Review Article
Laboratory Informatics

Ann Lab Med 2025;45:12-21
https://doi.org/10.3343/alm.2024.0323
ISSN 2234-3806 • eISSN 2234-3814

ANNALS OF LABORATORY MEDICINE

Laboratory Data as a Potential Source of Bias in Healthcare Artificial Intelligence and Machine Learning Models

Hung S. Luu, M.D.

Department of Pathology, UT Southwestern Medical Center, Dallas, TX, USA

Artificial intelligence (AI) and machine learning (ML) are anticipated to transform the practice of medicine. As one of the largest sources of digital data in healthcare, laboratory results can strongly influence AI and ML algorithms that require large sets of healthcare data for training. Embedded bias introduced into AI and ML models not only has disastrous consequences for quality of care but also may perpetuate and exacerbate health disparities. The lack of test harmonization, which is defined as the ability to produce comparable results and the same interpretation irrespective of the method or instrument platform used to produce the result, may introduce aggregation bias into algorithms with potential adverse outcomes for patients. Limited interoperability of laboratory results at the technical, syntactic, semantic, and organizational levels is a source of embedded bias that limits the accuracy and generalizability of algorithmic models. Population-specific issues, such as inadequate representation in clinical trials and inaccurate race attribution, not only affect the interpretation of laboratory results but also may perpetuate erroneous conclusions based on AI and ML models in the healthcare literature.

Key Words: Aggregation bias, Artificial intelligence, Clinical pathology, Diagnostic error, Health information interoperability, Logical Observation Identifiers Names and Codes, Machine learning, SNOMED CT

Received: June 26, 2024
Revision received: September 10, 2024
Accepted: October 18, 2024
Published online: October 24, 2024

Corresponding author:
Hung S. Luu, M.D.
Department of Pathology, UT Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390, USA
E-mail: Hung.Luu@utsouthwestern.edu

INTRODUCTION

Artificial intelligence (AI) refers to a variety of computational methods used to enable machines to model and mimic human judgment and perform tasks such as perception, reasoning, learning, and decision-making [1, 2]. Past AI and clinical support systems carried out only those tasks for which they were coded or trained using human knowledge that was encoded explicitly as rules in a knowledge base or relationships in an ontology to derive conclusions about input data using these rules or relationships. In contrast, modern AI and machine learning (ML) are not solely based on human-derived knowledge and use an algorithm to identify ("learn") repetitive data patterns present in example cases and match new cases to the patterns identified with human guidance [1].

ML algorithms are classified into three types according to the learning strategy: (1) supervised learning, (2) unsupervised learning, and (3) reinforcement learning [1]. Supervised learning is learning by example and requires labeled data points. Examples of supervised ML algorithms likely to be encountered in pathology are (a) classification, in which an algorithm predicts a label (e.g., a diagnostic category) given a set of input features

(e.g., an image), and (b) regression, in which an algorithm predicts a numerical value (e.g., the likelihood of tumor metastases) given input variables [3]. Unsupervised learning techniques do not use predefined labels or values; they attempt to identify patterns in unlabeled data and employ those patterns without labels [3]. For example, unsupervised learning in pathology can be applied to develop a new classification of mature B-cell neoplasms based on immunophenotype, gene expression profiles, or outcomes. Reinforcement learning awards points for progress toward a goal and provides tools to optimize sequences of decisions for long-term outcomes [1, 4]. For example, reinforcement learning in the practice of medicine can be applied to find the optimal sequence of diagnostic tests and treatment interventions for the diagnosis and management of neonatal patients presenting with fever of unknown origin. Reinforcement learning is currently rarely applied to pathology [5].

All ML techniques require large datasets. Electronic health records (EHR), electronic administrative records, and cancer registry data have all been used as datasets to predict survival in cancer patients using ML [6, 7]. ML techniques are used to optimize clinician workflow, predict outcomes, and discover insights from large medical datasets [2]. As the use of ML in healthcare increases, the sources and collection methods of the data upon which these algorithms are built must be carefully assessed. A large part of those data is likely to be laboratory data. In a study examining the frequency of laboratory tests ordered during patient encounters, 98%, 56%, and 29% of inpatient, emergency department, and outpatient populations, respectively, had one or more laboratory tests ordered [8]. Thus, biases in laboratory data can influence clinical decision-support tools based on AI and ML algorithms.

In this review, possible sources of bias resulting from the use of laboratory datasets in the training of AI/ML models are discussed. First, we explore the lack of harmonization of laboratory results across platforms and methods as a potential source of bias in AI/ML algorithms. Next, we discuss barriers at all levels of interoperability (e.g., technical, syntactic, semantic, and organizational aspects of interoperability) for laboratory data, along with potential solutions. Finally, we review population-specific issues related to laboratory data and their effects on AI/ML models.

LACK OF HARMONIZATION OF LABORATORY TESTS AND THE EFFECT THEREOF ON AI/ML MODELS

Clinical decisions are increasingly based on clinical guidelines that use a fixed laboratory test result value for treatment decisions [9, 10]. While patients, clinicians, and other healthcare stakeholders generally assume that results from different laboratories are comparable for a given test, different analytical methods and instruments can produce significantly different results [11, 12]. Laboratory test harmonization is defined as the ability to achieve the same result (within clinically acceptable limits) and the same interpretation (e.g., normal vs. abnormal) irrespective of the method or instrument platform used to produce the result [9, 11].

Application of cutoff values to non-harmonized test methodologies

Clinical studies may involve the use of a single method in a central laboratory; however, guidelines resulting from such studies cannot be effectively implemented until all other methods are harmonized with the central laboratory procedure [9]. Incorporating compromised clinical guidelines into AI models that apply to patients with noncomparable test results may lead to inappropriate or incorrect treatments, as demonstrated below.

Serum albumin is a widely used biomarker in clinical nephrology guidelines, and serum albumin cutoff values are used to define disease, assess risk, and guide patient care [13]. A decreased albumin level is a diagnostic criterion for nephrotic syndrome [14] and is used to select patients who might benefit the most from adapted hemodialysis modalities [15]. Patients with nephrotic syndrome have an increased risk for venous and arterial thromboembolic events when serum albumin levels decrease, and clinical guidelines advise the initiation of anticoagulation therapy in patients with nephrotic syndrome when the serum albumin level falls <25 g/L and its discontinuation when the serum albumin concentration recovers to >30 g/L [13].

Two types of laboratory methods are available for measuring albumin: colorimetric methods, in which a dye (bromocresol green or bromocresol purple) complexes with albumin, and immunological methods (turbidimetry and nephelometry), which are based on antibody binding to albumin [16]. van Schrojenstein Lantman, *et al*. [13] found that the between-laboratory variation of bromocresol green, bromocresol purple, turbidimetry, and nephelometry assays for albumin results at ~35 g/L was 3.2%, 2.9%, 4.2%, and 5.7%, respectively, reflecting differ-

ences among instruments, calibration protocols, and local performance.

Comparative studies have shown that the bromocresol green assay overestimates albumin concentrations. The mean bias of the bromocresol green assay versus immunoassays ranged from 1.5% to 13.9%, with a larger bias in patients with low albumin concentrations [16]. Using immunonephelometry as a reference method and the clinical guideline of an albumin concentration of 25 g/L as a cutoff, researchers have calculated that using the bromocresol green assay could result in 21%–59% of patients with membranous nephropathy and nephrotic syndrome not receiving the appropriate anticoagulant therapy [16].

Even with knowledge of the lack of harmonization among instrument platforms and methods, clinicians can do little to account for these variations. The absolute amount of divergence of the bromocresol green assay, when compared with other methods, varied from 0.2 g/L to 11 g/L in individual patients, making correction impossible [16]. In addition, published studies seldom disclose details on the assays used for laboratory measurements (Table 1), rendering it difficult for providers, and by extension, AI models, to evaluate whether a clinical guideline cutoff value is applicable and appropriate for a particular patient test result.

A recent application of ML to laboratory medicine was the development of an ML model to predict abnormal parathyroid hormone-related peptide results using a dataset of 1,330 patients from a single institution [17]. The model achieved an area under the receiver operating characteristic curve (AUROC) of 0.936 and a specificity of 0.842 at 0.900 sensitivity in the development cohort. When the model was directly applied to two external datasets, the AUROC significantly deteriorated to 0.838 and 0.737. The maximum mean discrepancy, which was calculated to quantify the shift of data distributions across different datasets, was larger for the second dataset, indicating a greater data shift compared with that in the original dataset. Retraining the model using the two external datasets improved the AUROC to 0.891 and 0.837, respectively. When external data are insufficient for retraining, a fine-tuning strategy with varying amounts of available external data also improves model utility (Table 1) [17].

Aggregation of non-harmonized test methodologies

Clinical studies involving multiple laboratories using different methods produce data that cannot be aggregated to develop guidelines until the results from the different instruments and methods are harmonized [9]. Jacobsen, *et al*. compared the accuracy of three commercial home-use capillary hemoglobin A1c (HbA1c) tests based on measurements of paired HbA1c venous samples in central reference laboratories and reported that none of these tests achieved the National Glycohemoglobin Standardization Program goal of ≥90% measurements within 5% of a Diabetes Control and Complications Trial venous reference [18]. While the difference in accuracy may be attributed to differences in the sample type (e.g., venous vs. capillary blood) and train models against venous blood versus capillary samples, another study demonstrated that capillary blood collection kits suitable for home use produced results that were comparable with those obtained from venous specimens when tested at central reference laboratories [19]. The difference in accuracy was, therefore, entirely attributable to the instrumentation and method, not the sample type.

In a study simulating the aggregation of blood glucose measurements from diverse sources, the ability to classify a patient's current status decreased as the reliability of measurements decreased; however, the ability to classify a change between two values as significant or irrelevant decreased the most [20]. Attempts to train ML models using clinical data with laboratory results derived from different methods and instruments are expected to face the same obstacles, which is further compounded by the fact that instrumentation and method data are not required and rarely included in EHRs and reported with test results [21, 22]. The International Consortium for Harmonization of Clinical Laboratory Results (ICHCLR) is an international organization that serves to prioritize measurands by medical importance and to coordinate the work of different organizations to avoid duplication [9]. According to the website maintained by the ICHCLR, 70 of the 143 analytes listed have achieved a harmonization status of adequate/maintain [23]. Harmonizing a greater number of clinical laboratory tests will contribute to improved healthcare and ensure that AI and ML models that call for the use of laboratory tests can be appropriately developed and implemented (Table 1).

LACK OF INTEROPERABILITY OF LABORATORY DATA AND THE EFFECT THEREOF ON AI/ML MODELS

Laboratory data provide an ideal dataset for AI and ML models in healthcare because they are generally electronically available, highly structured, quality-controlled, and indicative of many diseases [22]. However, the standardization of the data produced by laboratories is lacking, resulting in variations in local test

https://doi.org/10.3343/alm.2024.0323

Table 1. Current and desired states of laboratory data and AI and ML modeling

Source of bias	Current state	Desired state	Ongoing initiatives	Gaps/opportunities
Application of cutoff values to non-harmonized test methodologies	Methods used to derive cutoff values in clinical practice guidelines are seldom reported	Methods used to derive cutoff values are disclosed and transparent to allow applicability determination		Journals should require reporting of instrumentation and methods used in clinical trials and clinical practice guidelines
Lack of harmonization of laboratory tests	Only a subset of laboratory tests has been harmonized by manufacturers to produce comparable results across platforms	The majority of *in-vitro* diagnostic tests have undergone harmonization by the manufacturer	International Consortium for Harmonization of Clinical Laboratory Results	Increase reference material available for harmonization efforts
Instrument and method not reported with laboratory results	Laboratory results do not include the method or instrument used to derive the results, limiting result comparability evaluation of non-harmonized tests	Laboratory results are encoded with instrument and reagent kit identifier	SHIELD	Dissemination and uptake of standard ontology recommendations from SHIELD; EHR and laboratory system functionality to support standard ontology recommendations
Lack of standardization in the digital representation of laboratory results	Variability in how tests are named and results are reported	Accurate representation of laboratory test results using standard ontologies	SHIELD	Dissemination and uptake of standard ontology recommendations from SHIELD; EHR and laboratory system functionality to support standard ontology recommendations
Degradation of Artificial Intelligence Models when applied to different datasets and changing data representation	AI and ML models may lack generalizability when applied to a setting with different data representations and different result values due to the use of non-harmonized tests	Local evaluations are conducted to ensure that the model performs as expected upon retraining or fine-tuning the model with local data as needed, along with performance monitoring over time	CHAI	Dissemination and uptake of recommendations
Insufficient representation of women and minority populations in datasets and clinical trial results	Clinical trials do not routinely report results by sex, race, or ethnicity	Datasets include sex, race, or ethnicity in outcome reports and as a covariate in statistical analysis	NIH Revitalization Act of 1993	Increased compliance with National Institutes of Health policies
Bias in AI and ML models because of erroneous race data collected and conclusions inferred in the healthcare literature	Predictive models using race are vulnerable to bias that may perpetuate health disparities because of inaccurate race representation and sufficient data on minorities in datasets	Developers are mindful of potentially erroneous race data collection and inaccurately inferred conclusions and employ the concept of "counterfactual fairness" to ensure models do not unfairly disadvantage minority populations	The Alan Turing Institute Counterfactual Fairness Project	Dissemination and uptake of recommendations

Abbreviations: AI: artificial intelligence; ML: machine learning; SHIELD: Systemic Harmonization and Interoperability Enhancement for Laboratory Data; CHAI: Coalition for Health AI; NIH: National Institutes of Health; EHR, electronic health record.

https://doi.org/10.3343/alm.2024.0323

ANNALS OF LABORATORY MEDICINE

codes, test names, reference ranges for normal and abnormal values, and formats of test results and associated units [24-26]. Interoperability is broadly defined as the ability of two or more systems or components to exchange information and to use the information that has been exchanged. It is further defined by different components or levels of interoperability that distinguish between lower-level technical components and higher-level organizational components (e.g., technical, syntactic, semantic, and organizational aspects of interoperability.) [27]. In practice, interoperability is determined by a combination of health information technology product features chosen by the vendor (e.g., default settings) and implementation choices made by the institution (e.g., how to represent a particular data element or legacy data) [28]. The interoperability of laboratory data has implications for the accuracy and applicability of AI and ML models derived from the data.

Technical interoperability

Technical interoperability can be defined as the ability to ensure basic data exchange among systems [27]. A national survey of independent and hospital laboratories conducted in the United States in 2013 showed that 71% of hospital-based clinical laboratories and 80% of larger clinical laboratories were capable of sending test results electronically, compared with 58% of independent and 48% of smaller clinical laboratories, respectively. While 62% of clinical laboratories were capable of sending test results electronically, only 30% were doing so for ≥75% of their test results [29]. Improved exchange of patient information underpins many efforts to improve quality, safety, and efficiency in the US healthcare system. Strategies such as the Hospital Readmissions Reduction Program, Accountable Care Organizations, bundled payments, and Medicaid Redesign depend on access to comprehensive and timely patient information [30].

While health information exchange (HIE) connectivity among organizations has greatly improved since 2011 [31], the 2019 American Hospital Association Information Technology Supplement Survey revealed a gap between the adoption and actual use of different HIE models; the availability of different HIE methods (i.e., one-to-one and many-to-many exchanges) did not necessarily guarantee optimal interoperability among affiliated and unaffiliated hospitals. Hospitals reported a lack of technical capabilities to electronically send patient information (9%), cumbersome workflow (23%), and lack of provider technical capability to receive the information despite the adoption of EHRs (59%) as barriers to electronically sending patient health information [32].

The COVID-19 pandemic challenged the US public health information infrastructure and contributed to COVID-19 data inaccuracies and reporting delays because many public health departments still relied on faxes and e-mails to receive reportable disease information [33]. The sheer volume of reporting data during the pandemic exceeded the capacity to manually curate the data [33]. Many reference laboratories still rely on faxes and PDF reports to communicate results to client laboratories. These reports are not easily converted to discrete data and cannot be easily used by AI and ML algorithms. Electronic laboratory data, which can be used by AI and ML, represent only a fraction of the total volume of laboratory results produced and are skewed toward the proportion of the population that has access to hospital-based and larger clinical laboratories.

Syntactic interoperability

Syntactic interoperability refers to the format, structure, and structured exchange of health data, such as laboratory data, and is supported by international standards development organizations, such as Health Level Seven International (HL7) or Integrating the Healthcare Enterprise (IHE), which specify health information technology (IT) standards and their use across systems [27]. While laboratory medicine databases can be a rich source of data, they are often not suitable for the application of data science techniques because of the facts that they are created for regulatory requirements rather than research purposes and store data inefficiently and only for the required retention period [29]. In addition, current health IT systems often operate with a wide variety of data formats and custom specifications [27].

Attribute data must be optimally suited for research as summarized in the Findability, Accessibility, Interoperability, and Reusability (FAIR) Guiding Principles, which provide guidelines to make data findable (F), accessible (A), interoperable (I), and reusable (R) to the research community and emphasize enhancing the ability of machines to automatically find and use the data [34]. However, laboratory data are associated with specific risks and attributes that distinguish them from other data sources. As healthcare data, laboratory data must be subject to special protection. Regulatory requirements preclude freely accessible query functionalities. Structured laboratory data are associated with a risk of unauthorized data duplication because of the relatively small file size. In addition, post-analytics on laboratory data can be difficult as the IT systems of receivers (clinicians or researchers) must be able to handle the data formats supplied and must not alter or falsify their presentation [35].

https://doi.org/10.3343/alm.2024.0323

While laboratory medicine is claimed to be the largest producer of structured data among medical specialties [34], microbiology culture reports form a notable exception [36]. Although microbiology reports are semi-structured and often electronically transmitted, their free-text nature may challenge their incorporation into AI and ML models [36]. Using algorithms on unstructured, non-standardized data can introduce errors that distort test results. These errors are difficult to detect in large datasets owing to the immense volume of data, which complicates the process of anticipating, detecting, and correcting all possible errors [24]. Systematic approaches have been developed for microbiology laboratory result identification and extraction to produce structured data for secondary use [36].

Semantic interoperability

Semantic interoperability is the use of medical terminologies, nomenclatures, and ontologies to ensure that the meaning of medical concepts, such as laboratory data, can be shared across systems [27]. While the lack of standardization in the digital representation of laboratory data does not impact the daily operations of individual laboratories, this inconsistency poses a significant barrier when comparing or aggregating data across institutions. Moreover, this issue creates inefficiencies and ambiguity for secondary uses, such as in AI and ML applications [24]. Two international standard terminologies used for laboratory data are the Logical Observation Identifiers Names and Codes (LOINC) and Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) [10].

Before the development of the LOINC by the Regenstrief Institute—a non-profit medical research organization—in 1994, most laboratories used locally defined codes for each laboratory test [10, 24]. When exchanging laboratory test results electronically using HL7 standards, these local test identifiers were transmitted in the HL7 observation identifier (OBX-3) field to identify the test. Without specific knowledge of the laboratory-generated code by the receiver, the result was often subject to misinterpretation. The LOINC terminology was initially developed to provide universal identifiers for the OBX-3 field of the HL7 observation-reporting message and eliminate ambiguity regarding test orders and results [24].

While mapping local laboratory codes to a standardized terminology such as LOINC aimed to enable semantic interoperability and facilitate data sharing and aggregation across systems, two issues limit the effective use of LOINC to promote laboratory data interoperability. These issues include the limited accuracy of coding and lack of sufficient granularity in codes to impart adequate meaning to results for aggregation and interpretation.

To achieve data aggregation, LOINC must be mapped with absolute precision and accuracy. However, this is complicated by duplicate and overlapping codes, missing codes for some tests, and continuous updates that make accurate coding and code maintenance challenging for laboratories [37]. Lin *et al*. examined the accuracy of LOINC mapping at three large institutions and revealed that choosing different "method," "scale," and "property" attributes were the most common reasons for incongruent code choices [38]. A survey of 90 laboratory participants of the College of American Pathologists coagulation and cardiac markers proficiency testing program showed inaccurate code assignment for nearly 20% of tests [24]. Findings from a study of five medical centers and three test manufacturers revealed a mismatch between how medical centers use LOINC to encode laboratory tests and manufacturer-recommended LOINC coding for the same laboratory tests in 136 (41%) of 331 tests available [25].

An analysis of the accuracy of LOINC coding for 179,537,986 mapped results for 3,029 quantitative tests derived from data extracted from the Patient-Centered Outcome Research Network (PCORnet) revealed a coding accuracy rate of 95.4%, which was higher than the rates in previous studies [39]. The discrepancy could be partly explained by the fact that the study only included quantitative test results, and data were only available from two participants in PCORnet, a research consortium composed of more than 60 organizations known as "DataMarts."

Even when tests are correctly mapped to LOINC codes, the information conveyed by the codes may lack sufficient meaning to allow for accurate test result interpretation because multiple tests can be appropriately assigned to the same LOINC code despite producing very different quantitative result values because of differences in the instruments and methods used to produce the results [40]. The LOINC terminology uses a construct of six independent parts/axes that define the laboratory test performed and reported: analyte, kind of property, time aspect, sample type, precision, and method [41].

The ability to use LOINC as a method to standardize the representation of laboratory results from different institutions depends not only on the ability of participating institutions to correctly encode their local tests into LOINC identifiers but also on the specificity and completeness of the LOINC vocabulary [41]. Despite the importance of the instrumental platform in laboratory result comparability and interpretation, this information is not part of the LOINC vocabulary construct and is seldom reported with laboratory results (Table 1) [40]. Therefore, LOINC is

insufficiently precise to adequately represent tests that are not analytically standardized or harmonized [42].

Information on the instrumental platform used to produce results is potentially invisible to AI and ML models, as current protocols do not routinely code for or disclose this data with test results. Analytical techniques, reagents, and equipment calibration vary among in-vitro diagnostic platforms and test kits, leading to significantly variable results even when using the same methodology. Researchers increasingly recognize that data on instrumental platforms and test kits should accompany results to ensure not only interoperability in patient care but also to support secondary uses in research, public health reporting, and AI and ML applications [22, 40].

The Systemic Harmonization and Interoperability Enhancement for Laboratory Data (SHIELD), a public–private partnership in the US, has suggested that test results can be accurately represented using standard medical terminologies (Table 2), with LOINC representing the test and SNOMED CT codes representing specimen information and qualitative results. Quantitative values would ideally be characterized by their units of measure using Unified Codes for Units of Measure (UCUM). The specific device can be described using the Device Identifier component of the Unique Device Identification (UDI) system. When combined with LOINC, the Device Identifier could provide the additional specificity needed to prevent the aggregation of non-harmonized laboratory result data in AI and ML algorithms [40, 43].

Organizational interoperability

Organizational interoperability is the highest level of interoperability and represents the ability to successfully apply AI and ML algorithms developed using data from other institutions and organizations to a particular organization or site. An AI or ML model that behaves differently depending on its deployment site raises concerns about patient safety and effectiveness. Barriers to the transportability and interoperability of AI and ML models include programming complexities that limit the integration of different systems and databases and the diversity of clinical data sources in terms of how laboratory and clinical data are captured and stored [44]. Culture, workflow, and local technology are variables that must be managed to achieve organizational interoperability.

A study involving 68 oncology sites revealed a representation agreement rate of 22%–68% for six medications and six laboratory tests for which well-accepted standards exist. The agreement was the highest (68%) among sites that used the same EHR vendor system and the lowest (20%) among sites with different EHR systems. In the best-case scenario, in which standardized representations exist, and the institutions utilize EHR systems from the same vendor, approximately two-thirds of data types will be "understood" by the other site [28].

Even within a single hospital, ML models trained on data from one EHR system deteriorated significantly when tested on data from a new EHR system after the hospital changed vendors [45]. This illustrates how prediction systems are vulnerable to abrupt changes in the EHR system when little overlap exists between representations. Changes in laboratory instrumentation producing different baseline values are expected to have the same effect [40].

It has been argued that all prediction models are subject to an expiration date. The field of medicine is fast-changing and dynamic. Patient populations, standards of care, available treatment options, measurements, and data representation evolve

Table 2. Data model for digital representation of laboratory test results as proposed by SHIELD

Test data element		Standardized ontology
Test information	Test ordered	LOINC to identify the test ordered
	Test performed	LOINC to identify the test performed
Specimen information	Specimen type	SNOMED CT at minimum
	Specimen source	SNOMED CT
	Specimen collection method	SNOMED CT
Results	Quantitative	Needs to include units of measure with UCUM preferred
	Qualitative	SNOMED CT
Methodology	Test kit identification	Device identifier
	Instrument identification	Device identifier

Abbreviations: SHIELD, Systemic Harmonization and Interoperability Enhancement for Laboratory Data; LOINC, Logical Observation Identifiers Names and Codes; SOMED CT, Systematized Nomenclature of Medicine – Clinical Terms; UCUM, Unified Codes for Units of Measure.

https://doi.org/10.3343/alm.2024.0323

over time, and these changes can influence AI and ML models [46]. The Coalition for Health AI (CHAI) is a community of health systems, public and private organizations, academia, patient advocacy groups, and expert practitioners of AI and data science engaged in harmonizing standards and reporting for health AI and educating end users on evaluating efficacy and safe integration into healthcare settings before adoption [47]. CHAI promotes local evaluations to determine whether AI and ML models perform as expected with local data and conditions and conducts long-term monitoring to ensure that models remain effective and are adaptable to any changes (Table 1).

POPULATION-SPECIFIC ISSUES RELATED TO LABORATORY DATA AND THEIR EFFECTS ON AI/ML MODELS

The use of clinical guidelines and human knowledge in AI models has raised the concern of perpetuating or even worsening existing health disparities. Algorithms learn from decisions or classifications that may contain potential biases against specific populations [48]. For example, studies have suggested that the use of the Black race modifier with estimated glomerular filtration rate equations can lead to physicians failing to diagnose early stages of chronic kidney disease in the Black population, which leads to a delay in secondary prevention [49]. In addition, despite best efforts to reduce the underrepresentation of minorities in randomized control trials in the US, non-Hispanic Whites of European ancestry still comprise >90% of the patient population in clinical trials [50].

Genetic studies have been criticized for not fully accounting for non-European populations. Patients of African or unspecified ancestry have been diagnosed as having pathogenic genetic variants that were actually benign but were misclassified because of a lack of understanding of variant diversity in these populations at the time of testing [51]. Simulations indicated that the inclusion of African Americans in control groups could have prevented misclassification.

Biases in EHR data, including laboratory data, used to train ML models may arise because of differences in patient populations, access to care, or the availability of EHR systems. One example is the widely used Medical Information Mart for Intensive Care III EHR dataset derived from patients receiving care at the intensive care units in Beth Israel Deaconess Medical Center, which has a largely White patient population [52]. In a US study, uninsured Black and Hispanic or Latino patients, as well as Hispanic or Latino Medicaid patients, were less likely to have primary care physicians with EHRs than White patients with private insurance [53]. Race and ethnicity undeniably have long been part of medical decision-making. Most physicians would be expected to know that sickle cell disease is substantially more common in the African and Mediterranean populations than in northern European populations and *vice versa* for cystic fibrosis and hemochromatosis [54]. This clustering of certain diseases in certain populations can have implications for laboratory tests as well as for AI and ML algorithms and models.

Sickle cell trait (SCT) is associated with elevated levels of thrombin-antithrombin complexes (TAT) and d-dimer [55], and African-American patients with SCT had lower levels of HbA1c at any level of fasting or 2-hr glucose than those without SCT [56]. The use of Hb A1c is not recommended at all for patients with sickle cell disease; however, Hb A1c is still misused in 11% of patients with sickle cell disease [57]. This quality gap may benefit from AI and ML algorithms, but only in cases where the clinical data are accurate and complete.

One of the issues confounding the use of race and ethnicity in medicine is the question of whether the data are captured accurately. Subjectively assigned race recorded by hospital administrators or providers is often incomplete and therefore inaccurate [58]; 59% of patients were misclassified by healthcare administrators and physicians as single race when they self-identified as multiracial. Predictive models that use race to assess population health risks and disease conditions are effective only when developers consider the potential errors in race data previously collected and the conclusions drawn in healthcare literature [58].

The underrepresentation of women and minority groups and inaccurate collection of race in datasets can contribute to inaccurate and unfounded conclusions that can unfairly disadvantage certain patient populations. The application of AI and ML should be fair and benefit all patients. A framework for modeling fairness using tools for causal inference in ML algorithms has been developed [59]. “Counterfactual fairness” is the concept that an algorithmic decision demonstrates fairness to an individual when the outcome remains constant in both the actual world and a counterfactual world where the individual belongs to a different demographic group [59]. This concept should help guide the development of AI and ML algorithms.

CONCLUSIONS

As AI and ML in healthcare mature, efforts to derive actionable insights from large datasets, such as laboratory data, will push

the traditional boundaries for the use of this information. This article discussed the current limitations of laboratory data in terms of standardization and interoperability that limit the potential of AI/ML and big data. Issues such as the lack of harmonization of test methodologies and variations in the digital representation of laboratory data must be addressed before AI/ML models can provide meaningful and valuable benefits to patients and providers. Data collection for race and ethnicity and the adequate representation of minority groups in clinical guideline data must be framed as important front-of-mind concerns for AI and ML algorithm developers to avoid perpetuating health disparities and inaccurate conclusions in the healthcare literature.

ACKNOWLEDGEMENTS

None.

AUTHOR CONTRIBUTIONS

Luu H was involved in conducting the literature review; manuscript writing, editing, proofreading; and reference formatting.

CONFLICTS OF INTEREST

None declared.

RESEARCH FUNDING

None declared.

REFERENCES

1. Harrison JH, Gilbertson JR, Hanna MG, Olson NH, Seheult JN, Sorace JM, et al. Introduction to artificial intelligence and machine learning for pathology. Arch Pathol Lab Med 2021;145:1228-54.
2. Miller MI, Shih LC, Kolachalama VB. Machine learning in clinical trials: a primer with applications to neurology. Neurotherapeutics 2023;20:1066-80.
3. McAlpine ED, Michelow P, Celik T. The utility of unsupervised machine learning in anatomic pathology. Am J Clin Pathol 2022;157:5-14.
4. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med 2019;25:16-8.
5. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. Acad Pathol 2019;6:2374289519873088.
6. Gupta S, Tran T, Luo W, Phung D, Kennedy RL, Broad A, et al. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. BMJ Open 2014;4:e004007.
7. Al Fryan LH and Alazzam MB. Survival analysis of oncological patients using machine learning method. Healthcare (Basel) 2022;11:80.
8. Ngo A, Gandhi P, Miller WG. Frequency that laboratory tests influence medical decisions. J Appl Lab Med 2017;1:410-4.
9. Myers GL and Miller WG. The International Consortium for Harmonization of Clinical Laboratory Results (ICHCLR) – A pathway for harmonization. EJIFCC 2016;27:30-6.
10. Miller WG, Tate JR, Barth JH, Jones GRD. Harmonization: the sample, the measurement, and the report. Ann Lab Med 2014;34:187-97.
11. Tate JR and Myers GL. Harmonization of clinical laboratory test results. EJIFCC 2016;27:5-14.
12. Park J, Lee S, Kim Y, Choi A, Lee H, Lim J, et al. Comparison of four automated carcinoembryonic antigen immunoassays: ADVIA Centaur XP, ARCHITECT I2000sr, Elecsys E170, and Unicel Dxi800. Ann Lab Med 2018;38:355-61.
13. van Schrojenstein Lantman M, van de Logt AE, Thelen M, Wetzels JF, van Berkel M. Serum albumin measurement in nephrology: room for improvement. Nephrol Dial Transplant 2022;37:1792-9.
14. Kidney disease: improving global outcomes (KDIGO) glomerulonephritis work group. KDIGO clinical practice guideline for glomerulonephritis. Kidney Int 2012;2:139-274.
15. Mactier R, Hoenich N, Breen C. Renal association clinical practice guideline on haemodialysis. Nephron Clin Pract 2011;118(S1):c241-86.
16. van de Logt AE, Rijpma SR, Vink CH, Prudon-Rosmulder E, Wetzels JF, van Berkel M. The bias between different albumin assays may affect clinical decision-making. Kidney Int 2019;95:1514-7.
17. Yang HS, Pan W, Wang Y, Zaydman MA, Spies NC, Zhao Z, et al. Generalizability of a machine learning model for improving utilization of parathyroid hormone-related peptide testing across multiple clinical centers. Clin Chem 2023;69:1260-9.
18. Jacobsen LM, Bocchino LE, Lum JW, Kollman C, Barnes-Lomen V, Sulik M, et al. Accuracy of three commercial home-use hemoglobin A1c tests. Diabetes Technol Ther 2022;24:789-96.
19. Beck RW, Bocchino LE, Lum JW, Kollman C, Barnes-Lomen V, Sulik M, et al. An evaluation of two capillary sample collection kits for laboratory measurement of HbA1c. Diabetes Technol Ther 2021;23:537-45.
20. Bietenbeck A. Combining medical measurements from diverse sources: experiences from clinical chemistry. Stud Health Technol Inform 2016;228:58-62.
21. Code of Federal Regulations. §493.1291 standard: test report, 2004. https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-G/part-493/subpart-K/subject-group-ECFR9482366886d579f/section-493.1291 (Updated on June, 2024).
22. Dahlweid FM, Kämpf M, Leichtle A. Interoperability of laboratory data in Switzerland – a spotlight on Bern. J Lab Med 2018;42:251-8.
23. International Consortium for Harmonization of Clinical Laboratory Results. https://www.harmonization.net/measurands/ (Updated on March, 2024).
24. Stram M, Seheult J, Sinard JH, Campbell WS, Carter AB, de Baca ME, et al. A survey of LOINC code selection practices among participants of the College of American Pathologists coagulation (CGL) and cardiac markers (CRT) proficiency testing programs. Arch Pathol Lab Med 2020;144:586-96.
25. Cholan RA, Pappas G, Rehwoldt G, Sills AK, Korte ED, Appleton IK, et al. Encoding laboratory testing data: case studies of the national implementation of HHS requirements and related standards in five laboratories. J Am Med Inform Assoc 2022;29:1372-80.
26. Hauser RG, Quine DB, Iscoe M, Arvisais-Anhalt S. Development and implementation of a standard format for clinical laboratory test results. Am

J Clin Pathol 2022;158:409-15.
27. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2:79.
28. Bernstam EV, Warner JL, Krauss JC, Ambinder E, Rubinstein WS, Komatsoulis G, et al. Quantitating and assessing interoperability between electronic health records. J Am Med Inform Assoc 2022;29:753-60.
29. Patel V, McNamara L, Dullabh P, Sawchuk ME, Swain M. Variation in interoperability across clinical laboratories nationwide. Int J Med Inform 2017;108:175-84.
30. Vest JR, Unruh MA, Shapiro JS, Casalino LP. The associations between query-based and directed health information exchange with potentially avoidable use of health care services. Health Serv Res 2019;54:981-93.
31. Holmgren AJ, Esdar M, Hüsers J, Coutinho-Almeida J. Health information exchange: understanding the policy landscape and future of data interoperability. Yearb Med Inform 2023;32:184-94.
32. Chen M and Esmaeilzadeh P. Adoption and use of various health information exchange methods for sending inside health information in US hospitals. Int J Med Inform 2023;177:105156.
33. Arvisais-Anhalt S, Lehmann CU, Park JY, Araj E, Holcomb M, Jamieson AR, et al. What the coronavirus disease 2019 (COVID-19) pandemic has reinforced: the need for accurate data. Clin Infect Dis 2021;72:920-3.
34. Hulsen T, Friedecký D, Renz H, Melis E, Vermeersch P. Fernandez-Calle P. From big data to better patient outcomes. Clin Chem Lab Med 2022;61:580-6.
35. Blatter TU, Witte H, Nakas CT, Leichtle AB. Big data in laboratory medicine-FAIR quality for AI? Diagnostics (Basel) 2022;12:1923.
36. Yim WW, Evans HL, Yetisgen M. Structuring free-text microbiology culture reports for secondary use. AMIA Jt Summits Transl Sci Proc 2015;2015:471-5.
37. Carter AB, de Baca ME, Luu HS, Campbell WS, Stram MN. Use of LOINC for interoperability between organisations poses a risk to safety. Lancet Digit Health 2020;2:e569.
38. Lin MC, Vreeman DJ, McDonald CJ, Huff SM. Correctness of voluntary LOINC mapping for laboratory tests in three large institutions. AMIA Annu Symp Proc 2010;2010:447-51.
39. McDonald CJ, Baik SH, Zheng Z, Amos L, Luan X, Marsolo K, et al. Mismappings between a producer's quantitative test codes and LOINC codes and an algorithm for correcting them. J Am Med Inform Assoc 2023;30:301-7.
40. Luu HS, Campbell WS, Cholan RA, Edgerton ME, Englund A, Keller A, et al. Analysis of laboratory data transmission between two healthcare institutions using a widely used point-to-point health information exchange platform: a case report. JAMIA Open 2024;7:ooae032.
41. Baorto DM, Cimino JJ, Parvin CA, Kahn MG. Combining laboratory data sets from multiple institutions using the logical observation identifier names and codes (LOINC). Int J Med Inform 1998;51:29-37.
42. Chang T, Herman DS, McClintock DS, Durant TJS. The roadmap to interoperability and laboratory data: current state and next steps. J Appl Lab Med 2023;8:226-8.
43. CLSI. Semantic interoperability for in vitro diagnostic systems. 1st ed. CLSI report AUTO17. Clinical and Laboratory Standards Institute, 2023.
44. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17.
45. Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. Proc Mach Learn Res 2019;106:1-23.
46. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med 2023;21:70.
47. Shah NH, Halamka JD, Saria S, Pencina M, Tazbaz T, Tripathi M, et al. A nationwide network of health AI assurance laboratories. JAMA 2024;331:245-9.
48. Kurant DE. Opportunities and challenges with artificial intelligence in genomics. Clin Lab Med 2023;43:87-97.
49. Marzinke MA, Greene DN, Bossuyt PM, Chambliss AB, Cirrincione LR, McCudden CR, et al. Limited evidence for use of a Black race modifier in eGFR calculations: a systematic review. Clin Chem 2022;68:521-33.
50. Ma MA, Gutiérrez DE, Frausto JM, Al-Delaimy WK. Minority representation in clinical trials in the United States: trends over the past 25 years. Mayo Clin Proc 2021;96:264-6.
51. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018;178:1544-7.
52. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci 2021;4:123-44.
53. Hing E and Burt CW. Are there patient disparities when electronic health records are adopted? J Health Care Poor Underserved 2009;20:473-88.
54. Jorde LB and Wooding SP. Genetic variation, classification and 'race'. Nat Genet 2004;36(S11):S28-33.
55. Amin C, Adam S, Mooberry MJ, Kutlar A, Kutlar F, Esserman D, et al. Coagulation activation in sickle cell trait: an exploratory study. Br J Haematol 2015;171:638-46.
56. Lacy ME, Wellenius GA, Sumner AE, Correa A, Carnethon MR, Liem RI, et al. Association of sickle cell trait with hemoglobin A1c in African Americans. JAMA 2017;317:507-15.
57. Sivasankar S, Cheng AL, Lubin IM, Lankachandra K, Hoffman MA. Use of large scale EHR data to evaluate A1c utilization among sickle cell disease patients. BMC Med Inform Decis Mak 2021;21:268.
58. Witzig RS and Dery M. Subjectively-assigned versus self-reported race and ethnicity in US healthcare. Soc Med 2014;8:32-6.
59. Kusner MJ, Loftus J, Russell C, Silva R. Counterfactual fairness. Adv Neural Inf Process Syst; 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017