Transcript Details
This is a transcript of an educational program. Details about the program and additional media formats for the program are accessible by visiting: https://reachmd.com/programs/lipid-luminations/big-data-lipids-refining-cardiovascular-risk-profiles-nationwide/8070/

ReachMD
www.reachmd.com
info@reachmd.com
(866) 423-7849

Big Data in Lipids: Refining Cardiovascular Risk Profiles Nationwide

Narrator:

Welcome to ReachMD. You are listening to **Lipid Luminations,** produced in partnership with the National Lipid Association and supported by an educational grant from AstraZeneca. Your host is Dr. Alan Brown, Director of the Division of Cardiology at Advocate Lutheran General Hospital and Director of Midwest Heart Disease Prevention Center at Midwest Heart Specialists at Advocate Healthcare.

Dr. Brown:

You're listening to ReachMD and this is Lipid Luminations, sponsored by the National Lipid Association. I'm your host, Dr. Alan Brown, and with me today is Dr. Seth Martin, Assistant Professor of Medicine and Associate Director of the Lipid Clinic Ciccarone Prevention Center at Johns Hopkins University School of Medicine. Today our topic is going to be using big data and how to use a very large database, particularly in the area of dyslipidemia. Seth, thank you very much for taking time out of a busy meeting to join us here at the NLA Meeting.

Dr. Martin:

My pleasure, Alan. Thank you.

Dr. Brown:

So, big data has become a cliché. People talk about it all the time and like psychiatry, half of what they know is correct and nobody knows which half. So, maybe you can start by telling us, for the audience, what do we mean when people are using the term big data?

Dr. Martin:

Yes. It's a great question and I don't have the great answer yet. I think we're still learning what big data is. Some people have said it's like teenage sex, because everyone talks about it, everyone talks about it because other people are doing it, but no one really knows what it is in the first place. But big data, it can be big in terms of the, you can just have a lot of patients in your dataset. So, in databases that we've worked with, we have over a million people, so you can just have a lot of people. You can have a lot of data points. You might measure thousands and thousands of things, so you just have a very wide data file with all these different things in each person or you may be getting big data because you're following people very closely with a smartphone or something, getting data points every minute of the day and your data's becoming big in that way. So, I think traditionally we're thinking about these big data files in just that we have big large numbers of people because we're tapping electronic health records, or other big datasets, but it can be big in different ways.

Dr. Brown:

So, let me ask you this, is there a danger when you just collect thousands of data points on millions of people, and then you go back and you try to draw conclusions. I mean, is there retrospective reanalysis danger that maybe you missed a couple of the most important data points so you draw weird conclusions? How do you decide what's an appropriate query of such a nebulous, large database?

Dr. Martin:

I think your question gets at one of the big concerns about using big data, the quality of the data. Are you really asking questions that the data truly can answer for you? I think for our research we've asked very simple fundamental questions and using the big, even

though it's big with lots of patients, we had a very limited array of variables that we actually knew about these people, and it forced us to focus our attention and ask questions that we thought the data really could tell us, but I think we do run the risk of getting over-excited sometimes in this big data space and we have all of this data and coming up with, you know, starting to look at associations between this and that and this and that. And when you have a lot of people, everything is going to be statistically significant. It's something we've noticed. So then, interpreting what is the actual, is it a real relationship in the first place? What were the quality that went into that? Some people have said garbage in, garbage out. So, we want to be sure that the conclusions we're drawing really are based on data that can actually answer those questions and lead to those conclusions.

Dr. Brown:

Yes, I think my fellows, they love to have a big data set to go in and just pull out a bunch of crap and publish something and, obviously, the ideal situation seems to be, if you know what questions you're going to ask, if you could determine what data points would appropriately answer that question, and then build a gigantic database prospectively, right? That's just not possible in any case.

Dr. Martin:

Yes. You're right.

Dr. Brown:

So, when you go back and look at a large healthcare database and you have a hypothesis, do you use it sometimes just to generate is this a reasonable hypothesis and then lead to prospective trials?

Dr. Martin:

It's a good question. I think when we started getting into this big data thing with lipids we were very much thinking in the traditional scientific way that we're going to have a hypothesis or we have a very specific objective to examine the accuracy of one variable, or to look at, we're hypothesizing that this variable is going to be related in a positive or an inverse way with this variable, and we're to test that. What's interesting is I've thought about it more and interacted with more folks is the other approach, which is the not the traditional approach which is the totally agnostic analysis of the data, meaning to go in without any pre-set belief as to what it's going to show, and then let the data speak to you, and this gets more into this machine-learning agnostic analysis of the data and let it identify potentially associations or signals in the data that wouldn't have been identified via the traditional scientific route. How much of that will be meaningful information, will turn out to be true, is yet to be determined, I think, in large part, but it's an interesting way to think about this, about letting the computer just analyze it without any pre-set notions.

Dr. Brown:

Yes, probably an interesting sort of example would be, we've been using statins for all these years and it never dawned on anybody that there could even be a slightly increased risk of developing diabetes, but if you had such a large dataset and you might see that signal, looking at millions of people, whereas looking at a few thousand that signal wouldn't be obvious. Is that a reasonable analogy?

Dr. Martin:

Yes. Maybe we would, if we were doing that in realtime while these studies were being conducted, that we would have noticed that sooner than we otherwise would.

Dr. Brown:

If you're just tuning in, you're listening to ReachMD. I'm Dr. Alan Brown and I'm happy to have Dr. Seth Martin, Assistant Professor of Medicine and Associate Director of the Lipid Clinic at the Prevention Center at Johns Hopkins University. So Seth, we talked a little bit, about general terms, about big data and mining that data and trying to come up with some interesting observations. Let's talk about the study that you did with this large database of lipids looking at the accuracy of Friedewald-derived LDL cholesterol, so calculated cholesterol, tell me a little bit more about what you found?

Dr. Martin:

Yes, absolutely, yes. So, the very large database of lipids, what we did was basically download millions of people's lipid data from the Atherotech VAP test server, so basically doing ultra-centrifugation of lipids to differentiate out their LDLs and HDLs, and all components of the lipid profile. When we got the data, we didn't know who anyone was, we don't have their social security numbers, we just know their age, sex, and their lipid profile, and then we do have subsets of information that have other lab tests that were done, like a CRP level, for example. And what we realized is we have a very limited dataset. In fact, talking with a lot of colleagues, some people, especially our expert epidemiologists, would say, "This is unusable. I wouldn't even bother with this data. It's not collected in the

traditional prospective cohort study kind of way where we know all this level of detail about people and conduct it in this very standardized way." But we also saw a lot of potential value here in what we had in this huge database to be able to get very confident results, because we have so many people, and to look in more subgroups than are traditionally possible with the cohort studies. And so, we did persist and we realized we had to ask very fundamental, simple questions. And so, the first question that we ended up asking was whether the Friedewald estimate of LDL cholesterol is accurate, as compared to ultra-centrifugation directly measuring LDL? So something that we could, really we just needed the lipid profiles to test. Yes, we would want to know other things like what drugs they're on or off, but really in clinical practice we use this in someone whether or not they're on drugs. So, we don't absolutely need to know that to answer the question fundamentally.

Dr. Brown:

So what were your results? What did you come up with because just anecdotally in our practice we used to do direct LDLs when we could get them at the same price as a standard and we found there was a significant difference. In fact, we went back and most of the guidelines were based on calculated LDL. So probably there were some other particles that were being considered LDL, some remnant particles etcetera in patients with dyslipidemia, even when their triglycerides were less than 400. So tell us a little bit about what you're finding.

Dr. Martin:

Absolutely. Absolutely. And as you say, the LDL, by the traditional definition, includes what's biologic in our bodies. LDL plus IDL plus LPa and what's taken out of the equation is subtracting the estimated VLDL cholesterol, so triglycerides divided by 5, and when Friedewald and his colleagues, Fredrickson and Levy, did their original study back in 1972, they wrote in their paper that that estimate of VLDL cholesterol is not particularly accurate, but it was tolerable because it was a small component of the equation. People's LDLs were much higher in that era than they are now. So, what we found is that at low LDL levels, when that estimate becomes a bigger part of the equation, and when the triglyceride level is higher, even if it's not above 400, even if it's that 150 to 200 or 200-300 range, there can be quite a lot of inaccuracy in the Friedewald estimated LDL. It still works very well for most people, but it works until it doesn't and unfortunately when it doesn't work it's often in the settings just when you'd want it to work the most, in those higher-risk people that you're treating to low LDLs and may have higher triglycerides, have diabetes, have atherosclerotic cardiovascular disease. So it really breaks down just when you would want it to work the best, unfortunately.

Dr. Brown:

Which is why we kind of focus on non-HDL or measurement of particles.

Dr. Martin:

Yes. Yes.

Dr. Brown:

The problem I had with the direct LDL number, even though we can agree that it's probably not accurate, still a good barometer in those people that don't have elevated triglycerides and we don't really know what to do with the direct LDL number. So, do you agree that probably a more appropriate measure would be of all atherogenic lipoproteins? So looking at non-HDL, or ApoB, or LDL particle numbers?

Dr. Martin:

Yes. Yes. So, in general, I would agree non-HDL's a really good measure. It's nice that it's so simple, total minus HDL cholesterol. Both of those are directly measured. It's really reproducible. We think it can be done in fasting and non-fasting states. What we've also looked at it in our big data, in the very large database of lipids and found that there's a lot of discordance between whether someone, even if they have a lower LDL, often their non-HDLs are still raised, we've traditionally, the cutpoints used for non-HDL, if it's equivalent to the LDL of 70, we'd use a cutpoint of a hundred. We and others, like **Christy Valentine*11:31**, have shown that probably the equivalent non-HDL for that LDL of 70 is closer to 90, if you look at how many people on a population-percentile basis would be above or below those cutpoints. So, that was another interesting use of our big dataset to look at non-HDL.

Dr. Brown:

That's fascinating stuff and hopefully those kinds of bits of information will help us as we get the new guidelines and new recommendations, and I'm encouraged that we're going to see some numbers coming back into those recommendations. So, we only have like 3 minutes left, unfortunately, tell me what your aspirations are since you're into the big data analysis. Any other ideas that you have that you're excited about that you want to tell our audience about?

Dr. Martin:

What we've talked about was the first iteration of our database and now we're moving into the second iteration of the database. First one had 1.3 million. This one's going to have over 5 million people. So, we're getting bigger in our big data and we'd like to continue to ask fundamental simple questions like we've done in the past and also try some of the machine-learning techniques, phenomapping, looking for patterns in the data that we may not expect or predict in the firsthand and really try to understand the heterogeneity in folks' lipid profiles perhaps better than we have in the past. Also, in addition to that cross-sectional lipid data that we've had in this database, are now, have linked the database to mortality. Of course, as cardiologists and clinicians, we want to know more about, more than just someone's all-cause mortality, we want to know what's their risk of heart attack, strokes? We won't have that kind of data, but still, all-cause mortality is a useful parameter, very important to insurance companies, or so, when they're trying to decide about insuring someone. So we will be able to look at these, at the VAP lipid profile in different age and sex groups to understand the relations, in the modern era, to all-cause mortality.

Dr. Brown:

And do you predict that if you get a pattern that was wildly unexpected that that would lead to a prospective randomized trial? Or what are you going to do with that information when something pops out at you that you said, "Boy, I never dreamt of this?"

Dr. Martin:

Yes. That's a real, that's a great question. Sort of, what is the next step or the end game to all of this? I think it's less, less clear for those kinds of studies, frankly. We would hope that it could identify some novel pathway that might then be targeted in the future. I think the stuff that we did with LDL does have sort of direct relevance to kind of the decisions happening in clinic on a day-to-day basis, because the guidelines tell us to use those LDL levels in, whether we classify someone as being in a statin-eligible or statin-benefit group, or when we're looking at on-treatment levels. So, those LDL, the LDL work, I think, has direct relevance and we did, one thing I forgot to mention is, after that first study we then did a followup study published in *JAMA* that shows that we were able to develop a better estimate of the LDL cholesterol, using rather than that one-size-fits-all factor of 5 dividing the triglycerides, we had an individualize factor, and there are 180 different factors that may apply, given someone's lipid profile, and so we get a better estimate of LDL. So, it may be that we don't have to go to those direct LDL assays; we can use our novel estimate and still remember that it's important also to look at other things like non-HDL cholesterol.

Dr. Brown:

Well this is fascinating, Seth. I could talk to you for a long time but unfortunately we've run out of time. So, I can't thank you enough for taking the time away from a busy meeting to speak with us on ReachMD about big data and clarify for the audience, number one, what the topic really means, and then some good examples of how we can use that to help guide our therapy.

I'm Dr. Alan Brown. You've been listening to Lipid Luminations, sponsored by the National Lipid Association on ReachMD. Please visit ReachMD.com/lipids where you can listen to this and other podcasts and please make sure to leave your comments and share the podcast with your colleagues. We welcome your feedback and, once again, I'm your host, Dr. Alan Brown. Seth, thank you very much for joining us.

Dr. Martin:

Thank you for having me.

Dr. Brown:

And thank you all for listening.

Narrator:

You have been listening to Lipid Luminations, produced in partnership with the National Lipid Association and supported by an educational grant from AstraZeneca. To download this program and others in the series, please visit ReachMD.com/lipids. That's ReachMD.com/lipids.