

Chapter 13 / Capítulo 13

Primum non nocere: Software engineering
for medicine
ISBN: 978-9915-9876-8-2
DOI: 10.62486/978-9915-9876-8-2.Ch13
Pages: 211-224



© 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 13. Explainability and Trust (XAI)

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 that gets it right but doesn’t explain why is like a doctor who makes the correct diagnosis but cannot justify it. In both cases, the correct answer may be due to chance, intuition, or even an error disguised as a correct answer. Medicine is not based on faith but on reasons. In addition, reasons, to be valid, must be expressible.”

Imagine that an artificial intelligence algorithm recommends against administering anticoagulant therapy to a patient at high risk of thrombosis. Its recommendation contradicts the doctor’s intuition, who believes the patient would clearly benefit from the treatment. The doctor asks the algorithm why it recommends against anticoagulation. The response is a list of probabilities, weights, and statistical correlations that shed no light on the decision. Doctors face a dilemma: they follow their clinical judgment, which is based on years of experience and deep knowledge of the patient, or follow the algorithm’s recommendation, which likely processes more data than he has, but whose reasoning he cannot understand. In this situation, there is no obvious correct answer, but there is one certainty: the lack of explainability undermines trust and hinders shared decision-making.

Explainability is neither a luxury nor an academic frill. It is a clinical, ethical, and legal necessity. Clinically, without understanding why an algorithm recommends something, the physician cannot integrate that recommendation with their own knowledge of the patient. Ethical, because patients have the right to understand the decisions affecting their health, even when those decisions are influenced by algorithms. In the event of an error, the courts will demand answers about how the decision was made, and “the algorithm said so” is not an admissible defense. Explainable artificial intelligence, known as XAI, is the field that seeks to endow models with the ability to explain their reasoning in a way that is understandable to humans.

In this chapter, we explore the fundamentals of explainability as applied to digital health. We begin with the fundamental premise: it is not enough for the algorithm to be correct; the physician must know why. We will continue with the most commonly used techniques for explaining models, such as SHAP and LIME, which allow predictions to be broken down into the contributions of each variable. We address the current limits of explainability, especially in complex models such as deep neural networks. Finally, we discuss how to communicate risk and uncertainty to the clinician and the patient—an aspect as important as the technical explanation itself. The goal is not to turn the reader into an expert in XAI but to equip them with the questions they should ask when faced with any system that aims to influence their clinical decisions.

13.1 IT IS NOT ENOUGH TO BE RIGHT; THE PHYSICIAN MUST KNOW WHY

“A diagnosis without an explanation is an oracle, not medicine; and doctors should not practice divination.”

The history of medicine is full of remedies that worked without anyone understanding why. Willow bark for fever, used since ancient times, only revealed its secret—acetylsalicylic acid—when organic chemistry matured. Digitalis, extracted from the purple foxglove, relieved heart failure centuries before we understood sodium and potassium channels. Quinine from the cinchona tree saved lives from malaria, while parasitology was still a dream. This functional ignorance was not a medical virtue but a historical limitation that science gradually overcame. Modern medicine is not content with knowing that something works; it demands understanding why it works because that knowledge allows us to predict side effects, extend applications to new situations, combine treatments rationally, and correct them when they fail. Explainability is not an intellectual luxury but a practical tool for improving care.

Medical artificial intelligence should be held within the same epistemological standard. An algorithm that predicts sepsis with high accuracy is useful, but its utility is greatly enhanced if it explains which factors drive that prediction. Is recent heart rate the main driver? The trend in blood pressure? A specific pattern of lab values? The nonlinear interaction of multiple subtle signals? With this information, the physician can direct diagnostic attention to the identified elements, verify whether they make physiological sense in the patient’s context, explain to the patient why they are considered highrisk, and decide whether the algorithm’s pattern applies to this particular situation. Without it, the prediction is a black-box verdict: it is either right or wrong, but it does not teach, does not allow for questioning, and does not respect the intelligence of the clinician who must use it.

13.1.1 WHEN ERROR IS KNOCKED AT THE DOOR: THE NEED TO DIAGNOSE

All predictive models make errors; in medicine, those errors can result in harm. When an algorithm fails, explainability becomes imperative, not optional. Without understanding why the model was wrong, we cannot correct it; we cannot prevent it from occurring again, and we cannot even determine whether the error was attributable to the algorithm or to an exceptional circumstance in which no reasonable model would have anticipated. A model that does not explain its errors is a model from which we cannot learn and therefore a model that cannot be fully trusted.

Imagine a sepsis risk prediction system that, in a particular case, predicts high risk, but the patient progresses without complications. If the model explains that its prediction was based primarily on mild tachycardia in a young patient with no other signs, the physician can recognize that the model overemphasizes that sign in this context, adjust their interpretation, and report the case for algorithm refinement. If the model does not explain, the physician only knows that it failed, without knowing whether it was a systematic error or an exception and without the ability to take preventive action. Explainability transforms an opaque error into an opportunity for improvement.

The ability to learn from errors is especially critical in medicine, where biological variability ensures that unusual cases will occur. A model that explains this situation allows for the creation of a repository of borderline cases, an understanding of failure conditions, and iterative improvement. Without explanation, each error is an isolated event with no discernible pattern; with them, errors become data for the system's evolution. This capacity for guided self-correction is the hallmark of mature systems versus fragile ones.

13.1.2 DETECTING THE INVISIBLE: BIASES REVEALED BY THE LIGHT OF EXPLANATION

Explainability also serves as a detector of biases operating in the dark. An algorithm may systematically discriminate against certain groups without its creators being aware of it, especially when discrimination operates through proxy variables. If the model does not explain its decisions, this bias can remain hidden for years, causing cumulative silent harm. If the model explains, even partially, the explanations can reveal suspicious patterns: that the ZIP code has an unusually high weight in risk predictions, suggesting confusion between geographic location and socioeconomic status; that age is used differently for men and women, indicating possible gender bias; and that certain hospital names appear as strong predictors, revealing overfitting to the training context.

These signals, visible only when the model opens up its logic, can trigger deeper investigations and corrections before widespread harm occurs. Explainability is thus a mechanism for ethical oversight, not just technical oversight. Without it, algorithmic justice is an impossible aspiration; with it, it is a difficult but attainable goal.

An example of the “zip code” as a predictor of health risk illustrates the problem. In many health systems, the zip code strongly correlates with health outcomes, not by geographical coincidence but because it captures socioeconomic variables: access to healthy food, air quality, opportunities for exercise, and environmental stress. A model that uses ZIP code as a direct predictor may seem accurate but perpetuates structural inequalities. Only when the model explains its reliance on this variable and when clinicians question whether it is appropriate can a correction toward more direct causal variables emerge.

13.1.3 TRUST AS A CONDITION OF USE: WHY CLINICIANS IGNORE WHAT THEY DO NOT UNDERSTAND

Trust is a determining factor in clinical adoption, and explainability is a necessary condition for justified trust. Repeated studies show that clinicians ignore recommendations from decision support systems when they do not understand their reasoning, even when those recommendations are correct. This resistance is not an irrational obstinacy but professional prudence: a physician who blindly follows an algorithm they do not understand abandons their responsibility for clinical judgment becomes dependent on a tool that may fail without warning and loses the ability to explain their decisions to patients and colleagues.

Explainability does not guarantee trust; a physician may understand the explanation and find it insufficient, but its absence guarantees mistrust or, worse, unfounded trust. An algorithm that does not generate trust is an algorithm that is not used, no matter how accurate it may be.

in validation. Explainability is thus a condition for real utility, not a mere ethical embellishment.

The dynamics of trust are complex and deserve specific study. Research shows that doctors trust systems more when they explain using familiar concepts rather than technical model concepts, even when both explanations are equivalent in informational content. Familiarity with the format matters as much as the information. These findings suggest that effective explainability requires not only technical expertise but also a design centered on clinical users, which translates model outputs into language that resonates with medical reasoning.

13.1.4 LEGAL LIABILITY AND DECISION TRACEABILITY

The relationship between explainability and legal liability is a growing point of tension in algorithmic medicine. In the event of an error resulting in harm to the patient, courts will ask who is legally liable. If the physician followed an algorithmic recommendation without understanding it, does that limit their liability or increase it because of negligence in exercising their own judgment? If they ignored a correct recommendation and harm occurred, are they more liable for failing to follow the evidence or less so for not trusting a black box? Legal answers are evolving, but in any scenario, having clear explanations of what the algorithm recommended and why is essential for reasonably determining liability.

Without explanations, the legal debate devolves into speculation about the mental states and capabilities of opaque systems. With explanations, one can assess whether the model's reasoning was reasonable, whether the physician had grounds to accept or reject it, and whether the harm was preventable. Explainability is thus the infrastructure of justice, not just of clinical efficiency.

Emerging case law on medical artificial intelligence suggests that the “black box” will not be an acceptable defense for developers. If a model causes harm and cannot explain why it made the decision it did, courts will tend to attribute liability to the manufacturer for a design defect. This creates powerful legal pressure toward explainability that complements ethical and practical pressure. Opaque systems become legal risks, not just technical ones.

13.1.5 LEVELS OF EXPLAINABILITY: FROM TRANSPARENT TO APPROXIMATE

There are different levels of explainability, and not all the models achieve them equally. Some models are inherently interpretable: decision trees reveal their logic in their branching structure, logistic regression expresses its criteria through coefficients that quantify each variable's contribution, and association rules present explicit patterns. These models allow for a direct understanding of how they reach conclusions, generally at the cost of reduced predictive power for complex nonlinear relationships.

Other models—deep neural networks, gradient boosting, and support vector machines with complex kernels—are black boxes that require posthoc techniques to generate explanations. These techniques, such as SHAP and LIME, which we explore in the next section, approximate the model's behavior locally, revealing which factors influence a specific prediction without revealing the complete internal workings. They are like telescopes that show part of the model's universe, not the entire universe.

There is a trade-off between interpretability and predictive power that must be consciously decided. For high-risk decisions where understanding is a priority, interpretable models may be preferable even if their performance is slightly lower. With respect to complex patterns in which only deep architectures capture the signal, posthoc explanation techniques may be acceptable because of their limitations. The choice is a clinical decision, not merely a technical one, and must involve those who will use and rely on the system.

13.1.6 TOWARD EXPLANATIONS THAT SPEAK THE LANGUAGE OF MEDICINE

The future of explainability lies in deep integration between artificial intelligence and clinical reasoning. It is not just a matter of algorithms explaining their decisions but of those explanations aligning with how clinicians actually think. A clinician does not primarily reason in terms of variable weights or SHAP contributions; they reason in terms of pathophysiological mechanisms, coherent clinical histories, similar past experiences, and recognizable patterns. Explanations should be able to be translated into this language: instead of “variable X contributing 0.3 to the prediction,” prefer

“the pattern of tachycardia disproportionate to fever suggests an early systemic inflammatory response.”

Translating technical explanations into clinical narratives is a massive challenge that requires collaboration between engineers and clinicians, language models that rephrase explanations in understandable terms, and validation that these rephrasings remain faithful to the original reasoning. However, it is a necessary investment: explanations that do not speak the language of their recipients do not truly explain; they merely transfer opacity from one form to another.

For developers, incorporating explainability from the design phase is a responsibility, not an option. In terms of model selection, interpretability and performance should be considered. During development, explanations are consistent with clinical knowledge and useful for end users. At deployment, interfaces that present understandable explanations without overwhelming users with unnecessary details should be provided. In addition, maintain humility: an explanation is an approximation, not a complete revelation.

For clinicians, explainability is a tool to exercise judgment with more information, not to delegate it to the machine. Receiving an explanation should lead to the following questions: Does it make clinical sense? Does it align with what I know about this patient? Are there factors the algorithm did not consider? The explanation is *ne*, nor is it revealed truth but rather another input for clinical deliberation. The physician remains ultimately responsible for the decision, and the explanation should help them make it with greater knowledge and not abdicate their responsibility.

In that dialog between algorithmic explanation and clinical judgment, where each contributes what the other lacks, lies the potential for medical artificial intelligence that truly improves care, not just speeds it up.

13.2 EXPLAINABILITY TECHNIQUES: SHAP, LIME, AND OTHERS

The need to explain algorithmic decisions has spawned a flourishing field of research, with dozens of techniques promising to shed light on the inner workings of opaque models. Some have become *de facto* standards, adopted globally by developers who need to justify predictions to users and regulators. However, like any powerful tool, they have significant limitations and can be misused. Understanding what they are, how they actually work, and when they are appropriate versus when they are misleading is essential for clinicians who will rely on the generated explanations. This chapter explores the techniques most relevant to medicine, their fundamentals, their strengths, their critical weaknesses, and the dangers of interpreting them naively.

13.2.1 LIME: THE LOCAL APPROACH TO THE INCOMPREHENSIBLE

LIME, an acronym for local interpretable model-agnostic explanations ((1)), is an intuitively appealing and widely used technique. Its philosophy is elegantly simple: rather than explaining the entire model globally—a task that would be prohibitively complex for deep architectures—it focuses on explaining individual predictions one by one. To explain why the model produced a certain prediction for a specific patient, LIME generates small synthetic variations in that patient by slightly altering their characteristics and observing how the model's prediction changes. Using these perturbed data, a simple, local model—typically a linear regression—is constructed that approximates the behavior of the complex model in the immediate vicinity of that patient. The features with the greatest weight in this local model explain the original prediction.

LIME's fundamental strength lies in its model agnosticism: it can explain any algorithm, from deep neural networks to random forests or support vector machines (), without needing to know its internal architecture or have access to its parameters. This makes it universally applicable and conceptually accessible. Its implementation is relatively straightforward, and its output—a ranking of features by local importance—is easily understandable to nontechnical users.

However, LIME has critical weaknesses that must be understood. Explanations can be unstable: small variations in the random generation of synthetic data can produce substantially different explanations for the same patient, undermining confidence in any specific explanation. The quality of the explanation depends crucially on how the patient's “neighborhood” is defined and what constitutes “small variation”—a decision that has no single correct answer and whose choice can significantly impact the results. If the neighborhood is too small, the local model may

fail to capture the relevant behavior; if it is too large, the linear approximation may be poor. The LIME explains which features changed the prediction in the local context but not why the original model uses those features the way it does.

13.2.2 SHAP: GAME THEORY FOUNDATIONS FOR FAIR ATTRIBUTION

SHAP, which stands for SHapley Additive exPlanations ((2)), represents an approach with a more solid mathematical foundation than LIME does. It is based on Shapley values, a game theory concept that distributes the “payoff” of a cooperative game among players in a way that satisfies certain fairness axioms. In the context of explainability, the “game” is the model’s prediction for a specific instance, and the “players” are the patient’s features. SHAP calculates the marginal contribution of each feature to the difference between the prediction for that patient and the reference mean prediction for the entire population. The result is an additive decomposition where the sum of the contributions of all the features plus a base value exactly equals the model’s prediction.

The main advantage of SHAP is theoretical consistency: unlike LIME, SHAP explanations are stable and satisfy desirable properties such as efficiency (the contributions sum to the full prediction), symmetry (identical features receive equal contributions), and additivity for linear models. There are computationally efficient implementations for specific models, such as decision trees (TreeSHAP) and neural networks (DeepSHAP), which avoid the prohibitive cost of exact calculation of Shapley values. The associated visualizations—summary plots showing the distribution of global importance, force plots illustrating how specific features push the prediction up or down, and dependence plots revealing interactions—are informative and aesthetically clear.

The main disadvantage of SHAP is the computational cost of exact calculations, although modern approximations mitigate this. More fundamentally, SHAP assumes that all features are potentially interchangeable, which may not be realistic in medical data where certain variables are causally related. Interpreting SHAP values requires caution: they represent the average contribution across all possible feature permutations, not necessarily the actual causal mechanism. A high SHAP value for “heart rate” does not imply that changing the heart rate would alter the risk; only the model uses this variable to distinguish risks in this population.

13.2.3 DEPENDENCE PLOTS: FROM GLOBAL TO LOCAL

Partial dependence plots (PDPs) offer a global perspective that complements local explanations. They show how the model’s average prediction changes when a specific feature is varied, while all others are held at their marginal values. They are useful for understanding the general relationship between a variable and an outcome, revealing patterns such as nonlinearities or critical thresholds. For example, a PDP may show that the risk of sepsis increases monotonically with increasing lactate concentration up to a certain point, where it stabilizes—valuable information for calibrating clinical alarms.

However, PDPs hide heterogeneity: a variable may have a positive effect in one subgroup and a negative effect in another, and the PDP would show an average effect that does not represent any real patient. This limitation is particularly problematic in medicine, where interactions between variables are common. ICE plots (individual conditional expectations) extend PDPs by displaying dependency curves for each individual patient, revealing variation in effects. If lines from different patients are parallel, the effect is homogeneous; if they cross, there are significant interactions with other variables.

Interpreting ICE plots requires careful sampling: with thousands of patients, the graph becomes unreadable. Selecting representative subgroups or using curve clustering techniques can help. Visualizing heterogeneity is powerful for identifying subpopulations where the model behaves differently, suggesting a need for stratification or specific models. However, it also increases complexity: instead of a simple story about a variable’s effect, the clinician must navigate multiple possible stories.

13.2.4 SURROGATE MODELS AND THE ILLUSION OF UNDERSTANDING ()

Decision rules and surrogate trees represent a different family of techniques. They consist of training a simple and fully interpretable model—typically a decision tree with few branches—

to mimic the predictions of the complex model across the entire dataset or relevant subsets. The resulting tree can be examined to understand what patterns the original model has learned, presenting rules of the form “IF age > 65 AND creatinine > 1.5 THEN high risk.” This form of explanation resonates with rule-based clinical reasoning, making it intuitive for physicians.

The fundamental limitation is that the replacement tree only approximates the behavior of the original model. If the true model is highly nonlinear with complex interactions, the tree needed to approximate it faithfully may become so large and intricate that it loses interpretability. A tree with one hundred nodes is technically interpretable but practically incomprehensible. Furthermore, multiple different trees may approximate the complex model equally well but suggest different explanations, creating a problem of nonuniqueness. The surrogate tree explains which rules approximate the model, not necessarily which rules the model uses internally.

Saliency maps for medical imaging deserve special mention given the importance of images in diagnosis. Techniques such as Grad-CAM or layerwise relevance propagation generate heatmaps over medical images, highlighting regions that contributed most to the classification. (2) For a radiologist, seeing which areas of an X-ray triggered a diagnosis of pneumonia can be more useful than an abstract numerical explanation. These techniques are rapidly advancing toward more granular explanations, identifying not only regions but also specific patterns within them.

However, saliency maps can be misleading. They may highlight regions that are technical artifacts or spurious correlations rather than genuine pathological findings. A model that has learned that patient ID tags correlate with a certain diagnosis (due to training bias) will show saliency over the tag, not over relevant anatomy. Validating that highlighted regions have clinical significance requires medical expertise and systematic correlation studies with expert interpretation.

13.2.5 BEYOND THE TECHNIQUE: TOWARD MEDICALLY MEANINGFUL EXPLANATIONS

For developers, the choice of explainability technique depends on multiple interrelated factors: model type, data nature, the audience for the explanations, and the specific purpose. There is no universally superior technique; each has strengths and weaknesses that must be weighed. The most sensible practice is to use multiple complementary techniques, compare their results, and verify the consistency among them and with established clinical knowledge. If LIME and SHAP agree that heart rate is important for this patient, confidence increases; if they disagree, further investigation is needed.

For clinicians, understanding the limitations of these techniques is just as important as understanding their contributions. A SHAP explanation is not “the reason” why the model made its decision; it is a specific mathematical attribution that could be different under another reasonable methodology. A LIME explanation is a local approximation that may fail if the model is highly nonlinear in that region. Saliency maps may highlight regions without clinical significance. Explanations should be taken as hypotheses generated by the model, not as revealed truths, and are subject to clinical verification.

The distinction between explainability and causality is a common pitfall. SHAP may indicate that “heart rate” is important for predicting sepsis, but this does not imply that heart rate causes sepsis or that modifying it reduces risk. Explainability reveals correlations that the model has learned, not causal relationships in the real world. Clinicians, accustomed to causal reasoning, must be explicitly aware of this distinction so as not to misinterpret explanations as justifications for interventions.

The future of medical explainability lies in deeper integration with clinical knowledge. Instead of generic explanations in terms of variables, methods are being developed that generate explanations in clinical concepts: “This patient is at high risk because they present a pattern of tachycardia disproportionate to their fever, with elevated lactate, consistent with an early systemic inflammatory response.” These explanations, closer to medical reasoning, are more useful and generate greater justified confidence.

Interactive explanation techniques are also emerging, where clinicians can ask “what if creatinine were normal?” and see how the prediction would change. This capacity for counterfactual analysis allows us to explore the robustness of predictions and understand which variables are critical versus which are noise. However, it also requires caution: models lack true causal knowledge, and counterfactual predictions are extrapolations that may not correspond to biological reality.

13.3 CURRENT LIMITS OF EXPLAINABLE AI

There is a natural tendency, with any new tool, to attribute powers to it that it does not possess. With explainable artificial intelligence, this is precisely what has happened: a collective illusion has been created that techniques such as SHAP or LIME allow us to open the black box and truly understand why models make their decisions. The reality is more modest and, in a certain sense, more unsettling. The explanations we obtain are not transparent windows into the model's inner workings but rather partial, approximate, potentially misleading interpretations and are always mediated by mathematical choices that are neither unique nor neutral. Understanding the fundamental limits of explainable AI is an ethical and practical imperative: Without this understanding, we fall into a false sense of security that can be just as dangerous as the opacity we are trying to combat.

13.3.1 APPROXIMATION AS A FEATURE, NOT A FLAW

The first limitation, perhaps the most fundamental yet frequently overlooked, is that posthoc explainability techniques do not reveal the model's actual inner workings but rather generate local approximations of its observable behavior. When SHAP assigns specific importance to a variable, it does not show how the model actually uses that variable in its internal layers—calculations that may be opaque even to designers—but rather calculates a contribution on the basis of specific game theory axioms. Another reasonable technique, with different but equally valid mathematical foundations, might attribute different importance to the same variable. The explanation is not “the truth” about the model; it is one perspective among many possible ones, each with its own implicit assumptions.

This multiplicity of valid explanations is uncomfortable but inevitable. The complex model operates in a high-dimensional space with nonlinear interactions that do not admit a simple, unique decomposition. Any explainability technique imposes a simplifying structure—local linearity, additivity, and independence of features—which is an approximation, not a revelation. The danger arises when we forget this condition of approximation and treat the explanations as if they were faithful photographs of the model's interior. They are maps, and like any map, they distort the territory they represent.

13.3.2 INSTABILITY, LACK OF FIDELITY, AND THE VERIFICATION PROBLEM

The second critical limitation is the instability of explanations in the face of small variations. In LIME, the random sampling process used to generate neighborhoods can produce substantially different explanations for the same patient across different runs, even when the model's prediction barely varies. This instability is not an implementation bug but a characteristic of the stochastic approach. For clinical decision-making, an explanation that changes drastically with a different random seed is unreliable, but detecting this instability requires specific tests that are rarely performed in clinical practice.

The lack of fidelity constitutes a third fundamental limitation. An explanation is faithful if it correctly reflects the model's behavior, but measuring this fidelity is a circular problem: we have no independent access to the “truth” about how the model works against which to validate the explanation. Surrogate metrics—the ability of key features to alter predictions when modified—are imperfect. A model can generate plausible explanations, consistent with human expectations, that do not faithfully represent its internal workings. This “misleading plausibility” is particularly dangerous because it generates unwarranted confidence.

The difficulty of capturing complex interactions is a fourth major limitation. Modern models learn highly nonlinear relationships where the effect of variable A depends radically on the values of variables B, C, and D. Techniques that attribute importance to variables individually—including SHAP in its standard form—may miss these essential interactions. A variable may be critical only in a specific combination with another, and univariate explanations suggest moderate importance for both without capturing synergy. Techniques that explicitly model interactions exist but are computationally demanding and conceptually complex, limiting their adoption.

13.3.3 CAUSAL TRAP AND VALIDATION OF HUMAN PROBLEMS

The confusion between learned correlation and causal relationships constitutes a fifth limitation with serious practical consequences. An explanation may indicate that “heart rate” is important for predicting sepsis, but this does not imply that heart rate causes sepsis or that modifying it, if possible, would change the risk. The statistical correlation of the training data in the model revealed that a high heart rate may be a marker of an inflammatory response, or it may be a consequence of another unmeasured variable that is the true causal driver. Clinicians trained in causal reasoning may misinterpret such explanations as justifications for interventions targeting highlighted variables, with potentially harmful results if the relationship is not causal.

The difficulty of validating explanations with human experts is the sixth methodological limitation. In theory, expert clinicians should be able to judge whether explanations make medical sense. In practice, this assumes that experts can assess the plausibility of complex combinations of variables that may not correspond to familiar clinical patterns. Furthermore, experts have their own cognitive biases, potentially preferring explanations that align with their prior knowledge even if they are not faithful to the model or reject valid but counterintuitive explanations. Human validation is necessary but not sufficient and can introduce biases of its own.

The tension between explainability and privacy emerges as a seventh limitation in healthcare settings. Detailed explanations, especially when they reveal specific contributions of features to individual predictions, can leak information about training data. A sophisticated attacker could use sequences of explanations to infer patient characteristics from the training dataset or carry out membership attacks. This vulnerability is particularly relevant where privacy is a fundamental value, requiring that explanations be consciously designed to minimize the risk of information leakage, possibly at the expense of granularity.

13.3.4 LACK OF STANDARDIZATION AND COGNITIVE OVERLOAD

The lack of standards constitutes the eighth structural limitation. There is no standard equivalent to IEC 62304 or ISO 14971 for AI explainability. Each developer chooses techniques arbitrarily, presents explanations in proprietary formats, and makes inconsistent claims about their reliability. This lack of standardization hinders comparisons between systems, consistent regulatory evaluation, and the accumulation of evidence on which approaches work best in specific clinical contexts. Regulatory bodies are beginning to address this gap, but the pace of technological development outstrips that of regulation.

Cognitive overload is a frequently overlooked practical limitation. Providing detailed explanations, with multiple variables, interactions, and visualizations, can overwhelm clinicians who are short on time. A physician in the emergency department does not have the ability to study complex SHAP plots while attending to multiple critically ill patients. Explanations must be concise, contextually relevant, and presented at the appropriate moment in the workflow. However, determining what is relevant requires a deep understanding of the decision context, the clinician’s specific needs at each moment, and the opportunity costs of care. Designing effective explanations is as much a problem of information engineering as it is of mathematics.

The tenth and most insidious limitation is that explainability itself can create a false sense of security. A physician who receives a plausible explanation may place greater trust in the algorithm, even if the explanation is mathematically misleading or clinically inadequate. This “misplaced trust” can lead to the acceptance of erroneous recommendations without the critical scrutiny that would be applied to a prediction without an explanation. Misunderstood, explainability becomes a veneer of legitimacy for decisions that deserve to be questioned. It is a dangerous paradox: the tool designed to increase transparency can, in the hands of users unaware of its limitations, reduce critical vigilance.

13.3.5 NAVIGATING THE LIMITS: HUMILITY AND MINDFUL DESIGN

For developers, these limits imply that explainability must be approached with epistemic humility and explicit transparency regarding limitations. One cannot promise that explanations reveal the model’s inner workings; one must communicate that they are approximations subject to uncertainty. One cannot assume that a single technique is sufficient; multiple approaches must be used, their consistency compared, and the discrepancies explicitly noted. Decision-making responsibility cannot be delegated to the explanation; the clinician remains ultimately

responsible, and the explanation is an input for their judgment, not a substitute for it.

For clinicians, these limits require maintaining a critical attitude even in the face of more plausible explanations. An explanation is not revealed truth but a hypothesis about why the model predicted what it did. It must be contrasted with clinical knowledge, the specific patient context, and medical common sense. If the explanation does not make clinical sense, the model is likely failing or the explainability technique is inadequate for this case. Trust should not be placed in the explanation as an object but in the process of informed deliberation that the explanation facilitates.

More robust techniques are being developed through the aggregation of multiple explanations, more accurate methods validated against the model's empirical behavior, and approaches that integrate explicit causal knowledge to generate more medically meaningful explanations. Standards for evaluating and presenting explanations and guidelines for effectively communicating associated uncertainty are being explored. However, while these advances mature, prudence advises against attributing more power to explanations than they actually possess. The black box has not become transparent; we have merely learned to project interpretations onto some of its surfaces that must be treated with constructive skepticism. It is in this humility regarding what we truly understand—rather than in excessive confidence in explainability tools—that the possibility of safe and effective use of medical artificial intelligence lies.

13.4 COMMUNICATING RISK TO CLINICIANS AND PATIENTS

Let's imagine that an algorithm has calculated that a patient has a 15% risk of developing postoperative complications. That number, in isolation, is almost useless for decision-making. Is the risk high or low? Compared to what baseline? What does "15% risk" mean for this specific patient, at this point in their life, with this unique medical history and with these personal values and preferences? The answer lies not in the number itself but in how it is communicated, how it is contextualized, and how it is integrated into the conversation between clinician and patient. Risk communication is an art that combines science, psychology, and empathy, and the arrival of artificial intelligence at the patient's bedside does not simplify this complexity; it amplifies it with new technical and ethical dimensions that we must learn to navigate.

13.4.1 PSYCHOLOGY OF PROBABILITY: WHY NUMBERS ARE NOT ENOUGH

The first fundamental difficulty is that humans, including doctors, do not process probabilities in an intuitively correct way. We tend to think in binary terms: this will happen or it will not happen. A 15% risk may be interpreted as "almost certainly won't happen" by natural optimists or as "quite likely to happen" by those who have had previous negative experiences, depending on their cognitive biases, personality, and emotional context. Doctors, despite their scientific training, are not immune to these biases; they can also misinterpret probabilities if they are not presented carefully. That is why the way information is communicated is just as important as the number itself—and sometimes even more so.

Presentation formats dramatically impact understanding. Numerical probabilities can be expressed as percentages (15%), natural frequencies (15 out of every 100 patients similar to you), or relative comparisons (the risk is 1.5 times the average for your age). Research in health psychology consistently shows that natural frequencies are better understood than percentages are, especially among patients with lower levels of formal education. However, the frequencies can be less accurate when the numbers are small. Comparisons help provide context, but they can be misleading if the reference group is not chosen carefully. Visual formats—pictograms with human figures, bar graphs, and risk thermometers—complement numerical information and facilitate intuitive understanding, although they introduce their own design biases.

13.4.2 CLINICAL CONTEXTUALIZATION: THE MEANING OF NUMBERS IN REAL LIFE

The clinical context radically transforms the meaning of any probability. A 15% risk of postoperative complications may be entirely acceptable if surgery is necessary to remove a malignant tumor with a good prognosis and unacceptable if it is an elective cosmetic procedure. The same number has different ethical implications depending on the situation. Therefore, risk communication cannot be limited to presenting a figure; it must include a discussion of available

alternatives, the expected benefits of the intervention, the consequences of inaction, and the patient's values and preferences. The algorithm provides the calculated probability, but the interpretation of that probability is inherently clinical, personal, and situational.

The uncertainty associated with predictions must be explicitly communicated, even if this is technically challenging. A 15% risk calculated by a model has a confidence interval, variability due to model limitations, and uncertainty regarding the representativeness of training data for this specific patient. If the interval is wide—for example, between 5% and 35%—the point estimate may be deceptively precise. While communicating this uncertainty is ethically necessary for genuine informed consent, it is conceptually difficult. Confidence intervals are statistical abstractions that many patients—and not a few physicians—find confusing. Visual simulations showing the distribution of possible outcomes can help, but they do not resolve the tension between epistemic honesty and communicative clarity.

13.4.3 TIMING, CHANNEL, AND HUMAN MEDIATION

Risk communication must take into account the timing and channel of delivery. Communicating risk during a calm consultation with time for questions is not the same as displaying it on a website without clinical supervision. Health care systems are beginning to make algorithmic predictions directly available to patients through mobile apps and online portals without clinical oversight. This poses risks of misinterpretation, anxiety generated by decontextualized information, and hasty decisions based on partial understanding. Risk communication cannot be completely delegated to technology; the human clinical encounter remains irreplaceable for the emotional and comprehensive assimilation of information.

For developers, designing systems that effectively communicate risk requires multidisciplinary collaboration with experts in risk communication, cognitive psychologists, and experienced clinicians. It is not enough to simply display numbers; interfaces must be designed to automatically provide context, allow for “what-if” scenario exploration, and facilitate clinical conversation rather than replace it. We must avoid “presentism,” which displays results without adequate explanation, and “alarmism,” which exaggerates minor risks to grab attention. Interfaces must be tested with real users of diverse backgrounds, refined for comprehensibility, and evaluated for their impact on real-world decisions, not just in the laboratory.

13.4.4 CLINICAL TRAINING AND SHARED DECISION-MAKING

For clinicians, the advent of algorithmic predictions adds a layer of complexity to an already difficult task. They must understand the predictions technically, explain them to patients in an understandable way, and integrate them into shared decision-making processes. This requires specific training not only in interpreting metrics but also in risk communication techniques, managing emotions in the face of probabilities, and navigating divergent values. It also requires time, which is a scarce resource in contemporary clinical practice. If systems are not designed to facilitate this communication—by providing suggested language, answers to frequently asked questions, and visual support—predictions may be ignored because of information overload or used in an inappropriately simplified manner.

The relationship between risk communication and shared decision-making is intrinsic. The goal is not for the physician to convey information and the patient to passively accept it but for both, with their respective knowledge and perspectives, to reach a joint, informed decision. The algorithm can inform this decision by providing quantified estimates, but it cannot replace the deliberative process. Risk communication should facilitate this dialog, not shut it down with numerical authority. It should open up questions about values and preferences, not close them with apparent objectivity. It should acknowledge uncertainty as a human condition, not hide it behind false precision.

13.4.5 CULTURAL DIVERSITY AND PERSONALIZATION OF COMMUNICATION

Risk communication to patients from different cultures, educational levels, and prior experiences adds a crucial layer of complexity. What works for a patient with scientific training and prior experience with healthcare systems may not work for a patient with low health literacy or a history of distrust toward medical institutions. Systems must be flexible, allowing for different

formats and levels of detail on the basis of the patient's explicit or inferred preferences. The personalization of risk communication is an emerging field with the potential to improve equity and effectiveness but also carries the risk of algorithmic biases if inferences about preferences are incorrect.

The cultural dimension is particularly sensitive. Concepts of risk, probability, and personal responsibility vary across cultures. In some cases, direct communication of bad news is considered insensitive; in others, withholding information is considered paternalistic. Risk communication systems must be culturally adaptable and not universally identical. This requires ethnographic research on communication practices in different communities and technical flexibility to accommodate variations.

13.4.6 HORIZONS: TECHNOLOGY IN THE SERVICE OF HUMANITY

The future of risk communication in medicine lies in deeper integration between artificial intelligence and human communication, but with the aim of enhancing, not replacing, the clinical relationship. Systems will be able to generate personalized explanations tailored to the patient's level of understanding, format preferences, and situational context. They will be able to simulate scenarios showing how risk would change with different decisions, facilitating informed deliberation. They will be able to suggest questions the patient might want to ask, on the basis of profiles of similar users, enriching the dialog. However, these technical capabilities will not replace the clinician; they will free them from routine tasks so that they can devote more time to what only humans can do: listen with full attention, empathize with genuine suffering, and accompany patients through existential uncertainty.

In closing, it is worth remembering that behind every calculated risk, there is a specific person. A person with legitimate fears, fragile hopes, and uncertain plans for the future. A person who needs to understand the information but also needs to be heard in their full humanity. A person who deserves for us to speak honestly about what we know and what we do not know, clearly about implications, and with compassion about limitations. Algorithms can calculate risks with increasing precision, but they cannot communicate them with humanity. That task remains our inalienable responsibility, and we must do it well because it depends not only on the quality of individual decisions but also on the trust in the entire medical enterprise and on the very possibility of technologically assisted yet humanly centered care.

Once we have learned to predict, explain, and communicate risk, how do we design studies that demonstrate that all this effort truly improves clinical outcomes and not just technical metrics?

REFERENCES

- Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstration Session [Internet]. February 16, 2016 [cited May 14, 2026];97-101. Available at: <https://arxiv.org/pdf/1602.04938>
- Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst [Internet]. May 22, 2017 [cited May 14, 2026];2017-December:4766-75. Available from: <https://arxiv.org/pdf/1705.07874>