

Chapter 16 / Capítulo 16

Primum non nocere: Software engineering
for medicine
ISBN: 978-9915-9876-8-2
DOI: 10.62486/978-9915-9876-8-2.Ch16
Pages: 262-274



© 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 16. Ethics and Algorithmic Bias

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 not ethical or unethical in and of itself; it is so because of what it does, how it is used, and the consequences it has for people. An algorithm that discriminates is not necessarily a poorly programmed algorithm; it may be an algorithm that has perfectly learned the biases of the society that trained it. The ethics of artificial intelligence is not an optional add-on; it is a requirement of justice.”

There is a naive belief that algorithms are objective because they are based on mathematics, and mathematics has no ideology. This belief is dangerously false. Algorithms learn from data, and data reflect society with all its biases, inequalities, and injustices. An algorithm trained on historical data will learn historical practices, including those we now consider discriminatory. It will learn that women were treated one way and men another, that ethnic minorities were offered fewer options, and that the poor were allocated lower-quality resources. It then reproduces and amplifies those biases, presenting them as if they were objective mathematical truths.

The history of artificial intelligence in medicine is full of examples of algorithms that have been discriminated. Algorithms that underestimated pain in Black patients because the training data reflected the erroneous belief that they feel less pain. Algorithms that recommend fewer cardiac tests for women because the symptoms of heart attacks in women are underrepresented in the literature. The algorithms allocated fewer resources to poor neighborhoods because the historical data revealed that less resources were spent there. These are not technical failures; they are ethical failures with real consequences for people's health.

In this chapter, we explore the fundamental ethical aspects of medical software development. We will begin with the most basic question: Does your software work the same for men as it does for women? We will continue with socioeconomic bias, which is often invisible but deeply ingrained in the data. We address informed consent for the use of secondary data, an increasingly relevant topic in the era of big data. Finally, we discuss legal liability and liability insurance—uncomfortable but necessary questions about who is responsible when an algorithm causes harm. This is not a treatise on philosophical ethics but a practical guide to developing software that does not perpetuate injustices.

16.1 DOES YOUR SOFTWARE WORK THE SAME FOR MEN AS IT DOES FOR WOMEN?

“The most dangerous question is not whether your algorithm is biased but whether you have the courage to ask it out loud.”

At first glance, the question seems almost naive in its simplicity: does your software work the same for men as it does for women? It could be interpreted as a minor technical curiosity, a methodological detail that developers resolve with a quick statistical check before moving on to more pressing matters. However, behind this apparent simplicity lies one of the deepest and least explored chasms in medical software development. It is not enough to answer yes or no; one must systematically demonstrate that the answer is true; one must understand with biological and social precision why differences emerge when they exist; and one must be prepared, morally and technically, to act decisively when the answer, as is often the case, is that the algorithm works better for one group than for another. The uncomfortable reality is that for most current medical algorithms, especially those developed without conscious attention to equity, the answer is no—even if their creators, trapped in the blindness of their own data, ignore this uncomfortable truth.

The training data that feed machine learning algorithms are often deeply unbalanced in terms of gender representation, reflecting historical biases in medical research that for decades considered the male body the universal norm and the female body a deviant variant. Cardiovascular diseases, for ex, were studied predominantly in men, ignoring the fact that they manifest qualitatively differently in women. The symptoms that algorithms learn to recognize as indicators of acute myocardial infarction—severe chest pain, radiation to the left arm, and profuse sweating—are typically male, whereas the most common manifestations in women—insidious fatigue, vague nausea, and jaw or back pain—remain underrepresented in the training data. Ignoring these differences does not neutralize them; it renders them invisible, making algorithmic bias more difficult to detect and correct because it is not even actively sought out.

16.1.1 THE FOUR BARRIERS TO GENDER BIAS

The first frontier that hinders an honest assessment of performance by gender is the

availability and quality of disaggregated data. To determine whether an algorithm performs equitably, we need sufficiently large samples of men and women, with adequate clinical variability, and reliably labeled for diagnosis. This combination is surprisingly rare. In many historical datasets, women are systematically underrepresented, especially in specialties such as interventional cardiology, vascular neurology, or critical care pulmonology. The large clinical trials that generated reference data for training modern algorithms historically excluded women of childbearing age and underrepresented older women. When these biased data feed algorithms, underrepresentation becomes algorithmic invisibility: the system simply has not seen enough female cases to learn reliable patterns, yet it operates with undifferentiated confidence across all patients.

The second frontier goes beyond mere data quantity to address qualitative differences in disease manifestation. Pathologies are not abstract biological entities that affect all human organisms equally; they are expressed through hormonal, metabolic, and sociocultural filters that differentiate men and women. Acute myocardial infarction, again the paradigmatic example, presents different symptoms in approximately one-third of female cases compared with male cases. An algorithm trained predominantly on male data learns a presentation pattern that is literally dangerous when applied to women, generating false negatives that can cost lives. It is not that the algorithm is consciously sexist; it has learned an incomplete and historically androcentric version of clinical reality.

The third frontier is more insidious because it operates on the “fundamental truth” we use for training: the diagnostic labels issued by human clinicians who themselves operated under historical biases. If, for decades, women were underdiagnosed for certain diseases—because their symptoms did not fit standard male patterns and because their complaints were attributed to psychosomatic causes because they underwent fewer confirmatory tests—then the training data reflect this historical underdiagnosis as if it were biological variation. The algorithm learns that women are less likely to suffer from certain conditions, not because it is true but because that is how it was recorded in medical history. This vicious cycle is particularly difficult to break because the reference “truth” is contaminated at its source, and the algorithm, far from correcting the bias, systematizes and amplifies it by applying it with the apparent neutrality of mathematics.

The fourth frontier recognizes that gender identities do not operate in isolation but intersect with other demographic and structural variables. Being a woman is not the same as being a woman over eighty, a woman of African descent, or a woman with limited financial resources that hinder access to preventive care. Inequalities intersect and reinforce one another in complex configurations that a simple binary men–women analysis obscures. An algorithm that appears to work reasonably well “on average” for women may fail catastrophically for Black women, rural women, or women with disabilities. Intersectional analyses, although they require even larger samples and more sophisticated methodologies, are indispensable for a comprehensive ethical evaluation of algorithmic performance.

16.1.2 CONCRETE ACTIONS: FROM DIAGNOSIS TO CORRECTION

For medical software developers, addressing gender equity is not an exercise in political correctness but a technical and ethical obligation that requires specific and verifiable actions. The first step is the active curation of training data to ensure demographic representativeness that reflects the population to which the algorithm will be applied, with oversampling of historically underrepresented groups when necessary and with transparent documentation of limitations when full representativeness is not achievable. This means rejecting the convenience of readily available public datasets to actively seek out more diverse sources, even if this increases development costs and complexity.

The second step is the explicit and systematic evaluation of performance broken down by sex/gender, reported publicly regardless of the results. Developers must calculate performance metrics—sensitivity, specificity, predictive value, and calibration—separately for men and women and for intersectional subgroups when sample size permits. These analyses must be included in scientific publications and regulatory documentation, not as optional appendices but as central components of the product’s characterization. Transparency regarding performance disparities, while commercially inconvenient, is ethically mandatory and increasingly expected by regulators.

The third action is causal investigation of differences when they are detected. Not all performance disparities are equally concerning or require the same interventions. Some may be due to remediable biases in the training data, others to real biological differences that the

algorithm should learn to handle, and others to technical design issues such as inadequate feature selection. Diagnosing the etiology of the bias is a prerequisite for designing an appropriate remedy. If the problem is a lack of female data, the solution is additional data collection; if it is a qualitative difference in manifestation, the solution may be specific feature engineering; if it is biased in historical labels, the solution requires learning strategies that are more robust to systematic noise.

The fourth course of action, when differences are substantial and persistent, is to consider developing algorithms specific to population subgroups or at least explicitly incorporating sex/gender as an input variable rather than expecting the model to implicitly learn these differences. Sex-specific algorithms, although controversial because of their apparent essentialism, can save lives when disease manifestations are qualitatively different. Alternatively, modeling architectures that treat sex/gender as an explicit covariable may allow the algorithm to adjust its predictions in a differentiated manner without requiring completely separate systems.

16.1.3 CLINICAL RESPONSIBILITY AND THE FUTURE OF ALGORITHMIC EQUITY

For clinicians who use or are considering adopting medical software, a critical assessment of gender equity must be a standard part of their due diligence. Before they rely on a decision support system, they should demand documented answers to specific questions: In which population was the algorithm trained and validated? Was performance explicitly analyzed by sex/gender? Are there known differences, and how have they been addressed? If the manufacturer cannot provide this information or if the evidence suggests unresolved disparities, this should be a cause for concern and potentially for rejection. In daily clinical practice, clinicians must maintain active vigilance regarding potential discrepancies: if the algorithm appears to work well in certain patients but generates inappropriate alerts or dangerous false negatives in others, these patterns should be systematically reported as warning signs of algorithmic bias.

The issue of gender equity in medical algorithms is, of course, the tip of a much larger iceberg that includes age, ethnicity, socioeconomic status, sexual orientation, gender identity, disability, and residential geography. Each of these dimensions can introduce bias if not explicitly considered in the design and evaluation. However, sex is a particularly valuable starting point because it is universally applicable, because biological and clinical differences are relatively well documented in the medical literature, and because historical biases are particularly evident and well studied. A development team that learns to seriously address gender equity will be better prepared to identify and manage other dimensions of algorithmic distributive justice.

The future of ethically sustainable medical artificial intelligence lies in a design that is conscious of equity from the earliest stages of development, not as a quality control measure added at the end. This means that equity must be an explicit design goal, with quantifiable metrics and acceptable thresholds for disparity, just as accuracy or sensitivity are technical goals. This refers to the proactive selection of representative data, the continuous evaluation of performance across subgroups during deployment, and the organizational commitment to redesign when unacceptable inequalities are detected. This means that the diverse participation of affected communities in development, such as the implicit biases of developers—typically young men from urban areas with a technical education—does not crystallize into products that perpetuate their blind spots. In addition, this means, above all, recognizing that equity is not a state achieved through a single verification but a continuous process of monitoring, auditing, and iterative improvement.

We have reached a conclusion that transcends algorithmic technique: behind every detected or unnoticed bias are real people who suffer concrete consequences, sometimes fatal ones. A woman whose heart attack goes unnoticed because the algorithm fails to recognize her atypical symptoms may die at home because of a delayed diagnosis. A man who receives inadequate doses of medication because dosing studies were conducted predominantly on women may suffer preventable adverse effects. A person of limited means for whom the system assigns lower priority because historical data show that “less is spent there” may see their health compromised by algorithmic decisions that reflect structural inequalities. These are not ethical abstractions; they are concrete clinical injustices that developers have a moral responsibility to prevent. The question “Does your software work the same for men as it does for women?” is not a political or ideological demand; it is a fundamental ethical requirement and an indispensable clinical necessity. Answering it properly, with the methodological rigor it deserves, is one of the most important contributions we can make to a genuinely just 21st-century medicine.

16.2 SOCIOECONOMIC BIAS: EVIDENCE AND MITIGATION

Socioeconomic bias in medical algorithms is perhaps the most insidious form of algorithmic discrimination because it operates in the shadows of apparent technical neutrality. Unlike gender or race, which are relatively visible categories that conscientious developers can actively seek out to assess equity, socioeconomic status is an elusive, multidimensional variable that is often completely absent from the clinical databases that feed machine learning models. However, its influence on health is overwhelming: people from lower socioeconomic backgrounds experience poorer chronic health, less access to preventive services, less favorable clinical outcomes, and a life expectancy that can differ by more than a decade from that of their more privileged fellow citizens. When algorithms learn from data saturated with these structural inequalities, they neither correct nor ignore them; they systematize, amplify, and render them invisible under the guise of mathematical objectivity, creating vicious cycles of discrimination that are particularly difficult to detect and dismantle.

The problem of socioeconomic bias dramatically increased public consciousness, with a study published in 2019 that analyzed a commercial algorithm used by millions of patients in the United States for the allocation of healthcare resources. The system, designed to identify patients with greater medical needs and offer them additional support programs, was designed to perpetuate massive racial discrimination without its creators or users being aware of it. The cause was not an intentionally racist design but the use of a seemingly innocuous proxy variable: historical healthcare costs.⁽¹⁾ The algorithm logically assumed that those who had generated higher medical costs in the past were those with greater current needs. However, because black patients have historically been less often covered because of barriers to access, discrimination in clinical practice, and a lack of insurance, the algorithm systematically learned that black patients were less sick and needed less care when they were just as sick or sicker than their white counterparts with higher documented costs were. This paradigmatic case illustrates how socioeconomic bias can infiltrate algorithms without ever being explicitly named, carried by variables that appear technically neutral but are deeply imbued with structural inequality.

16.2.1 THE TUNNELS OF BIAS: HOW INEQUALITY BECOMES A PROXY

The pathways through which socioeconomic status influences clinical data are multiple, complex, and mutually reinforcing. People of lower socioeconomic status face systematic barriers to accessing healthcare: inadequate transportation, inflexible work schedules that conflict with clinic hours, a lack of insurance coverage, or simply the inability to pay copayments and medications. As a result, their illnesses are diagnosed later, at more advanced stages, and with poorer prognoses. These diagnostic delays do not reflect lower disease severity but rather reduced access to the screening system; however, the algorithm that learns from these data interprets the delayed presentation as lower risk or less need for intervention.

In addition to access, there are differences in the quality of care received. Low-income patients are less likely to receive state-of-the-art treatments, be referred to specialists, or participate in chronic disease management programs. Their doctors may be less experienced, work in settings with higher patient volumes, or have less time for individualized care. Clinical outcomes are worse, but these poorer outcomes are not attributable to patient biology but rather to the unequal organization of care. The algorithm, however, learns that certain patient profiles, correlated with poverty, have worse outcomes and may erroneously infer that they are less likely to benefit from expensive interventions.

Health literacy and communication styles also vary systematically with socioeconomic status. People with less formal education may describe their symptoms differently, using vocabulary that medical coding systems do not adequately capture. They may have less ability to navigate the healthcare system, less persistence in seeking diagnoses, and less trust in medical authority. All these factors are reflected in the data as “lower adherence,” “atypical presentation,” or “poor compliance”—labels that the algorithm learns to recognize as predictors of poor prognosis, without understanding that they are manifestations of structural inequality rather than individual characteristics.

The physical and social environment adds another layer of complexity. Low-income individuals are more likely to live in neighborhoods with environmental pollution, food insecurity, inadequate housing, and exposure to violence. Their jobs are more physically demanding, more precarious, and more stressful. All of these affect their health and are recorded in medical data,

but the algorithm lacks the ability to distinguish between what is due to biological disease and what is due to social inequality. The ZIP code becomes a powerful proxy for health, not because geography determines biology but because geography determines exposure to social determinants of health.

16.2.2 CASE STUDIES: WHEN PREDICTING RISK IS CONFUSED WITH PREDICTING PRIVILEGE

The example of hospital readmission prediction algorithms clearly illustrates how socioeconomic bias can completely reverse the logic of intervention. These systems, widely used in U.S. healthcare systems and expanding in other countries, are designed to identify patients at high risk of readmission within thirty days of discharge, allowing hospitals to allocate preventive transition resources to those who need them most. However, independent studies have shown that these algorithms systematically overestimate the risk of readmission in patients from high socioeconomic backgrounds and underestimate it in those from low socioeconomic backgrounds.

The reason for this paradox is that the algorithm is, in fact, predicting access to the healthcare system rather than actual medical needs. Privileged patients are more likely to be readmitted because they have better access to hospitals, because their primary care physicians are more available to refer them, because they have insurance that covers readmissions, and because the medical culture is more inclined to hospitalize them at the slightest sign of doubt. Low-income patients, on the other hand, may not be readmitted simply because they cannot afford transportation to the hospital, because they fear the costs even if they have formal insurance, because they cannot afford the time off work for recovery, or because they have learned that the system does not meet their needs. The algorithm interprets the absence of readmissions as an absence of risk; in reality, it may indicate abandonment of the system because of insurmountable economic barriers.

The accumulated evidence on socioeconomic bias in medical algorithms, although still limited compared with the literature on sex or racial bias, is consistently concerning. A 2021 systematic review revealed that the vast majority of predictive algorithms in medicine simply do not evaluate performance by socioeconomic status, as if this dimension were irrelevant to equity. When studies do disaggregate data by available socioeconomic indicators, they frequently find significant and substantial differences. Some algorithms are completely useless and perform no better than chance when applied to low-income populations, even though they work reasonably well in privileged populations. Implementing these systems without specific adjustments is not a neutral technological improvement; it is an active way of amplifying existing inequalities, allocating preventive resources to those who already have them and denying them to those who need them most.

16.2.3 DEVELOPMENT CHALLENGES: DETECTING THE INVISIBLE

For medical software developers, addressing socioeconomic bias presents specific challenges that have no purely technical solutions. The first challenge is the lack of direct data. Many historical clinical databases simply do not include information on income, education, occupation, or any direct measure of socioeconomic status. Hospitals do not routinely ask patients how much they earn, and when the information exists in linked administrative records, privacy and regulatory barriers hinder its use for research. Without this basic information, it is technically impossible to assess whether the algorithm performs equitably across the socioeconomic spectrum.

A partial solution is the use of proxy variables such as the home ZIP code, which allows for an approximation of socioeconomic status on the basis of neighborhood characteristics using census data. However, these approximations are crude: a ZIP code can include both mansions and public housing, and assigning average characteristics to all its residents introduces substantial classification errors. Furthermore, using the ZIP code as a proxy can introduce new biases, particularly racial ones, in contexts where residential segregation is pronounced. Another approach is to enrich clinical databases with information from external sources, tax records, household surveys, and data from social programs, but this raises complex issues of privacy, informed consent, and integration logistics that many projects cannot resolve.

The second challenge is the complexity of the causal relationships between socioeconomic status and health. Socioeconomic status is not simply a confounding variable that can be

statistically adjusted for using standard techniques; it is a fundamental social determinant of health that acts through multiple simultaneous mediating pathways. Statistically adjusting for socioeconomic status can obscure important effects or create interpretive paradoxes. For example, if we adjust a model for “access to care,” we may eliminate precisely the variable that explains why certain patients have worse outcomes, erroneously concluding that there is no disparity when in fact we have made it invisible through the adjustment. There are no simple statistical solutions; careful, e-nalysis and explicit theorization of causal mechanisms are needed, as well as frequent interdisciplinary collaboration with social epidemiologists and health economists.

The third challenge is intersectionality—the recognition that socioeconomic status interacts with other dimensions of inequality in ways that are not additive but multiplicative. Low-income women face different and potentially greater barriers than men do in the same economic situation, including unpaid caregiving responsibilities, employment discrimination, and gender-based violence. Low-income immigrants face additional challenges related to language, documentation, and legal status. Poor older adults face social isolation and increasing dependence. Analyses that consider only one inequality variable at a time miss these crucial interactions, and evaluation designs that capture intersections require massive samples and sophisticated methodologies that are rarely available.

16.2.4 MITIGATION STRATEGIES: FROM AWARENESS TO ACTION

Despite these difficulties, there are concrete measures that development teams can implement to mitigate socioeconomic bias, although none offer a complete guarantee. The first step is, as always, awareness of the problem as an active dimension of design. Many medical software development teams, dominated by technical and biological health science profiles, simply do not consider social determinants relevant to their work. They implicitly assume that their data are neutral records of biological reality and that any pattern the algorithm detects is true pathology rather than a social artifact. Basic training in social epidemiology, health determinants, and the history of health inequalities should be a standard component of any medical software developer’s curriculum—not as a humanities elective but as an essential technical skill.

The second measure is the explicit evaluation of performance across different socioeconomic groups, using available variables even if they are imperfect. If ZIP codes are available, performance should be analyzed by neighborhood income quintiles. If the insurance type is known, performance should be compared between public and private systems or between different coverage levels. If educational information is available, performance should be evaluated by educational attainment. When significant differences are detected, investigating their causes must be a priority: Are they due to biases in the training data that underrepresent certain groups? To real differences in disease manifestation that require specific models? To technical design issues such as inadequate feature selection? Without this diagnostic analysis, any attempt at correction is blind.

The third step is to consider the use of bias-correction techniques that have been developed in the fair machine learning literature. Rebalancing training samples to overrepresent historically underserved groups can improve performance for these groups, albeit at the expense of overall accuracy. The explicit inclusion of socioeconomic variables in the model—far from being eliminated for the sake of “neutrality”—can allow the algorithm to adjust its predictions in a differentiated manner. Adjusting decision thresholds by group, by setting different cutoff points for risk classification on the basis of socioeconomic context, can equalize false positive and false negative rates across groups. These techniques have known limitations; they may introduce new biases, require access to protected variables that are not always available, and raise ethical dilemmas regarding when it is legitimate to treat different groups differently, but they are preferable to inaction, which perpetuates the discriminatory status quo.

The fourth measure is the genuine participation of diverse communities in the development and evaluation of the software. People with direct experience with the inequalities the algorithm aims to address can identify problems, assumptions, and blind spots that developers in positions of privilege do not foresee. Usability and acceptability studies must deliberately include patients from different socioeconomic backgrounds, not just those who are the easiest to recruit in university or private settings. Patient advisory boards must include voices from historically marginalized communities. This participation is not a symbolic gesture of inclusion; it is a mechanism for detecting biases that purely technical methods cannot capture.

16.2.5 THE CLINICIAN'S RESPONSIBILITY AND THE ETHICAL BOUNDARY

For clinicians who use algorithms in their daily practice, socioeconomic bias is an additional reason to maintain a critical and active stance. When an algorithm recommends fewer resources for a patient from a poor neighborhood or classifies their risk as low when clinical intuition suggests otherwise, the physician must ask the following: Does this recommendation reflect actual medical needs, or does it reflect historical patterns of neglect toward this type of patient? Is the algorithm considering the specific barriers this individual faces in accessing care? If the answer is no, the clinician should feel justified—indeed, obligated—to disregard the algorithmic recommendation and rely on their professional judgment, explicitly documenting the reasons for the discrepancy to inform system improvements.

The future of the fight against socioeconomic bias in medical algorithms lies in greater integration of social determinants into routine clinical data. Initiatives such as the inclusion of Z codes in the International Classification of Diseases (ICD-11) for social factors influencing health—including issues related to housing, employment, education, and the social environment—are advancing slowly. Public health information systems are developing the ability to link with social registries, enabling more comprehensive analyses.⁽²⁾ Statistical methods for detecting and correcting biases in machine learning continue to evolve, with adversarial learning and causal modeling techniques promising greater sophistication.

However, the most fundamental change is cultural: the widespread recognition that health is not a purely biological phenomenon but a biopsychosocial one and that algorithms that ignore this multidimensional reality are inherently incomplete and potentially harmful. Developers have an inescapable ethical responsibility not to exacerbate existing inequalities but not to create systems that amplify structural discrimination. However, the responsibility to build more just societies and to eliminate the roots of poverty and exclusion that biased data generate in the first place is collective and political. Algorithms can make inequalities visible, they can monitor their evolution, and they can even mitigate their effects to some extent. However, they cannot, on their own, transform the power structures that produce them. That requires bold public policies, sustained investment in social services, and an organized fight against poverty—tasks that transcend software engineering but must constantly inform it.

If we have internalized the gravity of socioeconomic bias and the responsibility to mitigate it, the question inevitably arises of how these efforts toward algorithmic justice align with the rights of the people whose data we are using. How do we ensure that the use of data to detect and correct inequalities does not violate the autonomy of those who never explicitly consent to their medical records being used to train algorithms that will eventually make decisions about the care of others?

16.3 INFORMED CONSENT FOR THE USE OF SECONDARY DATA

The relationship between patients and the information they generate about their own bodies constitutes one of the most intimate and simultaneously most opaque aspects of contemporary medicine. Every consultation, every diagnostic test, every prescription, and every MRI image leaves a digital trail that is recorded in hospital information systems, stored on servers, potentially shared among institutions, and eventually used for purposes that go far beyond the individual care that prompted its initial collection.

This information is undoubtedly essential for the continuity of care for the patient who generated it, but it is also extraordinarily valuable for other purposes: epidemiological research, the development of predictive algorithms, the improvement of quality protocols, the planning of population-based healthcare resources, and even, in commercial contexts, the generation of economic profits for private companies. The conflict between the patient's fundamental right to control their personal information and the social or commercial interest in using it to advance knowledge or profit represents one of the most complex and least resolved ethical dilemmas of the digital era in healthcare.

Informed consent has traditionally been the preferred ethical and legal tool for managing this conflict. The logic is seemingly simple: before a person's data can be used for any purpose other than direct and immediate medical care, the individual must be informed of what is proposed to be done with their information and what the risks and benefits of that use are, and explicit, voluntary, and revocable authorization must be obtained. However, this conceptual simplicity quickly fades when we confront the realities of big data and medical artificial intelligence. How

do we inform a patient about future uses of their data when even the researchers themselves cannot foresee what scientific questions will be asked five or ten years from now? How do we obtain consent for algorithms that do not yet exist, trained with technologies that have not yet been invented, applied to clinical contexts we cannot yet imagine? How do we manage consent when data flow across multiple institutions, national borders, and cloud ecosystems where the physical location of the information becomes irrelevant?

16.3.1 BEYOND SPECIFIC CONSENT: ALTERNATIVE MODELS FOR THE BIG DATA ERA

Traditional consent, designed for clinical research studies with defined purposes, static protocols, and limited timeframes, proves profoundly inadequate for big data research where uses are emerging, analyses are iterative, and time horizons are indefinite. In light of this inadequacy, alternative approaches have emerged that attempt to balance patient autonomy with the feasibility of large-scale research, although none fully resolve the inherent tension.

Broad consent, also known as “broadband consent,” represents the most widespread approach currently used in biobanks and large population databases. In this model, the patient authorizes the use of their data for a broad, predefined category of future research, provided that such research complies with certain general ethical and institutional governance conditions. The patient does not know the specific uses to which their data will be put but trusts in an oversight framework that will ensure that any subsequent research is ethical and scientifically valid. This approach has the virtue of practical feasibility—it allows for secondary research without requiring repeated contact with millions of participants—but it raises fundamental questions about whether it genuinely meets the “informed” standard of ethical consent. Can consent that authorizes unknown uses be considered informed? Is it legitimate to waive the right to know the specific purposes of the research in advance?

Dynamic consent has emerged as a response to these concerns, using digital platforms to maintain an ongoing, two-way relationship between researchers and participants. Rather than a single, static act of consent, patients access digital portals where they can manage their preferences over time, receive notifications about specific new projects that use their data, and actively decide whether they wish to participate in each particular initiative. This approach more faithfully respects patient autonomy and the right to information but faces challenges in large-scale implementation: it requires sophisticated technological infrastructure, digital literacy among participants, and institutional organization capable of managing individualized preferences at a population scale. Furthermore, the risk of “consent fatigue,” where participants disengage because of the impossibility of continuously processing new requests, can erode the protection it aims to provide.

A third approach, more controversial but operationally necessary in certain contexts, is the waiver of consent authorized by research ethics committees. When a retrospective study uses large databases where contacting all patients would be practically unfeasible and the risks to participants are minimal because the data are adequately anonymized, ethics committees may authorize the use of data without explicit individual consent. This exemption is contingent on the study having significant social value, as there being no known prior objections from patients, and robust privacy safeguards being implemented. However, this practice is ethically contested, with legitimate arguments that consent should always be required even if it is impractical and that the convenience of researchers should not take precedence over the autonomy of participants.

16.3.2 THE COMMERCIAL FRONTIER: WHEN RESEARCH BECOMES A PRODUCT

In the specific context of medical software development, the issue of consent takes on additional layers of complexity because data are frequently used not only for academic research aimed at advancing public knowledge but also to build commercial products that will subsequently be sold to healthcare institutions. A patient may be genuinely willing to share their cancer data for university research aimed at improving diagnosis but not for a private company to develop an algorithm that it will later market, with profits that will not benefit the patient or their community. The distinction between research and commercial development is ethically relevant; it implies different distributions of benefits, different levels of public oversight, and different levels of exploitation, but this distinction is often blurred in current practice.

Many patients who generously authorized the use of their data for “medical research”

discover with dismay, years later, that their information fed algorithms owned by corporations that are now reaping millions in profits. This sense of betrayal—even though broad consent forms might legally cover these uses—erodes public trust in medical research and may cause future patients to refuse to participate. Medical software developers face the ethical responsibility to explicitly clarify, in the consent processes that feed their databases, when data may be used for commercial purposes and what the implications of such use are.

The issue of anonymization repeatedly emerges as a justification for consent waivers or less formal consent procedures. The promise that “the data are anonymous” is used to reassure patients and regulators about the risks of secondary use. However, perfect anonymization is technically increasingly difficult to achieve in the world of big data. Demonstrative studies have shown that combining relatively few seemingly innocuous demographic data—zip code, date of birth, and sex—allows for the unique reidentification of a surprisingly high proportion of individuals in “anonymous” databases. When these databases are combined with other commercially or publicly available sources of information, reidentification becomes nearly inevitable. Patients have a right to know that the promise of anonymity may be technically untenable and that “anonymous data” does not necessarily equate to “data without a risk of identification.”

16.3.3 PRACTICAL RESPONSIBILITIES: FROM CODE TO ETH

For medical software developers, these abstract ethical issues have immediate and nonnegotiable practical implications. The first is the obligation of due diligence regarding the origin of the data used to train algorithms. It cannot be assumed that data acquired from third parties, or even generated internally, are free of ethical and legal restrictions. It is necessary to systematically verify the consent obtained at the source, authorizations from ethics committees, specified usage limitations, and the expiration dates of authorizations. Ignoring these aspects is not only an ethical failure; it is a potentially catastrophic legal exposure in a data protection regulatory environment that imposes multimillion-dollar penalties for noncompliance, as demonstrated by recent fines under the European GDPR.

The second implication is the imperative for transparency regarding the current use of data. When a patient interacts with a healthcare system that uses their data to improve algorithms—whether through implicit feedback on its use or through explicit analysis of its results—they should be clearly informed of this secondary processing. When a healthcare institution shares data with an external developer, there should be a clear contractual agreement specifying permitted purposes, explicit prohibitions, audit mechanisms, and the consequences of noncompliance. This transparency is not merely a matter of legal compliance; it is a necessary condition for maintaining the public trust that underpins the entire digital health enterprise.

The third technical implication is the design of systems that materially respect patient preferences. If an individual has expressed opposition to their data being used for certain purposes—for example, commercial research or international sharing—the system’s architecture must be capable of verifying the exclusion of their data from the relevant training datasets. This requires sophisticated technical mechanisms for managing consent at scale: database flags, filters in processing pipelines, and audits to verify that exclusions are effectively honored. Implementing these capabilities is no small feat from a software engineering perspective, but it is increasingly expected by regulators and patients alike.

The fourth implication is active participation in the broader social debate on health data governance. Decisions regarding acceptable uses of medical data cannot be delegated exclusively to technologists, regulators, or markets. Developers, as actors with expert knowledge of technological capabilities and limitations, have a responsibility to contribute to the collective reflection on which applications are socially desirable and which should be restricted. They also have a responsibility for public education: to help patients and citizens understand, in accessible yet honest terms, the potential benefits and real risks of using big data in medicine so that decisions regarding consent are genuinely informed and not based on myths or unfounded fears.

16.3.4 THE ROLE OF THE CLINICIAN AND THE DIMENSION OF JUSTICE

For clinicians operating at the interface between patients and information systems, consent for secondary use of data is a matter of direct and growing responsibility. It is often they who collect the initial data, who explains procedures and their implications to patients, and who obtain

consent signatures on forms that patients often do not read in full. They are also the ones who may see their patients' trust betrayed if patients discover that their data have been used in ways they did not anticipate or would not approve of. That is why it is essential that clinicians deeply understand these issues, be able to explain them in a way that patients can understand during consultations and actively participate in ethics committees and institutional decisions regarding data policies.

The relationship between consent and social equity is particularly relevant and concerning. The most vulnerable patients—those with less formal education, language barriers, a reduced ability to process complex information, or less power to negotiate with medical institutions—are precisely those most likely to have their data used without truly informed consent. They are also those who stand to benefit most from research advances but who may suffer the most from its risks if the research perpetuates inequalities. Ensuring that consent is truly informed for everyone—not just for privileged patients who can afford lawyers and advisors—is a requirement of distributive justice that health systems must explicitly address through the design of accessible informational materials, the use of interpreters, and adequate time for deliberation.

Consent management technologies, natively integrated into electronic health record systems, will allow individuals to set detailed preferences regarding which specific uses they authorize: academic research, commercial research, no domestic use, no international use, no cancer research, and no mental health research. Citizen research platforms will transform patients from passive providers of data into active participants in the design and direction of research projects. Shared governance models, where representatives of patients, researchers, developers, and regulators collaborate on decision-making regarding data use, are gaining ground as alternatives to technical dictatorship or pure bureaucracy.

However, if we have established that medical artificial intelligence must be ethical, responsible, and fair in the distribution of risks, the pragmatic question arises of how these imperatives translate into viable business models. How do we monetize an algorithm that saves lives without resorting to the exploitation of sensitive data or creating inequalities in access? How do we build economic ecosystems where innovation is rewarded but social benefit comes first?

REFERENCES

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* (80-) [Internet]. October 25, 2019 [cited May 15, 2026];366(6464):447-53. Available at: [/doi/pdf/10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342)

Hau C, Grubber JM, Ferguson RE, Cushman WC, Ishani A, Glassman PA, et al. Social Determinants of Health ICD-10 Code Use by a Large Integrated Healthcare System. *Healthc* [Internet]. November 1, 2025 [cited May 15, 2026];13(21):2710. Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC12608873/>