

Chapter 10 / Capítulo 10

Primum non nocere: Software engineering
for medicine
ISBN: 978-9915-9876-8-2
DOI: 10.62486/978-9915-9876-8-2.Ch10
Pages: 167-182



© 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 10. Construction and Validation of Clinical Datasets

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.

“An algorithm is only as good as the data it is trained on. If the data are biased, the algorithm will learn biases. If the data are erroneous, the algorithm will learn errors. If the data reflect only a part of reality, the algorithm will be blind to everything else. That is why, before we talk about artificial intelligence, neural networks, or deep learning, we must address the fundamentals: how we build the datasets that will power these systems.”

Imagine that we want to construct an algorithm that can detect pneumonia in chest X-rays. We need thousands of images, each labeled as “pneumonia” or “not pneumonia.” However, who assigns those labels? A radiologist with twenty years of experience? A first-year resident? A medical student? What if two radiologists disagreed on an image? What is the truth? What if the image shows an ambiguous finding—one that could be pneumonia or something else? These questions, which may seem minor, are actually at the heart of the problem. The quality of the algorithm depends on the answers we provide, and patient safety depends on the quality of the algorithm.

Building clinical datasets is among the most underappreciated yet most critical aspects of developing artificial intelligence in healthcare. While technical aspects (which neural network architecture to use and which optimization algorithm) receive disproportionate attention, the dirty work of collecting, cleaning, and labeling data is often relegated to the background. Researchers compete to publish models with better metrics but rarely ask how those metrics were obtained, what the quality of the data was, or what biases might be hidden. This imbalance between attention to the model and attention to the data are dangerous because a brilliant model trained on mediocre data will produce mediocre results, and no one will notice until it is too late.

In this chapter, we explore the fundamental aspects of building and validating clinical datasets. We will begin with the clinician’s role in data annotation, a role that goes far beyond simply applying labels. We will continue with the challenge of interrater disagreement and how to measure it using the kappa coefficient. We address biases in data collection, which may be present even before labeling begins. Finally, we discuss sample size and statistical power, which are essential concepts for determining how much data we need for our model to be reliable. The goal is not to turn the reader into a machine learning expert but to equip them with the questions they should ask when faced with any clinical dataset, and the criteria for evaluating whether the data they are asked to work with are trustworthy.

10.1 WHO ANNOTATES AND HOW: THE ROLE OF THE CLINICIAN

There is a persistent fantasy in the world of medical artificial intelligence that is both appealing and dangerous. In this fantasy, a team of engineers receives hundreds of thousands of X-ray images and processes them through a crowdsourcing platform where medical students label them in a matter of weeks, and from this effort emerges a dataset considered “ground truth.” The numbers are impressive: one hundred thousand annotated cases, fifteen thousand hours of work, coverage of twelve different pathologies. With these data, a deep neural network is trained that achieves 94% accuracy in internal validation. An article is published in a high-impact journal, funding is secured for a startup, and the algorithm is prepared for deployment in hospitals. However, when it finally faces real patients, something goes wrong. The system confuses atypical pneumonias with heart diseases, ignores subtle findings that any resident would detect, and generates predictions that senior radiologists consider absurd. What happened? In the rush to scale up, quantity was confused with quality, speed with wisdom, and the opinion of those just beginning their training with the judgment of those who had been interpreting images for decades.

The reality of clinical annotation is vastly different from this industrialized fantasy. It begins with a question that seems simple but rarely is as follows: What exactly are we trying to capture? When we say we want to detect “pneumonia” on a chest X-ray, we are using a term that, in medicine, encompasses a wide range of conditions. There is typical bacterial pneumonia with its classic lobar consolidation but also viral pneumonia with diffuse interstitial infiltrates, pneumonia in immunocompromised patients who do not mount a visible inflammatory response, and evolving pneumonia that shows changing patterns. There are microbiologically confirmed pneumonias and others that are diagnoses of exclusion. Some images show unequivocal findings, while others are ambiguous even to experts. Before any student touches an image, before a single line of code is written, a group of experienced clinicians must sit down to define with surgical precision what constitutes each category, what criteria differentiate one from another, and how to handle cases

that do not fit neatly into any box.

10.1.1 ART OF DEFINING THE INDEFINABLE: PROTOCOLS THAT SAVE ALGORITHMS

The construction of a valid annotation protocol is perhaps the most underestimated task in the entire development of clinical artificial intelligence. A good protocol is not simply a list of categories; it is a living document that anticipates the complexity of real-world medical practice and provides clear guidelines for navigating that complexity. It must include precise operational definitions, representative examples of each category, explicit exclusion criteria, and rules for handling ambiguity. When two expert radiologists disagree on an image, both may be correct because the image is inherently ambiguous. A smart protocol does not force false certainty but rather establishes procedures to document that ambiguity and, when appropriate, allows for multiple labels or categories of uncertainty.

The process of developing these protocols must be iterative and collaborative. You cannot write a protocol in a two-hour meeting and expect it to work. Pilot tests with small case sets, analysis of interrater variability, refinement of definitions, and further testing are needed. Clinicians involved in this phase must be true experts, not only in their medical specialty but also in the ability to articulate their clinical reasoning. Many excellent physicians struggle to articulate why they see something in an image; their knowledge is tactile, intuitive, and developed through years of practice. The challenge is to translate that tacit knowledge into explicit rules that others can follow without losing the essence of what makes an expert an expert.

The handling of ambiguous cases is particularly critical. In real clinical practice, a physician may request additional information: How has the patient been doing? Do they have a fever? What do the blood tests show? In the context of artificial intelligence annotation, this additional information may or may not be available. If it is available, it improves the quality of the annotation but introduces a bias: the algorithm will learn to make a diagnosis on the basis of information that will not be available at the time of the actual prediction. If it is not available, the annotations may be less accurate but more realistic from a deployment standpoint. This decision is not technical but strategic and must be made consciously with a full understanding of the implications.

Quality controls during the annotation process are just as important as final checks are. It is not enough to review 10% of the cases at the end of the project; mechanisms are needed to detect deviations in real time. This may include inserting gold cases whose correct label is known, monitoring the annotation speed (too fast often indicates inattention), and statistically analyzing the label distribution (abrupt changes may indicate that an annotator has misinterpreted the instructions). When problems are detected, correction should be immediate and formative and not punitive. The goal is not to punish errors but to improve the process continuously.

10.1.2 THE EXPERTISE PYRAMID: BUILDING SUSTAINABLE ANNOTATION TEAMS

Choosing who performs the annotations is a balancing act between the ideal and the feasible. In a perfect world, every image, every record, every signal would be reviewed by the top expert available—someone with decades of experience and international recognition in their field. However, these experts are scarce and expensive and have clinical responsibilities that they cannot abandon. A senior radiologist may take fifteen minutes to analyze a complex CT scan; multiplied by thousands of cases, the project becomes economically and temporally unfeasible. The solution is not to sacrifice quality but to design annotation architectures that distribute expertise intelligently.

An effective model is the three-tier pyramid. At the base, annotators with basic medical training but no specialization (advanced medical students, specifically trained imaging technicians) handle clear and routine cases, which represent the majority of the volume. At the intermediate level, residents or junior specialists review cases in which the base-level annotators have flagged as questionable and conduct quality checks on the basis-level work. At the top, senior experts focus exclusively on the most complex cases, on resolving discrepancies among lower-level raters, and on defining and reviewing protocols. This structure allows experts' time to be used where it truly adds value, while the volume is handled by more abundant resources.

Supervision and mentoring are essential components of this pyramid. Entry-level annotators should not work in isolation; they must have access to consultations with higher-level staff, ongoing training sessions, and regular feedback on their performance. An annotator who

receives only written instructions and never interacts with an expert will develop bad habits that will become ingrained over time. In contrast, an annotator who is part of a team where they can ask questions, discuss difficult cases, and see how experts reason will continuously improve their skills. This investment in human talent development pays dividends in data quality.

Documenting who annotated what and with what level of confidence is crucial for subsequent data science. Not all labels in a dataset are equally reliable, and that's fine as long as we know it. A record indicating that a certain annotation was made by a first-year resident and marked as "uncertain" is more valuable than a label forced to be definitive. Machine learning algorithms can incorporate this uncertainty information, giving more weight to high-confidence annotations and treating uncertain annotations in a special way. Furthermore, this traceability allows for retrospective audits: if we discover that a specific annotator made systematic errors, we can identify and correct those cases without invalidating the entire dataset.

Annotator well-being is an often-overlooked variable that directly affects quality. Medical annotation is intellectually demanding and emotionally taxing work. Reviewing hundreds of images of severe pathology, reading reports on terminally ill patients, and making decisions that will affect the ability of algorithms to diagnose real people can lead to fatigue and stress. Protocols must include regular breaks, task rotation to avoid monotony, and strict limits on the number of cases per day. An exhausted annotator makes mistakes that they would not make when rested, and those mistakes multiply by the thousands in a large dataset. Investing in humane working conditions is not only ethical; it is pragmatic from the perspective of data quality.

10.1.3 BETWEEN LABORATORY PERFECTION AND THE NOISE OF REALITY: ANNOTATION CONTEXT BIASES

A fundamental paradox of annotation for medical artificial intelligence is that the very act of annotating creates an artificial context that can distort the data. When a radiologist knows they are annotating images for a research project—without the pressure of a hectic shift, without the responsibility of a diagnosis that will determine a real patient's treatment—their performance changes. They may be more meticulous, spend more time on each image, and look for findings that they might overlook in routine clinical practice. The result is annotations that represent an idealized medicine, a version of clinical practice that exists only in research laboratories. When the algorithm trained on these "superperfect" data is confronted with the reality of an overcrowded emergency department, where images are of lower quality, patients have multiple overlapping conditions, and time is scarce, the mismatch can be stark.

To mitigate this context bias, some projects have explored retrospective annotation based on real-world practice. Instead of asking doctors to annotate specifically for the project, they use the diagnoses that clinicians recorded in their daily work at the time of patient care. This approach has the virtue of capturing medicine as it is actually practiced, with all its noise and imperfections. However, it also has serious limitations: clinical diagnoses may be incomplete, influenced by administrative factors (coding for billing), or simply erroneous. A diagnosis of "pneumonia" in a medical record may reflect an initial suspicion that was later ruled out or a diagnosis of exclusion that was never confirmed. Using these data without additional validation can introduce systematic noise into the dataset.

The most promising solution is prospective annotation integrated into the clinical workflow. In this model, annotation is not a separate activity but a natural component of patient care. When a radiologist interprets an image in their usual PACS system, their structured report automatically becomes the annotation for the dataset. Artificial intelligence tools are embedded within the clinical environment, not on external platforms. This approach requires a sophisticated technological infrastructure and changes to workflows, but it produces high-quality data that reflects real-world practice. Furthermore, it enables continuous updating: the algorithm can be retrained with fresh data that reflect changes in epidemiology, new imaging techniques, or evolving diagnostic standards.

The relationship between clinicians and developers must evolve into true collaborative cocreation. Too many projects treat physicians as mere data providers, paying them by the hour for data entry without involving them in study design, interpretation of results, or implementation decisions. This dynamic is unsustainable and counterproductive. Clinicians who feel part of the team—who understand why certain data are collected and how it will be used and who have a say in the project's direction—contribute not only data but also invaluable clinical insights. They can point out biases in the design, suggest relevant variables that engineers might not

have considered, and anticipate implementation issues that are evident only from a practical perspective.

Proper recognition of clinical contributions is both ethical and practical. In academia, clinicians involved in data annotation should be listed as coauthors of the resulting papers if their intellectual contribution is substantial, not merely acknowledged in the acknowledgments section. In commercial projects, they should receive fair compensation that reflects the value of their expertise, not crowdsourcing rates. In all cases, there must be transparency regarding how the data will be used, who will have access to it, and what safeguards exist to protect patient privacy. This collaborative ethic is not only the right thing to do; it builds the trust necessary for top clinicians to want to participate in future projects.

AI-assisted annotation represents a promising evolution that does not replace clinicians but rather amplifies their productivity. In this model, a preliminary algorithm proposes labels for clear-cut cases, and the clinician focuses on reviewing and correcting, especially in cases where the algorithm has low confidence. Over time, as the algorithm improves, the proportion of cases requiring human attention decreases, allowing the volume of annotation to scale without sacrificing quality. However, this approach requires caution: if the initial algorithm has systematic biases, these can be perpetuated if the clinician tends to accept the system's suggestions. The design of these interfaces must encourage critical scrutiny, not passive acceptance.

Clinical annotation is not a necessary evil that we will eventually automate entirely but a fundamental activity of 21st-century medicine. As artificial intelligence becomes more deeply integrated into healthcare, the ability to generate high-quality labeled data becomes a critical competency for healthcare systems. Investing in training professionals in these skills, developing rigorous methodologies, and creating infrastructures that facilitate annotation is important for the future of medicine itself. With the development and execution of algorithms, neural architectures evolve, but the need to connect human medical knowledge with the potential of computing remains. In that connection, in that bridge between the clinician's mind and the bits of the dataset lies the true frontier of medical artificial intelligence.

10.2 DISAGREEMENT AMONG ANNOTATORS AND THE KAPPA COEFFICIENT

Imagine you have gathered three experienced radiologists in a room with a hundred chest X-rays. Your goal is to create the definitive dataset to train a pneumonia detection algorithm. Doctors work in silence, marking their diagnoses with the confidence that comes from experience. When they finish, you collect the annotation sheets and compare the results with anticipation. Seventy images show perfect agreement among the three: they all see pneumonia, or they all see a healthy chest. However, in thirty images, something is off. Fifteen showed an uncomfortable pattern: two radiologists agreed on the diagnosis, but the third strongly disagreed. In the other fifteen, it's total chaos: each doctor sees something different, and there's no majority to resolve the discrepancy. Your initial instinct is frustration. How is it possible that trained experts disagree so much? Did you choose the participants poorly? Are the images of poor quality? However, as you delve deeper into the discordant cases, you discover something more interesting: the images with total disagreement show atypical infiltrates—patterns that could be viral pneumonia, cardiac edema, or even old scarring. The radiologists are not incompetent; they are honest about the inherent ambiguity of medicine. Disagreement is not a flaw in the process; it's a window into the complexity your algorithm needs to understand.

This scenario illustrates an uncomfortable truth that many medical AI projects prefer to ignore: variability among expert observers is not a methodological flaw to be eliminated but a fundamental feature of clinical practice. When we force artificial agreement, when we arbitrarily choose one opinion as “the truth” and discard the others, we are not purifying our data; we are impoverishing it. We are teaching our algorithms that the world is binary and welldefined; when in reality, it is often gradual and ambiguous. However, before we can incorporate this variability constructively, we need to measure it systematically. We need to know when the raters agree because the case is clear and when they disagree because the case is inherently difficult. For this task, statistics offers us elegant tools, with the kappa coefficient being the most widely used and, when interpreted correctly, one of the most revealing.

10.2.1 BEYOND THE PERCENTAGE: WHY SIMPLE AGREEMENT IS MISLEADING

The most common mistake when evaluating agreement among annotators is to blindly rely on the simple agreement percentage. If two doctors agree on eighty out of a hundred cases, we conclude that they have 80% agreement and feel satisfied. However, this metric can be deeply misleading, especially in medicine, where categories are often unbalanced. Imagine a dataset for breast cancer detection in mammograms, where only 5% of the images show actual malignancy. Two annotators who simply marked all the images as “benign” agreed in 95% of the cases, achieving an apparently excellent agreement percentage without having made any real diagnosis. They have agreed by taking the easy way out, not because they share medical judgment. This phenomenon, known as chance agreement or prevalence agreement, makes the simple percentage virtually useless for evaluating the quality of a medical dataset.

The kappa coefficient, developed by psychologist Jacob Cohen in 1960, was created precisely to correct this distortion.⁽¹⁾ In its most intuitive form, kappa compares the observed agreement among annotators with the agreement that would be expected by chance alone, given the distribution of the categories. If the annotators agree that more than chance would explain, kappa is positive. If they agree exactly as chance would predict, the kappa is zero. If they agree that less than chance would predict (which would indicate systematic disagreement), the kappa coefficient can be negative. This adjustment for chance makes kappa a much more conservative and realistic measure than the simple percentage. A kappa of 0.70, which indicates substantial agreement, may correspond to a percentage agreement of 90% or more in unbalanced categories but only 60% in balanced categories. Interpretation requires understanding this fundamental distinction.

Conventional kappa interpretation scales, although arbitrary, provide a common language for discussing the quality of agreement. Values less than 0.20 indicate poor agreement, which is barely better than chance. Between 0.21 and 0.40, the agreement is slight, suggesting that the raters share some criteria but with considerable noise. The 0.41–0.60 range indicates moderate agreement, which is acceptable for some applications but concerning for others. Between 0.61 and 0.80, we find substantial agreement, which is generally considered adequate for most clinical purposes. Above 0.81, the agreement is nearly perfect. These thresholds are not natural laws but practical conventions, and their application must be contextual. A kappa of 0.50 may be acceptable for a population-based screening task where absolute accuracy is less critical but unacceptable for an algorithm that will guide aggressive cancer treatments. What matters is not the number itself but the conversation this number enables regarding the quality of our data.

Calculating Cohen’s kappa for two raters is conceptually simple but revealing. First, we calculate the observed proportion of agreement (P_o) simply by counting the cases where both the raters agree and dividing by the total. Afterward, we calculate the proportion of agreement expected by chance (P_e), which depends on the marginal frequencies: if rater A marks 70 positives and 30 negatives and rater B marks 60 positives and 40 negatives, chance predicts that they will agree on positives in 42% of the cases (0.70×0.60) and on negatives in 12% (0.30×0.40), for a total expected agreement of 54%. Kappa is then $(P_o - P_e)/(1 - P_e)$, which normalizes the observed agreement relative to the maximum possible agreement beyond chance. This formula proportionally penalizes agreement due to chance, making kappa zero when $P_o = P_e$ and one when the agreement is perfect.

When working with more than two raters, Fleiss’s kappa ((1)) extends this logic to multiple observers. Although the calculation is more complex, the interpretation remains intuitive: it measures how much beyond chance all the raters agree simultaneously. For ordinal categories where disagreements are not all equally serious (confusing “mild” with “moderate” is less problematic than confusing “mild” with “severe”), the weighted kappa is used. This variant assigns different weights to the various types of disagreement, better reflecting clinical reality where errors have different degrees of severity. Choosing the appropriate variant is not a minor technical detail; it determines which aspects of agreement we are measuring and, therefore, what conclusions we can draw about the quality of our data.

10.2.2 WHEN KAPPA WHISPERS: DIAGNOSING THE CAUSES OF DISAGREEMENT

A low kappa coefficient is similar to a fever in a patient: it is not the disease itself but a symptom that requires further investigation. The first cause we must consider is ambiguity in the definitions of the categories. If we ask raters to classify “pneumonia” without specifying

whether we include atypical, active, or resolved cases, each physician will apply their own clinical interpretation. One will be restrictive, requiring clear lobar consolidation; another will be broad, including subtle interstitial infiltrates. The discrepancies do not reflect incompetence but a lack of prior consensus on what we are measuring. The solution is not to dismiss the raters but to refine the protocol: consensus meetings where experts discuss borderline cases, written operational definitions with visual examples, and explicit rules for ambiguous situations. Often, we discover that what we thought was a simple category actually encompasses subtypes that we should distinguish.

The second common cause is the inherent difficulty of certain cases. Some medical images are objectively difficult to interpret, even for senior experts. A pulmonary nodule with poorly defined borders, pneumonia overlapping with cardiac edema, and a bone fracture in a suboptimal projection generate legitimate disagreement that reflects the epistemic uncertainty of the medicine itself. In these cases, forcing agreement by choosing the opinion of the majority or the most senior expert is methodologically dangerous. We teach the algorithm that there is a single truth where in reality, there is a spectrum of valid interpretations. The most sensible approach is to explicitly document these difficult cases and exclude them from training if the ambiguity is extreme or, better yet, design the algorithm so that it can express uncertainty when faced with similar situations. A model that says “I’m not sure; this case needs human review” is more valuable than one that blindly relies on an arbitrary label.

Over time, fatigue and performance degradation constitute a third major cause, especially in high-volume projects. Medical annotation is cognitively demanding work that requires sustained attention. After fifty cases, concentration wanes; after one hundred, errors become more frequent; after two hundred, the annotator may be operating on the autopilot, applying rules of thumb rather than clinical reasoning. Disagreement patterns can reveal this phenomenon: if we calculate the kappa by time blocks, we can detect whether agreement deteriorates as the day progresses. The solutions are prosaic but effective: mandatory breaks every hour, task rotation, strict limits on the number of cases per session, and interspersed quality checks that detect deviations in real time. A well-rested annotator produces higher-quality data than an exhausted annotator does, regardless of their level of expertise.

A lack of adequate training or experience represents a fourth source of variability. Not all physicians are equally prepared for all the annotation tasks. An excellent cardiologist may not be the best annotator for neuroradiology images; a first-year resident lacks the case history needed to recognize atypical patterns. When we mix annotators of different skill levels without stratification, the kappa reflects these differences in ability rather than the intrinsic difficulty of the cases. The solution lies in carefully curating the annotation team, assigning tasks according to expertise, and providing specific training before the project begins. Less experienced annotators can be valuable if their work is properly validated, but we must be transparent about their limitations in the final dataset.

Systematic biases among annotators can manifest as low Kappas with characteristic patterns. A radiologist trained at a conservative school may tend toward overdiagnosis, indicating cases as positive that others consider doubtful. Another, perhaps scarred by litigation experience, may be excessively restrictive. These biases are not random errors but consistent trends that can be detected by analyzing confusion matrices between pairs of annotators. When we identify these patterns, we have options: we can calibrate the annotators through joint discussion of discordant cases, adjust the instructions to correct identified trends, or even consciously incorporate these differences in judgment in s as a way to increase the robustness of the dataset. An algorithm trained with annotators from different “schools” of thought may generalize better than one trained with a single but potentially idiosyncratic criterion.

10.2.3 BEYOND KAPPA: INCORPORATING VARIABILITY INTO ALGORITHM DESIGN

The most important conceptual revolution in the analysis of disagreement among annotators is the shift from viewing it as a problem to accepting it as valuable information. Traditional machine learning approaches assume that there is a single “fundamental truth” for each case and that the algorithm’s goal is to learn to predict that truth. However, in medicine, this assumption is often false. When three expert radiologists give three different diagnoses for an ambiguous image, there is no single truth that two of them are ignoring; there is a clinical reality that is genuinely uncertain. Modern multiannotator learning methods incorporate this uncertainty directly into the model, producing more honest and clinically useful algorithms.

An elegant approach is to model annotator confidence. Instead of treating all the labels as equally valid, we simultaneously estimate the true (latent) label and the reliability of each annotator. If a particular radiologist shows high agreement with the consensus of their colleagues in clear-cut cases, their opinion is given more weight when they disagree in difficult cases. This approach, known as learning with multiple noisy annotators, not only improves the quality of training but also provides diagnostic information about the annotators themselves. We can identify who is an expert in which types of cases, who has correctable systematic biases, and who is simply unqualified for the task. This feedback improves both the current dataset and the design of future annotation projects.

Probability distribution as a learning objective represents another powerful innovation. Instead of training the algorithm to predict a single label (pneumonia yes/no), we train it to predict the distribution of responses from expert annotators. If, for a given image, 60% of the radiologists see pneumonia, 30% see atelectasis, and 10% see normal findings, that is, the pattern the algorithm learns to reproduce. When the algorithm encounters a new image and predicts a similarly dispersed distribution, it inherently signals that the case is ambiguous. This explicit uncertainty is clinical gold: it allows the system to request human review when confidence is low or to present the clinician not just with a label but with a spectrum of possible interpretations along with their probabilities.

Stratified kappa analysis by case subgroup reveals patterns that the overall value hides. We can calculate separate Kappas for early versus late-stage disease, for different radiological presentations, or for patients from different demographic groups. We often find that agreement is high in typical cases but breaks down in atypical cases or that certain disease subtypes generate more variability than others do. This information is invaluable for algorithm design: it tells us where we can rely on labels for standard supervised training and where we need alternative strategies such as semisupervised learning, anomaly detection, or simply excluding cases from training and later labeling them via expert consensus.

Kappa confidence intervals, often ignored, are crucial for proper interpretation. In small datasets, kappa can be highly unstable; a value of 0.50 might have a 95% confidence interval ranging from 0.20 to 0.80, making any firm interpretation impossible. The calculation of these intervals, using bootstrap methods or analytical approximations, should be standard practice in any data quality report. Furthermore, comparing Kappas across different conditions (is agreement with Protocol A significantly higher than with Protocol B?) requires appropriate statistical tests that account for sample variability. Numerical precision matters when decisions about which data to use for training models that will affect human lives depend on these comparisons.

Integrating the annotator as part of the system, rather than as an external source of truth, opens up fascinating possibilities for medical artificial intelligence in the future. Imagine systems where the algorithm and the clinician collaborate iteratively: the algorithm proposes preliminary labels, the clinician corrects and comments, the algorithm learns from these corrections, and the process continues. In this paradigm, disagreement is not a failure but the engine of learning. Kappa becomes a metric of evolution: as the system improves, agreement between the algorithm's prediction and the expert's review should increase, but agreement among human experts may remain stable or even decrease if the algorithm discovers patterns that humans are overlooking. This complex dynamic requires new metrics that capture the quality of human-machine interactions, not just static agreement.

10.3 BIASES IN DATA COLLECTION: HOW TO IDENTIFY THEM

In recent years, the word “bias” has proliferated in conversations about artificial intelligence with a frequency that risks emptying it of meaning. We accuse an algorithm of being biased if it discriminates by gender, another if it performs worse on minority populations, and a third if it reproduces historical stereotypes. However, we rarely stop precisely defining what constitutes bias in the specific context of computational medicine or how it manifests in the very fabric of datasets. Here, bias is not a political opinion or a personal preference disguised as code; it is a systematic distortion, a consistent bias in the way we collect, measure, label, or select data that causes our algorithms to learn patterns that do not exist in reality or that exist only in a conveniently limited version of it. A biased algorithm is not merely imperfect; it is dangerous precisely because its harm is not random but directional, most severely affecting those who already face barriers to accessing quality healthcare.

Identifying biases must come before any discussion of performance metrics, accuracy, or

ROC curves. It is tempting to start by measuring how well our model performs, but that question is meaningless if we do not first know what reality we are comparing it to. An algorithm that achieves 95% accuracy on a biased dataset is not successful; it is an accident waiting to happen. That is why, before training the first neural network, we must subject our data to forensic scrutiny that examines not only what it contains but also what it has excluded, how it was collected, and what invisibly embedded assumptions it carries with it. This chapter explores the most common biases in clinical data collection, how to detect them, and why mitigating them is an ethical responsibility of both developers and clinicians.

Selection bias is the most frequent and, in many ways, the most fundamental. This occurs when the data sample used to train the algorithm is not representative of the population to which it will subsequently be applied. This mismatch is not a minor statistical error; it is a structural flaw that invalidates the generalizability of the model. Imagine an algorithm for detecting diabetic retinopathy trained exclusively on fundus images captured in university hospitals in major cities in Northern Europe. The images are high-resolution and taken with standardized equipment from patients with regular access to eye exams. When this algorithm is deployed in rural health centers in sub-Saharan Africa, where equipment is different, the prevalence of comorbidities varies, and disease presentation patterns differ; thus, its performance decreases. This is not because the algorithm is inherently flawed but because it learned to recognize European retinopathy, not human retinopathy.

Identifying selection bias requires a comprehensive demographic mapping of our data. We must systematically ask ourselves which institutions do they come from? Which geographic regions do they represent? Which age groups, genders, ethnicities, socioeconomic levels, and health insurance statuses are included? More importantly, who is missing? Gaps are more difficult to detect than presences are, but they are equally informative. If our breast cancer dataset does not include women under 40 years of age, the algorithm will never learn the atypical patterns of cancer at younger ages. If there are no patients with disabilities for whom standard imaging is difficult, the system will fail. Representation is not just a matter of social equity; it is a technical requirement for the model's robustness.

Stratified performance analysis is the most powerful diagnostic tool for detecting hidden selection bias. Even if the overall accuracy seems acceptable, we must disaggregate by relevant subgroups: Does the algorithm perform the same in men and women? In young and elderly people? In different ethnic groups? In patients with and without multiple comorbidities? We often find that the model performs well in the demographic majority but systematically fails among minorities. These disparities are not minor flaws; they are manifestations of selection bias that must be documented and, where possible, corrected through targeted sampling strategies, reweighting techniques, or additional data collection in underrepresented populations.

Exclusion bias is a particularly pernicious variant of selection bias. It occurs when certain groups are systematically excluded from the data, whether due to explicit inclusion criteria or implicit barriers to access. Traditional clinical trials, which frequently exclude pregnant women, children, frail elderly individuals, and patients with multiple comorbidities, generate datasets that do not represent the reality of clinical practice. When we train algorithms with these exclusionary data, we create systems that know how to function in "clean" patients but are dangerously ignorant of the complexity that characterizes most of those seeking medical care. Identifying this requires critically examining inclusion criteria and asking: Who is left out, and why? Is this exclusion medically justified, or is it methodological convenience that we are transferring to the algorithm?

Measurement bias arises when the method of capturing variables is inconsistent across groups or contexts, introducing artificial differences that the algorithm learns as if they were biological. Consider the seemingly trivial example of blood pressure. If in some centers in our dataset it is measured with the patient sitting, in others lying down, and in others standing; if some use cuffs of inadequate size for large arms; if the prior rest time varies between 5 and 20 minutes; then the numerical values are not directly comparable. An algorithm that learns from these inconsistent measurements may associate differences in measurement technique with differences in cardiovascular risk, producing absurd recommendations such as considering it riskier to go to certain hospitals simply because measurements are taken differently there.

Detecting this bias requires an archaeological examination of data collection protocols. We must explore the instructions given to staff, the device operating manuals, and the equipment calibration records. It is laborious but essential: a single poorly calibrated thermometer can introduce a temperature bias that affects sepsis detection in an entire subgroup of patients. Temporal variations also matter: if a laboratory analyzer was used during the first year of the study

and a different analyzer with varying specifications during the second year, we artificially created an “improvement” or “worsening” that the algorithm would erroneously attribute to patients or treatments.

Labeling bias, closely related to the interrater agreement challenges we explored previously, occurs when the “ground truths” in our dataset are not as accurate as we assume. If labels always come from the same radiologists, the algorithm will learn their idiosyncratic biases as if they were objective reality. If a particular pathologist tends to overdiagnose certain lesions, the model will reproduce that tendency. If labels are automatically extracted from clinical reports using natural language processing, we inherit all the errors, ambiguities, and inconsistent styles of those who wrote those reports. Identification requires an audit of the labeling process: Who did the labeling? What was their training? Under what time pressure? Are there independent reference standards against which to validate?

Confounding bias represents perhaps the most subtle and dangerous form of distortion. It occurs when the algorithm learns spurious associations—correlations that do not reflect causal or clinically relevant relationships. The classic example, documented in the literature, is the pneumonia detection algorithm that learned to identify the disease not by lung patterns but by the presence of chest drains. At the training hospital, patients with severe pneumonia frequently receive chest tubes; the algorithm, which is efficient in its correlational learning, simply learned that a chest tube implies pneumonia. When applied in centers where chest tubes were less common, its actual detection ability plummeted, even though it maintained high “accuracy” in detecting plastic tubes.

Identifying these confounders requires interpretable feature engineering and in-depth clinical knowledge. We must examine what the algorithm is actually seeing: is it focusing on relevant anatomy or on technical artifacts? On pathological findings or on process markers (such as the time of day of admission, which may correlate with severity in triage systems)? Explainability techniques such as SHAP or LIME can reveal these spurious dependencies, showing that the model uses proxy variables instead of those that truly matter. Prevention requires careful design of the datasets: intentional variability in procedures across centers so that the algorithm cannot learn to rely on specific local practices.

Survival bias is particularly insidious in longitudinal studies where time plays a central role. This occurred when we included only those patients who survived to a certain point in follow-up, systematically ignoring those who died or were lost to follow-up earlier. A risk prediction algorithm for heart failure trained on data from patients who completed one year of follow-up was used. Those who died in the first few months—precisely those at highest risk—are absent from the dataset. The algorithm learns “low-risk” patterns that are actually “short-term survival”, and when applied to new patients, it systematically underestimates the danger of severe cases.

Detection requires analysis of loss rates during follow-up and comparison of characteristics between those who complete the study and those who are lost to follow-up. If patients who are older, have poorer renal function, or have lower treatment adherence disappear from the dataset at higher rates, we introduce a dynamic selection bias that distorts the entire risk distribution. Statistical methods for survival analysis, with their explicit handling of censoring, must be applied not only to the clinical endpoint but also to continued participation in the study itself. Furthermore, we must be transparent about these limitations: an algorithm trained on one-year-old survivors should not be applied to palliative care decisions for terminally ill patients.

Recall bias, also known as outcome-knowledge bias, particularly affects data collected retrospectively. When a researcher reviews medical records and understands the patient’s outcome, they unconsciously look more intently for certain factors in the cases than in the controls. A physician who extracted data from patients who developed sepsis may have recorded previous febrile episodes more diligently than patients who did not develop it did, not because they were absent from the latter group but because the physician did not look for them with the same attention. This bias is especially dangerous because it resides in the mind of the data collector, not in the data itself, making it virtually impossible to detect through subsequent statistical analysis.

The only effective defense is a prospective design with standardized protocols that are blinded to the outcome. When data are collected before the result is known and when staff follow strict checklists without interpretive flexibility, we eliminate the opportunity for memory to distort data collection. For unavoidable retrospective data, we must apply sensitivity adjustments: we assume that certain findings may be underreported in controls or use multiple independent reviewers who are blinded to one another. Transparency regarding the retrospective origin of the data is mandatory; presenting it as if it were prospective constitutes scientific misconduct.

Publication bias, while not directly affecting a specific dataset, distorts the knowledge

ecosystem upon which we build our projects. Studies with positive results, high-performing algorithms, and successful innovations are published more frequently and in higher-impact journals than failures and implementations that did not work. When we review the literature to decide which model architecture to use or which features to include, we are looking at a sample biased toward success. Algorithms that “did not work” in certain contexts, limitations discovered too late, and biases identified postimplementation remain in researchers’ drawers or in the proceedings of minor conferences.

This distortion of available knowledge makes us unfoundedly optimistic. We believe that certain approaches are more robust than they actually are because we do not see the cases where they failed. Awareness of this bias should foster an attitude of healthy skepticism: valuing negative studies, actively seeking out reports of failure, and assuming that any published algorithm likely performed worse in unreported contexts. For developers, this means not blindly trusting high-impact papers but contacting authors for unpublished data, independently replicating studies, and being transparent ourselves about the limitations of our own systems.

10.3.1 TOWARD A PRACTICE OF CONSCIOUS BIAS COLLECTION

Identifying biases is not a one-time task that we complete before training; it is a continuous process that accompanies the entire algorithm lifecycle. Before we collect data, we must actively map out which populations might be excluded and why to design proactive inclusion strategies. During data collection, we monitor representativeness in real time, adjusting the sampling if we detect deviations. After training, we evaluate not only overall performance but also equity across subgroups, explicitly documenting where the algorithm fails. This continuous monitoring requires infrastructure: data pipelines that enable automated stratified analysis, equity dashboards that alert us to emerging disparities, and protocols for periodic auditing even after deployment.

For clinicians who participate in artificial intelligence projects, bias detection is perhaps their most valuable contribution. Their knowledge of real-world practice allows them to identify situations where data may not be representative: they know that Friday afternoon emergency room patients differ from Monday morning patients, that summer images show different patterns than winter images do and that certain groups avoid going to certain centers because of linguistic or cultural barriers. They can anticipate who will be harmed by a biased algorithm and advocate for their inclusion in evaluations. A clinician who points out a potential bias is not hindering progress; they are exercising their professional responsibility to protect patients.

Mitigating identified biases requires a diverse technical arsenal. Targeted stratified sampling can correct underrepresentation. However, no single technique is a panacea; each assumes that we know what biases exist, and all have trade-offs that must be evaluated.

Epistemic humility must be our compass. Recognizing that our data are always imperfect and that our algorithms will always be imperfect is not an excuse for complacency but a starting point for continuous improvement. Every identified bias is a learning opportunity. Every failure in specific subgroups is a call to action. The question is not whether our system is biased—it is—but whether we know its biases, whether we have documented them, whether we have implemented safeguards, and whether we are committed to reducing them iteratively. In that context, honesty about our limitations lies in the only solid foundation for building trustworthy medical artificial intelligence.

10.4 SAMPLE SIZE AND STATISTICAL POWER

“The question is not how much data you can accumulate, but how much you need to learn without memorizing, to generalize without ghosts.”

It is one of the most frequently asked questions at any initial medical AI project meeting and, curiously, one of the worst answered. The head of the cardiology department wanted to know how many ultrasounds his team must review. The director of the data unit calculates budgets on the basis of annotation hours. The principal investigator drafts ethical protocols with figures that later prove impossible to meet. All too often, the answer comes from intuition, availability (“we have access to five thousand historical cases”), or the simplistic mantra of “the more, the better.” However, this approach is dangerously naive. An inadequate sample size—whether too large or too small—can ruin a project in ways that become evident only when it is too late: algorithms that memorize noise instead of patterns, that fail spectacularly in new patients, or that consume

scarce resources on unnecessary data collection. Understanding the fundamentals of sample size and statistical power is not a methodological luxury; it is a survival requirement for any serious project.

The complexity of this issue goes beyond simple formulas because it depends on multiple factors that interact in nonlinear ways. There are no universal magic numbers or even foolproof rules of thumb. What does exist are sound statistical principles, estimation methods adapted to the era of machine learning, and above all, the need to think critically about what we want our algorithm to learn and with what degree of confidence. This chapter explores these fundamentals, from classic empirical rules to modern sample efficiency techniques, with the aim of providing a framework for making informed decisions about how much data we actually need.

10.4.1 THE FOUR PILLARS THAT DETERMINE HOW MUCH IS ENOUGH

The first pillar is the intrinsic complexity of the problem we are trying to solve. Not all medical artificial intelligence tasks are equally dataintensive. A simple binary classification between normal chest images and those with evident lobar consolidation, where visual differences are dramatic, may require relatively few examples for an algorithm to learn the distinctive patterns. In contrast, predicting the response to a specific cancer treatment on the basis of hundreds of genomic, clinical, and immunohistochemical variables—where signals are subtle and nonlinear interactions abound—requires substantially larger volumes of data. The classic rule of thumb of ten events per predictor variable, originating from logistic regression statistics, provides a conservative starting point: if we include twenty variables in our model, we need at least two hundred events of the outcome we are predicting. However, this rule is a floor, not a ceiling, and for complex models, it may be insufficient.

The second pillar, closely related but distinct, is the prevalence of the event of interest. In medicine, we frequently encounter rare outcomes: serious complications occurring in 2% of cases and autoimmune diseases with an incidence of one in a thousand specific genetic mutations present in minute fractions of the population. A dataset of ten thousand patients may seem impressive, but if it contains only one hundred positive cases of the condition we want to detect, its usefulness for training is limited. Algorithms, such as medical students, need to see enough examples of the pathology to learn to recognize it. One hundred cases scattered among ten thousand controls may not provide the variety of presentations, degrees of severity, and subtypes that characterize the disease. The rule of ten events per variable then becomes a difficult goal: with one hundred positive cases, we can only include ten variables with reasonable statistical confidence, which may be insufficient to capture the complexity of the phenomenon.

Population heterogeneity constitutes the third determining pillar. A homogeneous dataset, where all patients are of similar ages, come from the same center, share limited comorbidities, and were treated with standardized protocols, may require fewer cases to characterize its patterns. However, these savings in sample size come at the cost of poor generalizability: the algorithm learns medicine from that specific context, not universal medicine. If our goal is to develop tools applicable to diverse populations, we need data that reflect that diversity, which multiplies the required sample size. An algorithm that works equally well in young adults and frail elderly individuals, in patients with and without renal failure, and across different ethnic groups with relevant genetic variations needs sufficient examples in each subgroup to learn their particularities. Stratifying sample size by relevant subpopulations is an essential but frequently overlooked practice.

The fourth pillar is the complexity of the algorithm we intend to train. Simple models, such as logistic regression with few variables or shallow decision trees, have relatively few parameters to tune and therefore require less data to estimate them reasonably. Deep neural networks, especially convolutional architectures for medical images or transformers for clinical text, can have millions of parameters. Each additional parameter is a degree of freedom that the model can use to adapt to the training data, including its specific noise. Training these models with small datasets invites overfitting: the algorithm memorizes the examples it has seen, including their idiosyncratic peculiarities, but fails to extract the generalizable patterns that would allow it to perform on new data. The paradox is that more powerful models, theoretically capable of capturing more complex relationships, are also more data-hungry and more prone to failure when that data is scarce.

10.4.2 STATISTICAL POWER: THE PROBABILITY OF NOT BEING MISLED

The concept of statistical power, although originating in the design of traditional clinical trials, extends remarkably well to the field of medical machine learning. In its classical formulation, power is the probability of detecting an effect of a certain magnitude if that effect actually exists, thereby avoiding Type II errors (false negatives). In the context of algorithm development, we can reinterpret power as our study's ability to distinguish a genuinely useful algorithm from one that has merely been lucky with the available data. A study with low power has a high probability of producing misleading conclusions: it may lead us to believe that an algorithm works when its good results are due to random fluctuation, or it may fail to detect that a promising algorithm actually has merit.

Power depends on four factors that we must explicitly consider: the size of the effect we expect to detect, the inherent variability of the data, the sample size, and the significance level we accept. In medicine, large and consistent effects are rare; most useful interventions have modest effects on outcomes that vary widely among individuals. This leads to sample size requirements that are larger than we would intuitively expect. An algorithm that reduces mortality by 20% in relative terms—from 10% to 8%—is clinically important, but detecting this difference with confidence requires hundreds or thousands of patients, not dozens.

For the developer of clinical algorithms, power analysis is not a preproject academic exercise but a tool for ongoing planning. Before we begin, we must estimate the sample size needed to detect a clinically relevant improvement over standard practice. During development, we must monitor whether we are achieving the necessary power for our comparisons. Finally, we must report the power achieved so that others can evaluate the robustness of our conclusions. Ignoring power is similar to building on invisible foundations: it may support the building, or it may collapse under the first independent test.

10.4.3 PRACTICAL METHODS FOR DETERMINING THE RIGHT SAMPLE SIZE

Learning curves are the most intuitive and visual empirical method for assessing whether we have enough data. The logic is elegant in its simplicity: we train the algorithm on increasing subsets of our data—first 10%, then 20%, 30%, and so forth—and plot how the validation performance evolves. Initially, with little data, the performance is low and variable: the algorithm has not seen enough examples to learn robust patterns. As we increase the sample size, the performance improves, and the variance between runs decreases. At some point, the curve stabilizes: adding more data yields minimal marginal improvements. This saturation point indicates that we have reached the effective sample size for our given problem and model.

Interpreting these curves requires caution. If the curve is still rising steeply when we exhaust our data, we need more. If it has stabilized but is at an insufficient performance level, the problem may be data quality, model complexity, or the inherent difficulty of the problem. If there is a large gap between training and validation performance, we are facing overfitting: the model has the ability to memorize but not to generalize, suggesting that we need more data or a simpler model. Learning curves are powerful diagnostic tools, but they require that we already have enough data to perform subsampling, which makes them more useful for evaluating whether we need more data than they need for initial planning.

Simulation-based power analysis offers an alternative when data are not yet available. In this approach, we generate synthetic data with known characteristics—prevalence, variable effects, and noise levels—that simulate our real-world medical problem. We train the algorithm with different simulated sample sizes and observe at what size we correctly recover the patterns we know exist. This method is computationally intensive and critically depends on our simulations accurately capturing reality, but it provides estimates tailored specifically to our algorithm and problem. It is especially valuable for novel architectures where traditional sample size formulas do not apply.

Stratified cross-validation is an essential tool not only for evaluating performance but also for determining the stability of our results with respect to sample size. If we observe high variability in performance across different validation folds, it is a sign that our sample is small and that the results are unstable. A robust model should behave consistently regardless of which specific subset of data is set aside for validation. Excessive variability suggests that we are in the data-scarce regime, where sampling noise dominates the true signal. In this regime, increasing the sample size will likely improve both the average performance and the reliability of our estimates.

For binary classification problems, sample size formulas that combine the factors of prevalence, acceptable margin of error, and desired confidence level exist. The use of the Fleiss formula for proportions, adapted to the context of algorithm validation, allows us to calculate how many positive cases we need to estimate sensitivity with a certain degree of precision and how many negative cases for specificity. For example, to estimate a sensitivity of 90% with a 5% margin of error and 95% confidence, we need approximately 138 positive cases. If the prevalence is 10%, this implies a total dataset of approximately 1,380 patients. These formulas provide concrete figures for planning, although they must be adjusted for the complexity of the model and the need for independent validation.

10.4.4 EFFICIENCY AND WISDOM: WHEN MORE IS NOT ALWAYS BETTER

The future of sample size in clinical artificial intelligence does not rely solely on the accumulation of more data but rather on the use of existing data more intelligently. Transfer learning represents a revolution in sample efficiency: instead of training a model from scratch with data specific to our problem, we start with a model pretrained on related tasks—such as general medical image recognition or clinical language processing—and adapt it to our specific task with relatively few examples. A vision model pretrained on millions of natural images can learn to detect pneumonia in X-rays with just hundreds of medical examples because it has already learned fundamental visual representations that need only fine-tuning. This transfer of knowledge across domains multiplies the value of every specific piece of data we collect.

Data augmentation techniques, applied with clinical judgment, can artificially expand our dataset by generating plausible variations in existing examples. In medical images, rotations, translations, scalings, and contrast adjustments within physiologically reasonable ranges can create “new” examples that train the model to be invariant to these transformations. In time series, small Gaussian perturbations that preserve the temporal structure can increase robustness. However, these techniques have limitations: they cannot create information that does not exist, and if applied without discretion, they can generate medically impossible examples that confuse the algorithm. Augmentation must be conservative and clinically validated, not mechanical.

Adaptive data collection designs allow us to adjust the sample size in real time as we learn. Instead of setting a number *a priori* that may be excessive or insufficient, we collect data in stages, continuously evaluating the achieved accuracy and the stability of the estimates. When the confidence intervals of our performance metrics narrow sufficiently, we can stop data collection, saving resources. If the accuracy is insufficient, we continue. This approach, inspired by adaptive clinical trials, requires a logistical infrastructure for continuous data collection and interim analysis but optimizes the balance between information and cost.

For clinicians involved in artificial intelligence projects, understanding these fundamentals empowers them to engage in realistic negotiations. When a team of engineers promises to develop an algorithm using 100 cases in three months, the informed clinician can point out that the statistical power will be insufficient to detect clinically relevant effects or that cross-validation will reveal unacceptable variability. They can advocate for adequate resources for data collection, which typically consumes 70% or more of the budget and time of a serious project. They can critically evaluate published results: an algorithm with an area under the curve of 0.95 developed using fifty patients should be met with professional skepticism, not naive admiration.

Quality can never be completely replaced by quantity, but an insufficient quantity invalidates quality. An algorithm trained on a small amount of data, no matter how carefully annotated, is like a doctor who has seen only ten patients in his life: he may have excellent theory, but his clinical judgment will lack the empirical basis necessary for confidence. The crucial difference is that the doctor, aware of their limited experience, will be cautious, while the algorithm, lacking self-awareness, will issue predictions with the same confidence as if it had been trained on millions of cases. Building systems that estimate and communicate their own uncertainty—that know when they are operating outside their statistical comfort zone—is among the most important ethical challenges of modern medical artificial intelligence.

If we have learned to properly scale our datasets and validate their internal quality, how do we ensure that our model maintains its performance when it leaves the laboratory and encounters real patients in different hospitals, with different equipment, and with different clinical practices?

REFERENCES

McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica* [Internet]. 2012 [cited May 14, 2026];22(3):276. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/>