PEDIATRIC AND MATERNAL-FETAL DIVISION NEWS
Volume 43 | Issue 1 |February 2024

FROM THE MIND OF THE CHAIR

Greetings everyone! Now that 2024 has started, I hope the beginning of the new year has greeted you with hope and health. Consistent with our mission to provide educational session for our membership, the PMF leadership group developed and submitted several proposals for the 2024 ADLM Annual Meeting. Along with a few other division-sponsored proposals, we hope to have several specific PMF-related sessions for you to attend. For those of you planning to submit a poster, please note the deadline for submission is February 15, 2024, at 5 pm US Eastern time.

Consistent with past newsletter issues, the letter "M" leads our topic of discussion. In this newsletter you will find an interesting and informative article on "M is for Machine Learning" authored by Nick Spies, MD of Washington University in St. Louis. Other content includes Excerpts from the Literature "Cotton ball urine collection: Is the juice worth the squeeze?" Authored by Stephen Roper, PhD. And last, but not least, we feature our Interview with a Distinguished Colleague, our 2023 awardee for Outstanding Contributions to Pediatric and Maternal-Fetal Clinical Chemistry, Alison Woodworth, PhD.

As this year continues, we have plans to provide educational opportunities (webinars) for our membership as well as preparing our division for upcoming changes to our organizational structure. Please feel free to contact me with any questions or comments regarding the PMF division. Enjoy 2024!

Sincerely,

Stanley F Lo, PhD
Chair, AACC PMF Division

Table of Contents

THE ABC'S OF PEDIATRIC LABORATORY MEDICINE:

M is for Machine Learning

Nick Spies, MD

Chief Resident
Clinical Pathology
Washington University
St. Louis, MO

What is Machine Learning?

Machine Learning (ML) refers to the discipline of building algorithms that can improve their performance on human tasks without explicit instruction (1). These algorithms and the tasks for which they are trained come in many forms and take many names, but their potential to impact the way laboratory medicine is performed – for better or worse – is becoming undeniable. We will take a brief foray into the world of AI and ML to discuss how these algorithms are built, validated, and implemented effectively and responsibly.

Fundamental Concepts and Key Terms

All machine learning applications should begin by defining the task at hand. While recent breakthroughs in generative AI have broadened the scope of machine learning tasks, most use cases within the laboratory still revolve around making discrete predictions.

Typical ML applications include regression (predicting a numerical outcome), and classification (predicting a label). If the dataset used to train the model included the labels being predicted, then it is referred to as supervised learning. In contrast, if no labels are provided, the task is unsupervised. The fidelity of the data and labels used to train the models is the single most crucial aspect of the entire endeavor. However, curating a pristine input dataset that is completely labeled by a veritable gold standard is difficult, if not impossible. This has fueled research in approaches that can use some labels (semi-supervised) or make their own labels (self-supervised).

Assessing Model Performance

As laboratorians, we are no strangers to critical appraisal of assays. While many of these skills transfer quite nicely, there are idiosyncrasies that bear mentioning. First is vocabulary. Where we would refer to sensitivity, positive predictive value, and measurement error, ML parlance uses recall, precision, and loss. Next, when establishing a threshold, it is crucial that we interpret performance in the context of the problem being addressed. While it is common to compare sensitivity and specificity via a receiver operating characteristic (ROC) curve, because many ML solutions aim to detect rare events, they often prefer precision-recall (PR) curves and metrics such as the F1 score or the Matthews Correlation Coefficient (MCC).

These metrics provide more relevant representations of real-world performance when classes are imbalanced, similar to the trade-off between reporting sensitivity and specificity versus positive and negative predictive value. This will be an important distinction when evaluating model performance across populations where disease prevalence diverges from that exhibited in the training dataset. For example, an algorithm for the detection of myelodysplastic syndrome (MDS) that demonstrates 99% sensitivity and specificity may provide stellar performance in an adult hospital setting, but next-to-no clinical utility in pediatric or obstetric settings, where the incidence of MDS is much lower. Conversely, algorithms for preanalytical errors such as hemolysis or IV fluid contamination may provide much more value in the pediatric hospital, where these errors are unfortunately much more prevalent, and the relative cost of repeat phlebotomy is greater.

Validating ML Models: Common Pitfalls

Equally important as choosing the right performance metrics, is applying them to the right data sets. During the development process, it is common to partition input data into a set for training the model, and a set for testing its performance. This partitioning comes in many flavors (cross-validation, bootstrapping, leave-one-out, etc.), and is commonly done as part of the internal validation of the model. However, achieving stellar results in this internal validation phase is no guarantee that the model will maintain that performance when applied to the real world.

A major pitfall in ML design is overfitting – when a model detects spurious patterns in the training data that are not correlated with real-world outcomes. One common cause, information leakage, occurs when the use of privileged information unavailable at the time of inference is used to aid in predictions (e.g. a model to predict diabetes that inadvertently has access to current hemoglobin A1c). While extreme cases like this are often trivial to spot, others are more insidious, such as calculating features using the full data set instead of the training set or having the same patient present in both the training and testing sets.

Robust model validation is essential for safe and effective implementation of ML algorithms, especially from the perspective of a laboratorian overseeing maternal-fetal or pediatric testing. Because data is such a valuable resource, and the relative abundance of laboratory data is much greater for non-pregnant adults, it will likely be commonplace for us to need to validate algorithms that were trained on these adults in our local populations. Familiarizing ourselves with some common causes of generalizability failures will help us feel more comfortable performing these validations, without having to wait for pediatric- or obstetric-specific models to be made available.

MLOps: Deploying Models to Production

Given all the hype, it would be fair to wonder why the vast majority of the models you have read about in newspapers and journals have not translated into tangible impacts in patient care. If we were to summarize the root causes of discrepancy in one word, it would be MLOps.

A portmanteau of machine learning and development operations, MLOps refers to all the engineering, information technology, governance, and implementation science that goes into taking a model off a researcher's laptop and deploying it into the live clinical systems (often called a production environment). While the full breadth of MLOps is beyond the scope of this article, it behooves us to explore a few key questions; how do we act on the model's predictions, how do we ensure their continued accuracy over time, and what can we do to explain them?

Considerations such as whether the machine learning predictions should have a "human-in-the-loop" or be integrated seamlessly into an automated workflow, what extent to which we must ensure that the model's performance is not deteriorating over time, and how we can validate and implement real-time explanations of model predictions on a case-by-case basis will require extensive exploration for each model. Additionally, current laboratory information systems lack the capacity and interoperability to allow for "plug-and-play" implementation. Integrating a validated model within the live clinical workflow to provide real-time predictions is currently a Herculean task feasible only for laboratories with a wealth of informatics expertise and IT resources (2).

The optimal way to approach these and other MLOps considerations remains an area of active research in the ML community as a whole, across both adult and pediatric populations.

Responsible AI and Algorithmic Fairness

Up until now, we have described performance assessment as it pertains to entire groups, such as the test set or validation set, without consideration of the relevant subpopulations within these groups. Failing to do so puts us at

risk of contributing to inequity, often with catastrophic consequences (3).

One does not have to search long to find claims of racist, sexist, and otherwise biased algorithms in the real world. Placing the blame on the algorithm may stem from an inherent desire for us to anthropomorphize artificial intelligence but is inherently a bit misguided. It may seem optimistic to state that algorithms are not actually as racist or sexist as some claim, but the reality is much more bleak. These algorithms are reflecting the patterns inherent in the data and labels used to train them. If a model is built to predict whether an employee is “CEO material”, and trained using historical examples of CEOs, we should not be surprised to find that it prefers older, white men. While ML solutions offer immense potential to improve the operational and clinical dilemmas facing the modern laboratory, they also present an equally powerful mechanism by which existing disparities can become even more entrenched and amplified (4).

These considerations ring especially true in pediatric and maternal-fetal medicine, where children and pregnant patients are commonly excluded from training datasets, resulting in algorithms that may not generalize to these populations. Shifting the framing of the conversation from biased algorithms to biased datasets reminds us of our responsibility to explore and address these issues before they can cause harm through an overly hasty implementation. Fortunately, the topic of algorithmic fairness in laboratory medicine has received a substantial degree of attention and thought recently (4), providing us with useful tools and frameworks by which to translate these principles into practice.

Conclusion

ML represents the potential to revolutionize the practice of laboratory medicine, but it is not without its drawbacks and dangers. As these solutions become commonplace in literature and laboratories alike, we must develop the skills to appraise, validate, implement, and monitor them in a safe, responsible, and equitable fashion.

References

1. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM J Res & Dev [Internet]. 1959 Jul [cited 2023 Dec 28];3(3):210–29. Available from: http://ieeexplore.ieee.org/document/5392560/
2. Wheeler SE, Block DR, Bunch DR, Gramz J, Leung EKY, McClintock DS, et al. Clinical Laboratory Informatics and Analytics: Challenges and Opportunities. Clinical Chemistry [Internet]. 2022 Nov 3 [cited 2024 Jan 6];68(11):1361–7. Available from: https://academic.oup.com/clinchem/article/68/11/1361/6764778
3. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447–53.
4. Azimi V, Zaydman MA. Optimizing Equity: Working towards Fair Machine Learning Algorithms in Laboratory Medicine. J Appl Lab Med. 2023 Jan 4;8(1):113–28.

Excerpts from the Literature

Stephen Roper, PhD
Associate Medical Director of Pediatric Laboratory Services, St. Louis Children’s Hospital

Associate Professor of Pathology & Immunology, Washington University School of Medicine, St. Louis, MO

Cotton ball urine collection: Is the juice worth the squeeze?

Collecting urine from infants is challenging. Speaking with experience as a father, I can confirm that young children rarely urinate when or where you want them to. Catheterization and suprapubic aspiration are options but may be unnecessarily invasive depending on the reason(s) for testing. Alternative urine collection techniques for infants include the use of adhesive bags placed around the genitals, however this method is imperfect; bags may leak,

adhesives can cause skin irritation, and [I am told] this approach does not always work well for females (1). Granted the broad utility of urine chemistry testing, there is a need to validate more practical collection methods for infants.

The idea of using an absorbent medium to collect urine is not new. For more than 30 years, studies have investigated this practice for various tests (chemistries, microbiology, microscopy, etc) with mixed results (2, 3). The basic idea is that an absorbent material is placed in the front of an infant's diaper. Once urination has occurred and the material is saturated, it is compressed to extract the specimen for laboratory testing. Obvious limitations to this approach are lack of sterility, the possibility of contamination with feces and/or fibers from the absorbent material, and the inability to collect timed specimens. However, many of these drawbacks also apply to adhesive bags and the relative simplicity of using an absorbent medium is attractive to healthcare staff and parents.

The June 2022 issue of the Journal of Pediatrics features a contemporary, multi-site study investigating the use of cotton balls in diapers to obtain urine for routine chemistry testing (4). Thomas et al. designed this study to assess several variables that could potentially influence urine chemistry results, including different brands of cotton balls and diapers, as well as various chemistry analyzers. The researchers began by generating 20 aliquots of a urine from a pool of residual specimens submitted for urinalysis. Each aliquot was divided across 6 treatment groups and 1 no treatment control (reference specimen). The treatment groups covered various combinations of cotton balls (Curity and Centurion) and diapers (Huggies, Pampers, LUVs) used at the authors' institutions, as well as a no diaper control. For the treatment groups, 5-10 cotton balls were placed in a diaper, saturated with approximately 20mL of urine, and covered with a ziplock bag. Following incubation at 37°C for 1 hour, urine was extracted by placing the cotton balls in a 20mL syringe and applying pressure with the plunger. Extracted urine was separated into aliquots and frozen. Specimens were shipped to 5 labs employing various chemistry analyzers (Ortho Vitros 4600, Beckman Coulter AU5822, Roche Cobas 6000, Abbott Architect c8000, and Siemens Dimension Vista 500) where sodium, potassium, creatinine, urea, calcium, magnesium, phosphorus, albumin, and total protein were measured. Data was analyzed, relative to the no treatment control, using regression analysis, Bland-Altman plots, and t-tests.

Review of correlation and percent bias data did not suggest any brand-specific trend for the cotton balls or diapers included in this study. Analysis of individual analytes, however, did reveal several important differences related to the collection technique. For all treatment groups and the no diaper control, albumin and total protein measurement were found to have a substantial inverse negative bias when urine is soaked in cotton balls (range: -18% to -53%). As well, an inverse positive bias was evident at the lower end of the concentration range for urine calcium on all chemistry analyzers. Sporadic mild to moderate differences in magnesium and potassium concentrations were also apparent on some instruments for select cotton ball and diaper combinations. Overall, the most consistent measurements between treatment and control groups were observed for sodium, urea, creatinine, and phosphorus.

The study by Thomas et al. addresses an important, yet underappreciated challenge with pediatric specimen collection. The technique evaluated has the potential to simplify random urine collections from infants and improve clinical staff and parental satisfaction with lab services. Unfortunately, findings indicate that cotton ball urine collections can influence total protein, albumin, calcium, magnesium and potassium results. Given this information, labs wishing to accept cotton ball specimens for chemistries should evaluate this practice locally and proceed with caution. Outstanding questions include what, if any, effect does this collection method have on other tests commonly ordered in infants (urine drug test, for example). As well, it would be worthwhile to investigate the effect of delayed urine extraction from cotton balls to account for variability in the timepoint at which specimens are retrieved from diapers.

References

1. Kaufman J, Temple-Smith M, Sanci L. Urine sample collection from young pre-continent children: common methods and the new Quick-Wee technique [eng]. Br J Gen Pract 2020 Jan;70 690:42-3. Epub 2019/12/28 as doi: 10.3399/bjgp20X707705.
2. Ahmad T, Vickers D, Campbell S, Coulthard MG, Pedler S. Urine collection from disposable nappies [eng]. Lancet 1991 Sep 14;338 8768:674-6. Epub 1991/09/14 as doi: 10.1016/0140-6736(91)91242-m.
3. Ritchie D, Broadbent R, Medlicott N, Reith DM. In vitro validation of a method for neonatal urine collection and analysis [eng]. BMJ Paediatr Open 2019;3 1:e000482. Epub 2019/07/20 as doi: 10.1136/bmjpo-2019-000482.
4. Thomas SN, Stieglitz HM, Hackenmueller S, Suh-Lailam B, Pyle-Eilola AL. Use of Cotton Balls in Diapers for Collection of Urine Samples Impacts the Analysis of Routine Chemistry Tests: An Evaluation of Cotton Balls, Diapers, and Chemistry Analyzers [eng]. J Pediatr 2022 Jun;245:179-83.e8. Epub 2022/03/07 as doi: 10.1016/j.jpeds.2022.02.051.

Interview with a Distinguished Colleague: 2023 Award for outstanding contributions to pediatric and maternal-fetal clinical chemistry

By Stanley F Lo, PhD

Alison Woodworth, PhD, DABCC

Clinical Director, Global Laboratory Services
CTI Clinical Trial and Consulting Services

When you reflect upon your professional career in clinical chemistry, do you have any words of wisdom for those beginning their own career?

I would encourage clinical chemists that are early in their careers to get involved! ADLM has many opportunities to volunteer as do many other laboratory medicine based organizations. I started at the local section level and had the pleasure of serving on many committees from the local to international levels. These opportunities to interact with professional colleagues and friends are invaluable. Committee work helps to develop leadership skills and a portfolio for future career advancement. Most importantly, developing connections with professional colleagues gives us a group of experts with whom we can share ideas, ask questions, and collaborate.

You have recently made a slight change in your career path. In making this change, what has been the most significant change to your practice as a clinical chemist?

I did recently change my career path. After spending 16 years as an academic clinical chemist, I have shifted gears and now direct a laboratory at a Clinical Research Organization. While the day-to-day work is quite similar, we perform laboratory testing to evaluate patients as a part of clinical trials. This is another area where Clinical Chemists can thrive and show their scientific and clinical expertise! The most significant change in my new job is that this is a global laboratory operation, so we are mindful of global harmonization and the intricacies of laboratory testing and associated regulatory oversight for large clinical trials around the world. It's an exciting field and quite rewarding as we can help to identify new treatments for rare diseases and other complex health challenges!

In the area of clinical chemistry, what sort of changes are you expecting in the future?

I think we will experience significant change in the field of clinical chemistry in the future. I believe that AI and large data analytics will cause a shift in what we consider a diagnostic test. Complex diseases will now have better and more sensitive diagnostics by harnessing AI based laboratory testing algorithms. In addition, COVID has completely shifted the landscape of laboratory testing. I think we can expect the continuation of the trend to have lab testing performed at the point or care or even at home. The use of smart phones will allow physicians and patients to monitor complex diseases more closely from home. Hospitals and clinics' connectivity to EHRs and big data will all for easier remote patient monitoring. Finally, I believe that we will see a shift from reactive to proactive diagnostics and the laboratory can lead the way! Through access to large networks of patients through connected EHRs and Insurance databases, complex disease monitoring can be driven through AI and big data analysis. Patient need never miss an appointment or necessary laboratory testing as algorithms can be created to reach out to patients and facilitate just in time laboratory testing.

2023-2024 PMF Division Executive Board

Chair: Stanley Lo, PhD

Chair Elect: Joe Wiencek, PhD

Past Chair: Angela Ferguson, PhD

Members At Large:
Kara Lynch, PhD
Stephen Roper, PhD
Amy Pyle-Eilola, PhD
Sydney Strickland, PhD

Secretary: Laura Smy, PhD

Treasurer: Erin Schuler, PhD

Newsletter Editor: Sarah Wheeler, PhD

Newsletter Editorial Board:
Emily Garnett, PhD
Stephen Roper, PhD

Fellow Representative: Erving Laryea, PhD