shift. It occurs when the distribution of the input variables changes, but the fundamental relationship between those variables and the outcome they predict remains constant. An image of a sepsis detection model trained on hospitalized adults, where a heart rate of 120 beats per minute is considered significant tachycardia, is used. When the same model is applied to a pediatric unit, many healthy children have rates of 130–140, simply because of their different physiologies. The model, which applies the thresholds learned from adults, generates massive false alarms. The relationship between tachycardia and sepsis remains valid (septic children also experience increased heart rates), but what constitutes “tachycardia” has changed. The model has not learned this contextual invariance; it has learned absolute numbers that no longer apply.

This type of shift is particularly insidious because the model may continue to have seemingly reasonable metrics in terms of globally calculated sensitivity and specificity while systematically failing in specific subgroups. A cardiovascular risk prediction model trained on a population with a high incidence of obesity may underestimate risk in lean populations because it has learned associations between variables that depend on metabolic context. Detection requires stratified analysis: comparing variable distributions between training and deployment and assessing whether performance remains consistent across different value ranges.

The prior probability shift, or change in prevalence, represents the second fundamental type.

This occurs when the proportion of positive and negative cases changes between the training and application contexts, drastically altering the clinical significance of the predictions. A model trained to detect breast cancer in population-based screening, where the prevalence is 0.5%, may perform brilliantly in that context, with modest but acceptable positive predictive values. If the same model is applied in a high-risk clinic where the prevalence is 15%, without recalibration, it will generate an avalanche of false positives that will overwhelm the diagnostic system. The technical sensitivity and specificity may remain intact, but the positive predictive value will plummet, making each alert from the model statistically less credible.

This phenomenon explains why screening models rarely work directly in high-risk settings without adjustments. Probability recalibration, using techniques such as Platt scaling or isotonic regression on local data, can restore the model's utility. However, this recalibration requires labeled data from the new setting, which can be difficult to obtain quickly. Proactive planning includes designing models that report not only binary predictions but also well-calibrated probabilities and establishing automatic or semiautomatic recalibration procedures as part of deployment.

This conceptual shift is the most profound and dangerous of the three. It occurs when the relationship between predictors and the outcome itself changes, fundamentally invalidating the model's logic. The COVID-19 pandemic provided dramatic examples: pneumonia mortality prediction models developed in 2019 failed spectacularly in 2020,⁽¹⁾ not only because the virus was new but also because the relationships among age, comorbidities, and outcomes differed. Treatments have evolved rapidly, changing the prognosis for similar conditions. ICU admission criteria changed, altering the severity of patients across different levels of care. The medical world changed more in months than it did in years, and models trained in the past became not only inaccurate but also potentially dangerous if used to guide decisions.

Detecting a concept shift is particularly challenging because it is not revealed simply by comparing variable distributions. It requires monitoring of actual predictive performance: continuously comparing predictions with observed outcomes. When the model systematically overestimates or underestimates risk in specific groups, when the calibration curves persistently deviate from the diagonal and when the performance metrics degrade monotonically over time, we are faced with evidence of a concept shift. The answer is not recalibration but rather retraining with new data or, in extreme cases, removing the model until it can be rebuilt with updated logic.

11.2.2 THE ORIGINS OF CHANGE: WHY THE WORLD REFUSES TO STAY THE SAME

The causes of dataset shift are multiple, intertwined, and often invisible until the damage has already occurred. Demographic differences between populations are the most obvious but not the only ones. Age, sex, ethnicity, socioeconomic status, and access to prior care all modify disease presentation and the distribution of normal values. A model trained on an insured urban population may fail in an uninsured rural population, not because of a technical flaw but because the epidemiology is different, the stages of presentation are different, and the untreated comorbidities are different.

Differences in clinical practice operate at subtle but crucial levels. Referral criteria determine who arrives at the hospital: A facility with strict triage in primary care will see more carefully selected patients than one with direct access. Treatment protocols alter outcomes: hospitals with an aggressive anticoagulation policy have different rates of embolism than those with a conservative policy do, altering the relationship between risk factors and events. Thresholds for performing diagnostic tests introduce detection biases: if troponin tests are performed only on patients with typical chest pain, the model learns that atypical chest pain implies low risk—a relationship that may not hold where broader screening is practiced.

Technical differences in data acquisition are particularly relevant for imaging and signal models. Equipment from different manufacturers, of different generations, or with different calibrations produces images that, although visually similar to the human eye, have distinct statistical characteristics that algorithms capture and erroneously associate with pathology. A computer vision model trained on images from a specific manufacturer may “learn” artifacts from that equipment as signs of disease, failing catastrophically with images from another manufacturer where those artifacts are absent or different.

Temporal changes affect even models that remain in the same hospital where they were trained. Clinical guidelines evolve, modifying diagnostic and treatment practices. New drugs alter the natural history of diseases. Epidemics and pandemics alter the prevalence and presentation of conditions. Changes in information systems, seemingly minor, can alter how data are recorded.

A model is, at best, a snapshot of a reality in motion; at worst, a snapshot of a reality that no longer exists.

For the developer, the dataset shift is not a remote contingency but a certainty that must be accounted for from the design phase. No model will remain indefinitely without degradation. The question is not whether it will happen but when, in what form, and how to detect it before it causes harm. This requires a continuous monitoring infrastructure, a comparison of the distributions of the input variables between training and production, an evaluation of the predictive performance in real time, and the establishment of alert thresholds for degradation. It also requires exhaustive documentation of training characteristics—not just size and metrics but also demographic distributions, clinical protocols, equipment used, and time period—everything that allows others to assess transferability.

11.2.3 DETECTING THE SHIFT: SURVEILLANCE SYSTEMS FOR MODELS IN PRODUCTION

Early detection of dataset shifts requires a monitoring architecture that goes beyond global performance metrics. Comparisons of distributions of predictor variables between the training dataset and current data in production using statistics such as the Kullback–Leibler divergence or Kolmogorov–Smirnov tests can reveal covariate shift before it degrades performance. Monitoring probability calibration, using monthly updated calibration plots, is used to detect prior probability shifts. A performance analysis stratified by relevant subgroups (age, sex, hospital unit, and shift) revealed degradations masked by stable overall averages.

Anomaly detection techniques in the feature space can identify cases that the model never encountered during training, signaling potential concept shift. When a patient presents a combination of variables that falls outside the training distribution, the model is operating in an uncharted territory, and its predictions should be treated with extreme caution or automatically rejected. These “prediction rejection” systems are essential safeguards in high-risk applications.

Prospective continuous validation, where a fraction of the model's predictions are systematically compared with actual outcomes, provides definitive evidence of sustained performance. Unlike variable monitoring, which is an indirect proxy, continuous validation directly measures what matters: is the model still predicting correctly? This practice, which is standard in the tech industry for recommendation and systems, is still in its infancy in medicine but should be mandatory for models that guide important clinical decisions.

11.2.4 ADAPTING TO CHANGE: MITIGATION AND SURVIVAL STRATEGIES

Mitigating dataset shifts requires a technical and organizational arsenal. Periodically updating the model—retraining it with recent data that incorporate the changes that have occurred—is the most direct strategy. However, this update is not trivial: it requires new quality labels, validation of the updated model's performance, and regulatory processes if the model is a certified medical device. The frequency of updates depends on the rate of change in the domain: models for stable conditions can be updated annually; models for rapidly evolving fields, such as oncology with new drugs, may require updates every six months or even quarterly.

Calibration adjustment allows models to be adapted to new prevalences without full retraining. If only the proportion of positive cases has changed, techniques such as Platt scaling or isotonic regression can restore probability calibration with relatively few labeled data points from the new context. This approach is faster and more cost-effective than full retraining, although it does not address covariate shift or concept shift.

Designing robust models from the outset, using invariant or adversarial learning techniques, can reduce susceptibility to certain types of shifts. If we know that age affects the normal values of a variable, we design the model to be invariant to age or to explicitly incorporate age as a moderating covariate. If we anticipate variation between imaging systems, we train the model using data from multiple manufacturers to force it to learn genuine pathological signals rather than specific artifacts. This design robustness comes at the cost of increased complexity and requires more diverse training data, but it pays dividends in terms of the model's longevity.

For the implementing clinician, dataset shift is a reality with which they must live. When faced with an externally developed model, they must demand local validation in their population before placing full trust in it, even if it involves a small sample that at least calibrates expectations. Once implemented, they must maintain professional skepticism: predictions that do not align

with clinical judgment deserve attention, not automatic dismissal or blind acceptance. When the model and the physician disagree, both may be wrong, but the model lacks the ability to recognize its error; the physician does.

This also has implications for the scientific evidence we generate. Clinical trials of algorithms should be designed to detect heterogeneity of effects across centers, not just to average overall results. They should include sufficiently long follow-up periods to capture temporal degradation. They should be transparent about the characteristics of participating centers, allowing future implementers to judge transferability. Evidence regarding an algorithm's efficacy is a snapshot, not a certificate of perpetual validity; it must be dated and contextualized as such.

11.3 THE PROBLEM OF GENERALIZATION

Generalization is the unspoken promise we make every time we train an algorithm. We show it a thousand cases of a disease, and we expect it to recognize the thousand and first, the two thousandth, and the case that arrives tomorrow from a different hospital with a different team and a different history. It is the difference between memorizing the answers to an exam and understanding the subject matter and between imitating the form and capturing the essence. Unfortunately, most medical AI projects meet this demand. The history of this field is written with the skeletons of algorithms that published dazzling metrics in *Nature* or *The Lancet*, only to crumble when others tried to use them in the real world. These are not anomalies or frauds; they are lessons that the field is still learning to digest. This chapter examines real-world examples of generalization failure, unravels their mechanisms, and explores how we can build systems that not only work in the laboratory but are also trustworthy in patients' lives.

11.3.1 WHEN THE MODEL LEARNS THE HOSPITAL, NOT THE DISEASE

One of the most frequently cited—and therefore most instructive—cases is that of the pneumonia detection algorithm developed by researchers at Mount Sinai Hospital and the National Institutes of Health (NIH). Published in 2017 in *PLOS Medicine*, the model achieved an area under the curve (AUC) of 0.93 in the joint internal validation, a figure suggesting an almost-supernatural ability to identify the disease. However, when external researchers tested it on data from the Indiana University Network for Patient Care, the AUC plummeted to significantly lower values. The difference was not statistically significant; it was a revelation.⁽¹⁾

Forensic analysis revealed that the algorithm had learned to detect not pulmonary pneumonia but the digital fingerprints of the hospital where it was trained: specific marks from the institution's X-ray equipment—the Konica Minolta devices in the inpatient units—generated images with an inverted color scheme compared to the Fujifilm units in the emergency department—standard patient positions on the stretchers—and most embarrassingly, the ability to distinguish with 100% accuracy between portable X-rays from the inpatient units and the Mount Sinai emergency department. The model was an expert on Mount Sinai Hospital, not on pneumonia. When images from other hospitals lacked these institutional signatures, the algorithm was blind.

This failure illustrates a fundamental mechanism: deep learning models are correlation detectors without causal understanding. They do not know what pneumonia is; they know which pixel patterns cooccur with the diagnosis in their training data. If those patterns include specific technical artifacts, differences in equipment manufacturers, positioning protocols, or even the more frequent presence of chest drains at certain centers, the model will happily incorporate them. The solution is not more manual feature engineering but deliberate training design: multicenter data from the outset, adversarial domain techniques that penalize the model for using center-specific information, and rigorous external validation that detects these dependencies before deployment.

11.3.2 THE FRAGILITY OF ARTIFICIAL SENSES: IMAGES, TEXT, AND SIGNALS

Medical image processing algorithms are particularly vulnerable to generalization failures because of their dependence on acquisition characteristics. Diabetic retinopathy detection systems provide a telling example. Multiple studies have shown that models trained on high-quality images captured with specialized cameras under optimal conditions failed spectacularly

when confronted with images from portable cameras in resource-limited settings. Differences in lighting, resolution, field of view, and capture angle were sufficient to cause identical lesions to go undetected.

This problem has profound implications for the promise of democratizing eye care through artificial intelligence. If models work only with high-end equipment available in urban centers of wealthy countries, they are not expanding access; they are replicating existing inequalities. The solution lies in intentional training with technological diversity: deliberately including low-quality images from older equipment under suboptimal conditions to force the model to learn pathological signals that persist through technical noise. It also requires specific validation in the intended implementation contexts, rather than assuming that performance on high-quality images will translate to other scenarios.

Clinical natural language processing faces analogous challenges of linguistic generalization. A model trained to identify heart failure in notes from one hospital may fail in another simply because doctors use different vocabularies. A facility where “exertional dyspnea” is the standard term versus one where “shortness of breath when walking” predominates; one where abbreviations are common versus one where words are written out in full; one where the style is telegraphic versus one where it is narrative. The model learns statistical associations between words and diagnoses, but these associations are institution-specific dialects.

Generalization in medical NLP requires normalization and robustness strategies. Contextualized embedding techniques that capture meaning beyond the exact form of words. Training is performed with data from multiple institutions to expose the model to linguistic variety. Attention mechanisms that identify which text fragments are truly informative versus which are institutional noise. Validation at centers with different documentation practices was performed to verify that the model understood the clinical content, not just the local style.

11.3.3 THE LITMUS TEST: COVID-19 AND GENERALIZATION DURING THE CRISIS

The COVID-19 pandemic provided the most instructive large-scale experiment on generalization in the history of medical artificial intelligence. In early 2020, dozens of research teams developed predictive models for diagnosis, prognosis, and resource management. The urgency justified the rush; the need for tools—any tools—seemed to take precedence over methodological rigor. A subsequent systematic review of 232 published models revealed that more than 70% had a high risk of bias and that most were never externally validated. Of those that were, many showed massive performance degradation.⁽²⁾

Mortality prediction models for COVID-19 illustrate this pattern. Developed using data from early outbreaks in China or Italy, where the pandemic struck specific populations with particular triage practices, these models failed when applied to other continents, other points in the pandemic, or other stages of understanding the disease. The variables that were predicted in March 2020—certain inflammatory markers and CT scan patterns—changed significantly as treatments, virus variants, and population immunity evolved. The models, anchored in the past, became obsolete oracles.

This massive failure was not the fault of artificial intelligence as a discipline but rather of its careless application. Given the rush to publish, the lack of external validation, the assumption that a model that works in Wuhan will work in New York, and the confusion between pandemic correlation and causality. The lessons are clear: in rapidly changing situations, temporal validation is as important as geographic validation; models must be designed for continuous updating, not for static deployment; and epistemic humility—recognizing what we do not know—is more valuable than algorithmic confidence.

11.3.4 BUILDING FOR THE REAL WORLD: PRINCIPLES OF ROBUST GENERALIZATION

Generalization is not a magical property that some models possess and others do not; it is the result of deliberate design decisions. The first principle is intentional diversity in training. A model trained on data from a single hospital will learn that hospital. A model trained on data from fifty hospitals across different regions, sizes, populations, and practices will be forced to learn what is common to all of them: the patterns of the disease itself. This diversity comes with logistical and financial costs, but it is an essential investment, not a luxury. Multicenter consortia, international collaborative networks, and diverse data repositories are critical infrastructures for

robust medical artificial intelligence.

The second principle focuses on causal and physiological relationships versus superficial statistical associations. Some relationships are universal: severe hypoxia can damage organs in any human. Others are context specific: the association between a genetic marker and drug response varies across populations. The incorporation of explicit medical knowledge into model design—architectures that reflect physiological understanding and variables that capture causal mechanisms—can improve generalization by anchoring the model in stable biological realities. Causal learning techniques, although still emerging, promise models that identify interventions invariant across contexts.

The third principle is technical robustness against measurement variations. Appropriate normalization of variables, data augmentation techniques that simulate plausible acquisition variations, and architectures that are inherently invariant to irrelevant transformations are needed. A vision model that learns to be insensitive to rotations, scales, and contrast changes within physiological ranges will be more robust than one that memorizes specific orientations. Data augmentation is not a trick for small datasets; it involves training in generalization through exposure to variation.

The fourth principle is that external validation should be the standard, not the exception. Every model must be tested on data that did not share the same origin as the training data: different hospitals, different populations, different teams, and different time periods. This validation must be transparent: publishing not only average metrics but also breakdowns by subgroup, analyses of specific failures, and explicit limitations. A model that performs well on average but fails in specific subgroups is not generalizable; it is potentially dangerously biased.

For the developer, these principles imply that the first model will never be the final one. This implies planning for external validation from the study design stage, establishing multicenter collaborations, and setting aside independent data. It implies humility regarding the model's limitations and transparency regarding its conditions of validity. This implies that impressive metrics are not confused during development with real clinical utility, which is demonstrated only through generalization.

For the clinician, these examples serve as a guide to professional caution. When faced with any new model, the following questions must be answered: In which populations was it validated? Are they comparable to my practice? Is there evidence of performance in different contexts? Are limitations and failures reported, not just average successes? If the answers are unsatisfactory, skepticism is a professional obligation. This is not because the models are inherently bad but because the evidence of their usefulness is insufficient. In addition, in medicine, acting without adequate evidence violates the fundamental principle of *primum non nocere*.

11.4 CLINICAL MLOPS: TRAINING, DEPLOYMENT, AND MONITORING

The history of software has taught us, through decades of costly mistakes, that writing code is merely the beginning of a much longer odyssey. Maintenance consumes the lion's share of the budget, emerging updates disrupt plans, monitoring reveals problems that tests never captured, and the orderly retirement of obsolete systems demands discipline that is rarely rewarded. This reality, now widely accepted in traditional software engineering, is only just appearing on the horizon of medical artificial intelligence. From this realization arises MLOps, a set of practices that professionalize the entire lifecycle of predictive models, from conception to retirement. In medicine, where a silent failure can result in harm to patients, these practices are not a technical option but an ethical imperative. This chapter explores how to operationalize the promise of sustainable clinical models—ones that not only start out with good metrics but also age with dignity and utility.

11.4.1 BEYOND TRAINING: THE COMPLETE LIFE CYCLE

The life cycle of a clinical model begins long before the first line of code or the first gradient descent. It begins with a rigorous definition of the clinical problem: what medical decision do we want to improve? What information will be available at the time of the decision? What outcome is predictable and relevant? A poorly defined problem here—predicting “readmission” without specifying the time window or “improvement” without standardizing how it is measured—does everything that follows. This conceptual design phase, often undervalued in the face of the appeal

of modeling, determines whether the model will have real utility or merely publishable metrics.

Identifying data sources and obtaining ethical and legal permissions are parallel and often slow steps. In regulated environments, access to clinical data requires ethics committee approval, confidentiality agreements, and technical security guarantees that can take months. Patience here is not a virtue but a necessity: shortcuts in data governance can legally invalidate the entire subsequent project. The design of the validation study, including the reservation of external data from the outset, must be completed before touching the training data. The temptation to “see what we have” before planning is natural but dangerous: it contaminates the independence of the test sets and compromises the validity of the entire validation.

The training phase, although the most visible in the scientific literature, is in MLOps and involves just one link in an iterative chain. Yes, the algorithms are explored, the hyperparameters are optimized, and the architectures are validated. However, in a mature MLOps environment, this phase is automated through continuous integration pipelines: the modeling code is versioned, experiments are tracked, and the best models are registered in central repositories along with all their metadata (hyperparameters, metrics, dependencies, and data used). Training is not a one-time event but a permanent capability, ready to be reactivated when new data or new insights demand it.

11.4.2 DEPLOYMENT AS A PROCESS, NOT AN EVENT

Deploying the model into clinical production is the moment of greatest tension. A system that works in the lab may fail for reasons that have nothing to do with the quality of the algorithm: unacceptable latency in generating predictions, incompatibility with legacy clinical information systems, user resistance to interfaces that disrupt their workflow, or simple integration errors in the data pipeline. That is why, in clinical MLOps, deployment is not a binary switch but a gradual ramp of exposure.

The canary deployment strategy, borrowed from traditional software engineering, adapts elegantly to medicine. First, the model is deployed in a test environment that replicates production but without real patients, verifying technical integration. Then, it is activated in “shadow” mode: it makes predictions in real time but without showing the predictions to clinicians, allowing a comparison of what it would have recommended with what actually happened. Next, it is exposed to a small group of volunteer users, gathering qualitative feedback on utility and usability. Finally, it is scaled to noncritical services, where the model assists but does not determine decisions, before it is considered for application in high-risk contexts. Each transition requires explicit approval gates, documentation of decisions, and the ability to immediately roll back if problems arise.

Deployment infrastructure in clinical environments presents specific challenges. Many hospitals operate legacy systems that were not designed to integrate machine learning models. The latency requirements are strict: a prediction that takes thirty seconds to arrive may be useless in an emergency. High availability is critical: if the prediction service decreases, does the clinical system have a reliable fallback? Traceability is mandatory: Each prediction must be logged with a timestamp, model version, input data, and result for subsequent auditing. MLOps platforms such as Kubeflow, MLflow, or specialized commercial solutions provide infrastructure to manage this complexity, but they require adaptation to specific healthcare workflows.

11.4.3 CONTINUOUS MONITORING: KEEPING AN EYE ON THE DEPLOYED MODEL

Monitoring is at the heart of clinical MLOps. A deployed model is not a finished product but an evolving sociotechnical system that requires multidimensional monitoring. Monitoring predictive performance is the most obvious aspect: do sensitivity and specificity maintain the levels validated initially? This requires a systematic comparison between predictions and actual outcomes, which entails an infrastructure to collect posthoc data. In some cases, this is immediate (did the patient develop sepsis within the next 24 hours?), in others it requires follow-up (did they survive at six months?), and in others, it is complex (did their quality of life improve?). Without continuous validation mechanisms, the model operates in the dark, without knowing whether it remains useful.

Monitoring the quality of the input data involves detecting dataset shifts before they degrade performance. Comparing variable distributions between training and current data using divergence statistics, hypothesis tests, or anomaly detection methods can generate early

warnings. If the average age of patients suddenly increases, if a laboratory changes equipment and the results range, or if a pandemic alters the mix of pathologies, the system must notify decisionmakers for evaluation.

Monitoring clinical use reveals whether the model is being adopted, ignored, or misused. Do physicians follow the algorithm's recommendations? Do they systematically override them in certain subgroups? Are there differences in adoption between departments or shifts? A model with excellent technical performance that clinicians do not use is a total failure; a model they use blindly without question is a potential danger. Usage data, which are collected in compliance with privacy regulations, provide insights into both perceived utility and the need for additional training.

The detection of deviations must be accompanied by response protocols. When performance falls below predefined thresholds or when significant changes in input distributions are detected, automated alerts are triggered. However, the response cannot be purely algorithmic: it requires human investigation of causes. Is it actual model degradation, or a change in data quality? Is it an evolution in clinical practice or a technical issue in the pipeline? Does it require retraining, recalibration, or temporary withdrawal? Documented runbooks, which establish procedures for different scenarios, are essential for a coordinated response between technical and clinical teams.

11.4.4 CONTROLLED EVOLUTION: UPDATES, GOVERNANCE, AND WITHDRAWAL

Model retraining is a natural response to model degradation, but it is not trivial in regulated environments. Each new version requires validation, documentation of changes, and potentially regulatory review if the model is a certified medical device. The frequency of updates depends on the rate of change in the domain: models for stable conditions can be updated annually; models for rapidly evolving areas, such as oncology with new drugs, may require semiannual or even quarterly cycles. MLOps pipelines automate technical retraining, but the decision to deploy a retrained model for production must be made by a human and documented.

Governance establishes who has authority for each decision in the lifecycle. Who approves the initial deployment? Who authorizes an update? Who decides on a withdrawal? In mature organizations, these roles are formally defined: clinical validation committees, quality managers, and patient safety teams. Documenting decisions are not bureaucracy but essential traceability: if it is later discovered that a model caused harm, we must be able to reconstruct who approved it, on the basis of what evidence, and who had the authority to stop it.

The orderly retirement of models is just as important as their deployment. When a model ceases to be useful—due to obsolescence, irreversible degradation, or because it has been replaced by a better alternative—it must be retired from production in a controlled manner. This includes notifying users, transitioning to alternative systems, archiving historical data according to retention policies, and drawing lessons learned for future projects. A model left in production, continuing to generate predictions that no one reviews, is a latent risk.

The relationship between MLOps and regulation is becoming increasingly close. IEC 62304, which we discussed in previous chapters on the medical software lifecycle, provides a framework adaptable to predictive models. The FDA's guidelines on medical artificial intelligence explicitly require postmarket monitoring processes and controlled updates. The European MDR Regulation, as interpreted for medical software, includes vigilance requirements that MLOps operationalize. The alignment between technical MLOps practices and regulatory compliance is not a coincidence but a convergence toward patient safety standards.

For the developer, clinical MLOps involve expanding skills. Expertise in deep learning architectures is no longer sufficient; an understanding of cloud or on-premise infrastructure, data pipelines, continuous integration, and systems security is needed. Close collaboration with hospital IT teams is needed, as they are often unfamiliar with the specifics of model deployment. Communication with clinicians is necessary to understand workflow needs and design useful interfaces. Compared with a laboratory researcher, the profile of the clinical MLOps engineer is hybrid and closer to that of a mission-critical software engineer.

For clinicians, MLOps provides tangible assurances of professionalism. Knowing that the model they consult is monitored, that there are processes to detect degradation and that there are channels to report issues increases legitimate trust in the tool. However, it also assigns responsibility: clinicians are the system's human sensors, often the first to notice when predictions "don't feel right." They must have mechanisms to report these intuitions, which can then be confirmed or refuted through statistical analysis. Vigilance is a shared responsibility, not

one delegated entirely to algorithms.

Ultimately, clinical MLOps is not an end in itself but a means to a higher end: maintaining models that truly help clinicians make better decisions for patients. A model deployed without monitoring is risk in disguise. A model that is never updated is planned obsolescence. A model without clear governance is institutionalized uncertainty. MLOps provides the framework to manage these risks to keep models alive, useful, and worthy of the trust we place in them. In the future, where artificial intelligence will be as ubiquitous in medicine as the stethoscope, having these processes in place will not be a competitive advantage but a condition of ethical existence.

If keeping traditional models up to date is already complex, how do we prepare for the arrival of massive-scale artificial intelligence systems, such as large language models, which were not designed for medicine but will inevitably be integrated into it?

REFERENCES

- Subasri V, Krishnan A, Kore A, Dhalla A, Pandya D, Wang B, et al. Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models. *JAMA Netw Open* [Internet]. June 2, 2025 [cited May 14, 2026];8(6):e2513685-e2513685. Available from: <https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2834882>
- Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ* [Internet]. April 7, 2020 [cited May 14, 2026];369. Available at: <https://pubmed.ncbi.nlm.nih.gov/32265220/>