# On Developing Better Practices for Reproducible Data Analysis in Laboratory Medicine

**Article:**
Patrick C Mathias.
*On Developing Better Practices for Reproducible Data Analysis in Laboratory Medicine.*
J Appl Lab Med 2023;8(1): 229–31. https://doi.org/10.1093/jalm/jfac097

**Guest:** Dr. Patrick Mathias is an Assistant Professor and Associate Medical Director of the Informatics division in the Department of Laboratory Medicine and Pathology at the University of Washington School of Medicine.

Randye Kaye: Hello, and welcome to this edition of *JALM* Talk from *The Journal of Applied Laboratory Medicine*, a publication of the American Association for Clinical Chemistry. I'm your host, Randye Kaye.

Clinical laboratories produce increasingly immense amounts of data. When laboratorians have ways to effectively extract and analyze that data, they're well-positioned to assess and improve the quality and efficiency of their services. Within basic science research, there's been a growing awareness of the need for rigorously designed experiments to ensure data reproducibility, which refers to the ability to achieve identical findings using the original data, code, and documentation. A Laboratory Reflections article in the 2023 Special Issue of *JALM,* on Data Science and the Clinical Laboratory, emphasizes that reproducibility should be expected and laboratory medicine as well. Today, we're joined by the article's author, Dr. Patrick Mathias.

Dr. Mathias is an Assistant Professor, who serves as the Vice Chair of Clinical Operations and an Associate Medical Director of the Informatics Division in the Department of Laboratory Medicine and Pathology at the University of Washington School of Medicine. Dr. Mathias' interests include improving clinical information systems to improve the ordering and interpretation of laboratory tests and applying analytics to improve laboratory quality and extend the lab's impact on clinical care. Welcome, Dr. Mathias, let's start with this. Why is data analytics valuable to the practice of laboratory medicine?

Patrick Mathias: Yes. So data analytics, data analysis, or analyzing data, really provides a foundation for our core operations in lab medicine. Every test that we offer is validated. That requires some collection of data to determine test characteristics, like accuracy, precision, sensitivity, specificity, and you know, that underlies the performance of our testing. In addition, many of our quantitative tests require some data analysis to go from a raw signal to results. We perform a lot of routine activities for quality assurance, like quality control, and that is inherently an activity of analyzing data to understand whether a test is performing as expected. And so, these are

really basic foundational concepts in laboratory medicine. And I think what we're seeing more recently is an expansion of these skills, these techniques, to other areas where we can improve the care that we're giving our patients.

So, we can look at using data analytics to improve, to assess the quality of our laboratory services, look for improvement opportunities, and integrate with other clinical data that might live outside of the laboratory, together to really provide useful information to our colleagues for patient care. And so, I'd say that analytics really is both foundational to laboratory medicine, but it also exists on kind of a frontier of doing more with our data and helping to improve how we take care of patients through laboratory testing.

Randye Kaye: Well, thank you. Let's go a little bit more specific here. What applications of analytics are the most promising for the field?

Patrick Mathias: I think there's a lot of opportunities for expanding on how we use analytics. There are a lot of opportunities to improve how laboratory tests are ordered. For example, I do a lot of work in the laboratory stewardship space. So, that really focuses us on thinking about how we use our laboratory testing in the best way possible, making the best use of our resources to provide high quality patient care. This can include some things like decreasing some activities, like decreasing the use of expensive or laborious testing where it doesn't improve patients' care. It can also include increasing the use of testing that positively impacts patient outcomes. For example, identifying patients who would most benefit from a screening or diagnostic test. That area of stewardship is one area and as an informaticist, I am frequently working with information systems. I might be a little bit biased, but I think we have a lot of opportunities to gain insight from our EHR data on how tests are being ordered and how to modify the EHR, the Electronic Health Record, to encourage the best use of lab testing.

So in my institution we are, at University of Washington we're increasingly making routine changes to order sets, preference lists, order panels, these Electronic Health Record constructs that help promote laboratory use or decrease unnecessary use. We are very frequently using analytics to identify what are our opportunities to improve in those areas. And then tailoring things like clinical decision support to suggest what tests are best for the patient and you know, what tests are either under ordered or over ordered.

There are other opportunities as well. I think there's a lot to be said for using your data and understanding how your lab is operating on a daily basis and every laboratory has an opportunity to improve how they're operating by really gathering data on the lifecycle of samples, looking for

opportunities to streamline workflows, or better align staffing to workload. And that's particularly important in our current environment where many labs don't have any staff as we need to get the work done. And so, using analytics to address that operational need can be a really important tool. And finally, you know, looking for the horizon, I think we're getting a lot more experience with machine learning, artificial intelligence. These are another set of tools. These are complementary to some of the more common descriptive analytics that I think many laboratory directors, managers use, and I think we'll continue to find more targeted applications where these techniques are more effective than some of our a simpler constructs, plots, or models.

And so, beyond solving more simple operational problems, I think there's really a role for artificial intelligence to help develop broader predictive models that incorporate lab data to predict outcomes and help support our clinical colleagues in taking the best care of their patients. I think it's important though, to recognize that just like laboratory testing, there's a really important need to evaluate those models rigorously and carefully before we use them in practice. So, yeah, I see a lot of applications really across the field, and I think there's been a lot of activity over the history of lab medicine in applying analytics, but we're starting to get some of these newer tools that might have an important role down the road for us.

Randye Kaye: That's a lot of information and a lot of applications. Let's talk a bit about reproducibility though, because you emphasize that concept in the article. Can you expand on what you mean by reproducibility and why it's so important?

Patrick Mathias: Absolutely. I think sometimes when we look at some of the fancier cutting-edge work, we might lose sight of really some of the basic principles that are important for us to keep in mind, as laboratorians and as professionals who are analyzing data.

So, I think reproducibility is one of those things that you can easily take for granted, but it's something that really should be at the forefront of the work that you're doing. And the definition in this context of data analytics really means that you can generate the same results based off of the same data set over multiple times. And so, you can take a data set and, for example, hand it to multiple people, and each of those people that you hand it to will give you the same results. Again, it seems very simplistic. It seems like a very simple concept, but I think there needs to be more attention paid to that basic concept across not just laboratory medicine, but really across all of medicine, and how we analyze the data that's in front of us.

This concept of reproducibility is in contrast to replication, and the scientific context replication means that you can get the same result or finding based off of an independently generated data set. So, the bar for reproducibility is actually easier to achieve than replication but it takes some effort. So despite data analysis, data analytics, being foundational to lab medicine, I think reproducibility is really important for not only others to understand the work that you've done, but even for yourself to understand, clearly understand ,what analysis was performed. Yeah, I think very often use this example of you've done some data analysis for a method validation study, and you want to come back to that six months later and revisit what you did and make sure that you understand exactly what was done to bring that test online. Reproducibility kind of speaks to that concept of being able to come back, six months, a year, or multiple years later and exactly understand what was done.

Randye Kaye: I see. Now, you may have already answered this at least in part, but are there any more specific principles of reproducibility that are most important for laboratorians?

Patrick Mathias: Yeah. I think there are these broader principles of reproducibility that are increasingly popular in the scientific literature and across the practice of science. I think there are a few things that are a little bit more specific for laboratorians that we can keep in mind, and I think a lot of the data that we generate actually comes from the laboratory information system or these other clinical information systems. And so, if you are able to develop standardized ways to extract that data and then do the work of validating, or ensuring that the data that you've pulled really reflects what you expect it to reflect in the database or in the record, doing that in the more standardized way, particularly as you do more analytics work, will really help improve your efficiency and help make each subsequent analysis that you do more efficient and effective because that data is predictable. You're extracting in a very predictable way. So, I think that is one skill to keep in mind that's laboratory-specific.

The other thing that I think is not necessary laboratory-specific, but I think is another important principle that really goes hand in hand with extracting your data in a standardized way, is developing what it's called a "data dictionary." The idea of a data dictionary is almost exactly as it sounds. For each data element that you have extracted from a system, for example, your laboratory information system, you have a definition in some underlying data around what that data element means. So, while it seems very simplistic and might seem like it is overkill, I think it is underrated how much that can help you in not only understanding the data that you're pulling from the lab information system, but also when

communicating to others who may want to use that data in the future.

Randye Kaye: Finally, Dr. Mathias, what steps can laboratorians take to improve the reproducibility of their data analyses?

Patrick Mathias: Yeah. I think it is one very simple step. And again, this article is really focused on some simple steps that every laboratorian can take. I think one simple step that can seem deceptively simple but it is really important and is often overlooked, is separating your raw data from your data analysis. I think, in the laboratory, unless you have learned some other tools, it's very common at all levels of the laboratory, whether you are collecting data on the bench, you are overseeing a validation, you're reviewing that data as a director. All levels, we very frequently interact with spreadsheets, tools like Microsoft Excel that really encourage the manipulation of raw data. And very often the analyses are done by working within the spreadsheets to analyze your data. If that's your primary tool for analysis, I think it's very important to develop a habit of generating a separate file from analysis that really stays separated from your raw data. I think that's really a stepping stone toward greater reproducibility. Even better, if you supplement that step of really separating your analysis from your raw data with including some detailed instructions on what is actually performed in the analysis, that's an even better practice, perhaps aspirational, but it can be really critical to ensuring that someone else can come in and do the analysis that you had done.

There are more advanced tools such as programming languages. Very frequently in the data analytics, data science world, people are using programming languages like R or Python and there are analysis formats called notebooks that help reinforce this reproducibility. So they inherently separate out your raw data from your analysis and they encourage the integration of both writing some code in that programming language, interspersed with some text that can help explain that data analysis. So this is a, I think really helpful tool, particularly if you are doing a larger volume of data analysis as part of your responsibilities, and I would encourage the audience to really explore these programming languages, these tools, such as the R programming language. There are a lot of online resources such as free courses, that are available to learn these tools, some of the conferences in the laboratory medicine space also provide workshops, and I do want to put in a plug for the work that AACC is doing in the data analytics space. And so, I'm part of the Data Analytics Steering Committee and we've been working with multiple other committees in the AACC to develop a data analytics curriculum that can help provide guidance really for all the membership of AACC, as well as others who are

interested in -- other laboratorians who are interested in learning how to do more effective data analytics.

And so, as we continue to develop that, we hope it can provide a nice resource for anyone is interested and there is more information on the AACC website. There's a data analytics page there, which will continue to develop as we develop more resources.

Randye Kaye: Wonderful, thank you so much for those resources. I am familiar with some of them as my husband teaches data analytics as an adjunct professor. I'm actually heard of it. I'm going to have him read these entire issue so he can have some stories to tell his students. Thank you so much for joining us today, Dr. Mathias.

Patrick Mathias: Thank you very much. Yeah. Just let me know if there's anything else I can help with, and thanks for the opportunity.

Randye Kaye: That was Dr. Patrick Mathias from the University of Washington describing the *JALM* article "On Developing Better Practices for Reproducible Data Analysis in Laboratory Medicine." This article is part of the January 2023 Special Issue of *JALM*, on Data Science and the Clinical Laboratory. Thanks for tuning in to this episode of *JALM* Talk. See you next time, and don't forget to submit something for us to talk about.