National Antimicrobial Resistance Monitoring System (NARMS) Technical Workshop Transcript

Introduction and Background- Presenter Dr. Errol Strain
Time- 0:00 – 27:50

Welcome everyone this morning! I'm Errol Strain, a bioinformatician with the FDA NARMS program. It's my pleasure to talk about accessing and using data generated from NARMS monitoring of enteric pathogens. You'll hear experts today at the CDC, USDA, NIH, and FDA about how to interact with various data types that are publicly accessible through NARMS. Our goal as bioinformaticians, data scientists and microbiologists working with this data is to put the right information in front of the people who can act on it. So, I would say, throughout the day, we appreciate your questions and feedback about the tools, visualizations, and all of the data structures that we put together. So this is, to the best of my knowledge, the first time NARMS has put together a technical workshop to compliment the public meeting. We are having a public meeting tomorrow and following Thursday. So why are we having a technical workshop? I would think the public meeting really, you know, you're going to hear a lot about the high level strategic plans and things going on in NARMS, high level findings and stories about the data. Part of that is those stories may not be your story. You may have different questions you want to look at, you might want to look at the latest data coming in from NARMS and trends in specific pathogens and specific antimicrobials that are not shown at the public meeting. Perhaps you have your own local analysis you want to incorporate some sequence data that you've put together and integrate that into NARMS and start to look at attributions. You may have some findings and see some interesting AMR results and see where they fit in to the larger picture of things like genomics surveillance across the U.S. and around the world. For this, the reason we have the technical workshop is there are multiple sources for NARMS data depending on the question you're asking. There's things like annual reports, we have various ways of sharing genome data and phenotypic data, there are multiple dashboards available from the FDA and CDC that we will show you today.

I would like to caveat part of the training and what we're trying to do. Unfortunately, this one day training will not make you a bioinformatics expert. It will give you a taste on how to interact with various types of NARMS data and if this does pique your interest please reach out because we might be able to help with other more in-depth trainings or have a more in-depth review on how to look at the real-time data going through the pathogen detection portal or more about the dashboards and visualizations we're doing in Tableau.

So, for today's workshop, we have broken it up in three different parts on it. In the morning, we're going to have experts from the different agencies to talk about their use of whole genome sequence data and the positive isolates that are found through the various monitoring and surveillance systems. These isolates are submitted to a public data base at NCBI that processes and report phylogenetic trees. In parallel there could be an analysis done at CDC and FDA where we grab that data and incorporate it into resources like NARMS Now.

One question you may want to ask when you're looking at whole genome data, like NARMS surveillance, is what have we found in the past two months on it?

Similarly, if there's historical NARMS isolates, you may want to monitor clusters to see if there are any new related human clinical cases, or, if not, if anything changed in the cluster related to historical isolates from all sources?

Later on starting this afternoon, I think that's where you will start to see where to look for trends and tracking larger changes and resistance over time. This differs from this first section where the focus is on isolates themselves. For this second session we focus on larger changes, for example, in the bottom right, showing *Salmonella* and ceftriaxone resistance, through different dashboards at the FDA and CDC.

Towards the end of the day, we'll talk about integrating NARMS data with other data sources. Mike Feldgarden from NCBI will demonstrate a little bit on how to use AMRFinderPlus and the NCBI Pathogen Detection Portal. The nice part of this for us is starting to make AMR findings associated with molecular mechanisms of resistance public. For a number of the pathogens and antimicrobials of interest, including the foodborne pathogens surveilled in NARMS, we don't always have to wait for the phenotypic data to come through. We can look at the AMR predictions. This can help you set up alerts if you have specific matrices, such as ground turkey or chicken or beef or pork, that you're interested in tracking. You can set up alerts it will send an email when a particular finding comes in.

And so, before we get started with the fun stuff and the experts talking today, I'm going to give you a little overview of the sample types coming in to NARMS to tee up some of other presentations. I'll also give a very quick intro to whole genome sequencing for those who are not familiar with it. I just joined CVM three years ago so I will say I'm still learning about it. I think you will hear today and tomorrow about how NARMS started in 1996 as an interagency collaboration between FDA, CDC and NIH. For those who are new to NARMS, one of the best resources to understand the why the program is guidance for industry 152. So, whatever your favorite search engine is, you can search up this guidance for industry and it really walks you through qualitative antimicrobial resistance for new animal drug applications with the focus on foodborne pathogens. For me coming in, why specific drugs were of interest or why a particular pathogen was monitored under NARMS, this helped me set the framework for understanding and interpreting a lot of the data collected as part of the NARMS monitoring program.

I would also like to mention, there's a newer concept paper relevant to this discussion as over the course of the next two days you'll hear about NARMS potential expansion in the one health arena. A lot of the focus for NARMS had been on foodborne pathogens and in the 2020 concept paper they explore extending it to one health framework, potentially to non-foodborne pathogens.

And about GFI #152, this guidance is one of the key areas of me learning about NARMS and Appendix A is a great place to start. They walk through the logic behind why they have classified different antibiotics as critically important, highly important, or important as it relates to the impact or importance for human health. You can see for here, for example, is it a foodborne pathogen? Is it responsible for human disease? Is it the sole treatment available for that disease on it and how does it rank across the different antimicrobial categories. So here, I'm showing a small portion of the table and guidance. GFI #152 is only about 25 pages and large font so I would encourage everyone to take a look. It really helped me set the framework for understanding NARMS, what the data types are, and to some extent, how I think how we should be sharing this data.

I would also like to mention, not only are the antimicrobials tested by NARMS provided, but also the breakpoints and cutoffs are also on the FDA and CDC websites. I think one thing I will emphasize a little bit in my talk this morning around it is NARMS has been collecting data since 1996. Things are generally stable but breakpoints and sample types may change over time. At NARMS we're trying to do as much as we can to make the raw data accessible and in this case, the raw data from the phenotypic testing means the MICs. If the break point did change over time you can go back and take a look at the new data and the historical data in that context. Okay, and just quickly hitting on the methodology used to collect samples from the different agencies. So this is the FDA retail meat surveillance and this is just an example of a typical month of collection. Each month representatives from NARMS laboratories go out and buy 8 samples of 8 packages of chicken breast, 8 of ground Turkey, 8 of ground beef, along with seafood and chicken gizzard at grocery stores. They will bring it back to the lab and work up the different pathogens based upon what is being monitored for NARMS. For *Salmonella*, all 8 samples of chicken and ground beef will be tested. For this one, this would be tested for *E. coli*. Just to highlight here once again, we're talking about real-time data sharing so positive isolates for these three pathogens that are collected in the NARMS labs are sequenced on site.

I will touch on sequencing in a minute but once again, for that data, as they are collected from NARMS, as quickly as we can we upload it in NCBI and that can help us do things like rapid detection of emerging pathogens and also make

NARMS data useful even beyond AMR attribution to source.

For FSIS, I mentioned NARMS is a USDA collaboration, originally with the agricultural resource service and now primarily with FSIS. FSIS leads the NARMS component for food animals, and this includes isolates collected through the pathogen reduction or HACCP testing, and starting around the 2013 or 2014, also, fecal samples from food animals. FSIS has done an amazing job of making their sampling protocols available online. There's more than I can quickly share here but you can see the different sample types collected over time through the program. Another point to emphasize, and you'll see this a couple of different times today is that around ten years ago is when a transition happened. If you remember your food pathogen surveillance history, around 2013, 2014 is the time that we started to use whole genome sequencing for surveillance of foodborne outbreaks.

Starting with *Listeria* in 2013 and evolving to *Salmonella*, you can see this is the time across the agencies when you get the start to get larger amounts of whole genome sequence data. So here, you're seeing a little bit around FSIS transitioned to use whole sequence. We will hear more about it from our presenters later on.

For CDC and their data types, you will hear from Jason in a minute, it's in close collaboration with the PulseNet for *Salmonella*, *Campylobacter*, and *E.coli*. As I mentioned PulseNet for these pathogens in 2016 started to use WGS. Initially NARMS at CDC was phenotypically testing a subset of those collected through the PulseNet collaborations. Now with WGS surveillance of human illness but now we're able to, arguably, using whole genome sequencing to better predict AMR and understand the spread and tracking and attribution and source of human illnesses.

Also, you'll hear about this over the course of the next few days, about the NARMS expansion to One health. So, we have retail meats, we have food animals at slaughter, and we have human illnesses. Under a One Health framework we're going to start looking at things like surveillance of surface water. Again, you'll hear more about it, but I want to emphasize that today's focus is on the technical data sharing so as part of that, certainly NARMS is thinking about which sample the to collect and where to collect and also how to share them.

We're working with groups now that are doing things like wastewater surveillance and making sure whatever samples we collect, we're able to share them through resources like NCBI. So, in addition to making as much of the raw data available from NARMS sample collection as we can, we're also committed to making our water data accessible. Obviously, it will not exactly fit into the classic NARMS framework with retail meats, but we are thinking about how to share metagenomic data too.

Briefly, I want to mention that NARMS has been around a number of years and things are generally stable, but methods or breakpoints may change over time. If you are working with the NARMS data and you see something unusual, before you write the paper, just reach out to us if you have questions. An example I have here. If you look into the number of *Salmonella* we get from retail chickens as part of our sampling protocol, here, you can see looking back from 2018 to 2008, we get two or three times as many positive for same number of samples. If read the annual report, you can see we made changes to the methodology and we have incorporated overnight enrichments and increased the sample size for the meats.

So once again we're trying to make this data as accessible as we can but if you have questions about it, reach out to us or to the CDC if you see something unusual. If you have questions, it's not always easy for us or possible for us to incorporate some of these subtle changes in the tabular data we share.

Briefly, I would like to touch on whole genome sequencing and antimicrobial resistance. A nice aspect of WGS is that you're not waiting for the sensitivity results to be generated. For retail meats we often see a delay of between 2 weeks and 3 months depending on the lab and the workflow and the agency before data makes it into NCBI pathogen detection. If you have some knowledge of bioinformatics and microbiology, you can interpret the raw data at NCBI that includes the AMR predictions. Antimicrobial Sensitivity Testing (AST) reporting often labs 6 to 24 months. In general, for the pathogens of interest and mechanisms that we commonly see for the foodborne pathogens surveyed by NARMS, we can predict resistance and predict the AST very well with this data.

Another nice aspect of AMR prediction, you can see on the right not only through NARMS but through collaborations like PulseNet, is that nearly half a million *Salmonella* are available on NCBI. So just briefly, oops, sorry there. That was, I forgot I had some audio on that. So just quickly, I would like to cover some of the examples of a NARMS data flow.

As I mentioned there's multiple sources of data for NARMS and it would be great if we had one sort of one resource but it's an interagency program and depending on the question you're asking, you may need to go to different places on it. You can see on the top kind of the typical workflows for FDA labs. Each lab that is collecting the one month's worth of data, they're shipping to CVM. That information will be incorporated into the dash boards like Pathogen Detection and Resistome Tracker. Similarly on the bottom, we know through collaborations like PulseNet, data is going into the CDC and they're developing their own dashboards around resources like NARMS Now and at the same time they're putting it up NCBI. So once that data that has been submitted and NCBI as you'll hear about this afternoon, they will process it and typically within one or two days, we'll get the predicted AST as the set AMR genes. And once again, it's an amazing collaboration from my point of view because all of the different participating agencies largely share their data in real-time with NCBI. We see a small number of errors or mistakes but we can catch it quickly and it puts us in a much better position to catch emerging hazards where something has originated in the U.S. or came to the U.S. Here in the bottom right corner we see the larger integrated reports or summaries. Once again, there is some lag real-time collection of the data.

This is a one slide introduction to it if you are not familiar with whole genome sequencing, that's a term you'll probably hear multiple times today but what it means is on the left, we have a pathogen like *Salmonella* that has a single chromosome inside of the cell along potentially with some accessory genetic elements like plasmids. This genome will be roughly 5 million base pairs in contrast to human genomes which are about 3 billion base pairs. We'll take a culture of the *Salmonella*, grow it up, extract the different chromosomes and plasmid accessory elements, break it up in small pieces. Small pieces can sometimes mean between 10,000 and 40,000 for some of our long read efforts like back bio which you'll hear about later today, or closer to 500 or 800. I would like to emphasize that the NARMS labs, USDA labs, CDC labs where you try to harmonize on protocols and ensure data is comparable quality so you can see on the right, what is the typical, for retail NARMS sites, sequencing *Salmonella* on it.

We check the average insert size and we see that each of the reads is about 236 base pairs. In this case the insert size may be a little small but this is in the range of 500 base pairs for the first read. It's acceptable data quality for the NARMS program. If you do have more questions about this on it, the CDC PulseNet has a lot of good information about data quality and standards. For those who are not able to participate in PulseNet for the non-public health labs there the FDA NARMS and FDA GenomeTrakr groups have collaborated to publish thorough QA-QC guidelines. So, if you want to analyze data or interpret whole genome sequencing data according to NARMS standards, for example, you can see quality scores and assembly lengths here that give you a pass or fail on whether or not you should analyze this data and if you should even look at it for AMR.

The NARMS program is the world leader in making whole genome sequencing and antimicrobial resistance phenotypes publicly available. Unfortunately, only about one to two percent of the public genomic data has an associated antibiogram with it. So we're getting reports about AMR genes found in all organisms but without knowing the context of the gene, the genetic background, and whether or not that finding of the AMR gene would lead to a phenotype resistance, it’s of limited utility. It's important to have good training sets and good baseline datasets together and I would just like to emphasize, part of the reason why I joined the NARMS program is a world leader in generating data.

What you can see by this is historically, and currently what we do for NARMS, all of the isolates go through sequencing and phenotypic testing for the pathogens that we survey through NARMS. We know generally the molecular mechanisms would underlie that phenotype resistance. So, through sequencing, we can predict what that phenotype is going to be. We work to standardize the metadata. This is important because break points may change over time. If you're a data scientist or a software development and you want to develop an API to interact with the data, FDA is probably not the right place for that but because they have made it available through NCBI, you can go and grab the detailed antibiogram data, the MIC, the measurement and what platform was used. If you want to build these tools around the AMR predictions, that raw data is available. On the left, you can see, if you want to wait for the pathogen detection, we have an isolate but not a NARMS isolate but an isolate from FDA testing of a fish paste, a *Salmonella* Kentucky with extreme multi-drug resistant isolate. The nice thing about these is if there's a certain gene or resistance that is of concern, we can roll out alerts directly with the labs doing the sample collection. They don't have to wait on NCBI. As soon as the sequencing is done and they can upload their data to a shared cloud resource. Within a few hours, we get reports for the resistance from these organizations. So, these resources are available and I would like to emphasize, if you're generating your own data, you can use the exact same tools these labs are using.

I will briefly talk about on the phenotyping we're doing and the questions around not only predicting the phenotype but also potentially predicting the origin of these different organisms out there. We're using new tools like machine learning and artificial intelligence. The genomic data provides a wealth of information.

We're trying to use it. We want to understand, yes, this AMR gene generally predicts resistance but using machine learning for other promoters or other aspects of the genetic background to whether it's important for an intermediate or resistant phenotype. The raw data is public, our AST data is public, and, if you're familiar with resources like Github, you can download the code used to predict the MIC. We’re working on building better predictions of MIC and extending the models to a wider range of pathogens. I think we're out of time. So I will stop sharing but I would be happy to take questions as you get your screen sharing set up.

Thank you for the introductions. There's a question from Heather. She wanted to know if you can share what the links were for the AST metadata in NCBI.

Yes, we can share that. If you're familiar with NCBI, the antibiogram are linked to the BioSamples. I think a little later today, if that question is not answered, we'll hear more about NCBI data and BioSamples but once again, for the NARMS data subject to sensitivity testing, those antibiograms are submitted and linked to the data type. So thank you, Heather!

Okay, I am not seeing any more questions on it. I think with that, it is my pleasure to introduce Cong Li. So take it away. Cong.

Application of the NCBI Pathogen Detection Browser for FDA NARMS Retail Meat Sample Surveillance – Presenter Cong Li
Time- 27:50 – 54:25

Thank you! I actually joined CVM in 2012. That is one year before the WGS program. So basically, my whole career here is related to the WGS program. So it's exciting. I feel very lucky to work. I see the program grow in this field and I myself, witnessed how useful it is for the surveillance program by the WGS data. So today, my talk, my technical session talk is how we use NCBI pathogen detection browser for the FDA NARMS surveillance program. So my talk has two sessions. The first one is how we submit the data and the second section is how we use the pathogen detection browser for similarity of the isolates and AMR.

So the first, as individual submitter, we need to submit all of our WGS data to the NCBI. So, it's like, the NCBI is more like a library. So for us, it's like an author we write books and we submit them to the library and the library then reorganizes it and puts in more information for the readers in order for the readers to make sense of the data. So, for the data submission, we first need BioProject and second there, is the BioSample, and third is elements associated with the BioSample. So, the BioProject is more like a folder that holds all of the related BioSamples in terms of organism and the study and the date, anything you can associate it with the folder you would like to have. So the basic information is who submitted it and what kind of data type and what purpose and what title. So here are several examples from our NARMS program. This is the BioProject for the NARMS program. So the first four are organized by organism and the last one is organized by the study. The last one, the reason it’s organized this way, is because there are many organisms, even some unknown, so it’s not very easy to organize that way and by the -- it makes sense to put them together. So after you create the folder, you can put all of the BioSamples under that particular folder. So BioSample, when you submit the data, you want to tell the readers, what kind of metadata is associated with all of the sequence and like, who, where, who collected it and where you collected it. What is the host and what type of organism and what kind of information you're going to submit.

So the NCBI has some limiting information that would want you to submit with your sequencing samples so there are two packages. One is NCBI package and the other one is GSC package. NCBI package is what we used. The other package was developed by the European BioTechnology Information group. So here is an example of the BioSample submission with the NCBI package and you have to have a sample name under which BioProject and what kind of organism. You don't really have to have an organism if you are not sure. And then what kind of strain does this isolate sample come from and who collected it. So all of the asterisk, the catalog, you have to put in some information. You don't have to know it. And if you don't know it, you can say not available. So under the BioSample, there is an important data type that is antibiogram as Errol mentioned. There was only one or two percent of the BioSamples have this so I think we are a big contributor for this type of data. So there are currently 300 drugs included in their submission template. So when you submit it, if your drug is not there you have to first request it. So this is all a drop down menu so the people's format, the submission format is uniform so you later can make sense of this easily. And also, Errol mentioned that MIC has a different standard so this one you have to say which you used for the cutoff to see if it's resistant or not.

So after all of this metadata, you will be ready for the sequencing data submission. And there are many different types and you can only, the list, you only need to submit the SRA data and the SRA is short read archive and it can be directly downloaded from MiSeq or NextSeq. There are also other higher quality of the data which you can assemble genomes and you can submit certain type of sequences such *Salmonella* genomic islands, pathogenic islands, and also annotations. But NCBI will also do an assembly for you, especially for short reads you don't need to submit your assemblies.

So after we submit all of the data and we want to know what we can do with all the data. We want to check all of the other authors that all write the books. We want to know how other people's isolates are associated with your own isolate.

You can search for our *Salmonella* BioProject and there are a total of seven thousand isolates and the assemblies are close to 6000. They also populate the publications associated with these BioSamples and BioProjects. You can also search for BioSample, for example, and so this is one of the samples we submitted, and here is the antibiogram and at the bottom you will see the SRA and nucleotide assembly, and this one happened to have two types of reads. One is from the PacBio and one is from the Illumina short reads. So this is associated with the raw reads you can try your own assembly, but we have submitted assemblies as well. So if you click nucleotide there are two contigs associated with this BioSample, so one is the plasmid and another one is the chromosome. They are all closed circular DNA. If you click this, which is plasmid sequence, all of the annotation is here as well.

So this is what we can see but with our data. So next one we would want to see all the data. So next one, we would want to see what we can -- not everyone is a bioinformatician so not everyone knows how to handle the data but that is not -- no worries because the NCBI offers pathogen detection browser. This is actually a portal embedded with a lot of tools that basically tools for everyone who is interested in the data to see, to make sense of it. So let me see if I can pull up a demo here.

So this is their homepage for the Pathogen Detection browser. And here’s what Errol mentioned. There are a half million of *Salmonella* strains and a quarter of a million of *E. coli* already today. To support this browser, there are embedded tools and data base under. So there is a MicroBIGG-E. You can search for the individual isolates. So this is a sample that you just saw and we looked it up in MicroBIGG-E. So this actually gave all of the resistance genes, the stress genes, and the virulence genes. Not only the will they tell you which contig the gene is located and the position it is, you can download this and look at the content more closely if you're interested in transposon and integrase and gene cassette and so you can look at it yourself. You don't need to do the assembly and know some, do any of the analysis with it.

So this portal gave you a way to search. For example, you can search for all of the new *Salmonella* submitted to NCBI. So the pathogen detection browser updates the database twice a week so every three days you will see the new data show up. So, you can pull up the search term as *Salmonella* and then the new, this is the Boolean value so if you see one, this is the newest submission that shows up here and you will see all of the new *Salmonella* showed up here.

And I would give another example here. So this is another search for *Salmonella* and AMR and CTX-M-65. Probably many of you are familiar with this gene. So this has a big plasmid with third generation ESBL genes. And so we want to know how many are there and so this is, as you can see, it is Infantis strains and you can see other information here. The more interesting part is they put, the way they organize it is they put the cluster here so this is the biggest cluster so you can actually click on the cluster. It will show all of the strains which will satisfy your search term and all of the other information associated with this. Like, AMR genotypes and virulence genes and stress genotypes. So for the left is the isolates selected. So all of these isolates are actually satisfying your search. And this part is the SNP tree and it's basically using KSNP method to cluster the strains. So this is all closely related strains which has SNPs of 50 nucleotide distance. So all of them are here. If there are so many of them it's hard to see in this browser. What you can do is actually download what we can see first. Before we download, you can choose more information. So on the left is what you can see with what is displayed here. On the right, you will see more information you can put in. So for example, the N50 is not here and you can actually put it wherever you want. You can say, okay, and it will be there so you can see how well the assembly is. And after you choose all your data type, you can download this data sheet which has thousands of the data points to your desktop. So you can download and you can choose the tab delimited table or you can choose the excel table. You can see there are more than 10 thousand strains. So here, you will download this. You can open this. So all of the strains under this SNP cluster will show up here and including all of the data you chose to read. And they also gave you all the assemblies. You can download it directly. You don't really need to assemble the data. And you can also watch the isolates. So you want to know what isolates come in every time? There is an update and so you can actually choose, you can give a name and so this is what we have and we're watching and there are a lot of functions here you can explore and it's quite useful so we use it daily.

So this is actually reflecting the real-time how the data looks like as long as you submit the data and two or three days later, it should populate on this browser. You can see your isolates. So you can also see, let me see here. You can also put in the search for only new ones to see how this cluster go, how this cluster is involved so you can see all of the new isolates show up here. So there are matched isolates, 7 of them and matches clinical sample is 4, match the environmental sample is 3. This cluster shows up here. So it's a quite useful tool for us.

So that is about all the functionality I can show you. But there are also, probably, it's good to show the other database you can explore and so let's go to the home page. So MicroBIGG-E, we already showed to you. This is the reference gene category catalog so this has, I believe, more than 8,000 resistant genes and mutations. So those are all manually curated from publications, and it's all validated so it is -- you can search for the name. Such as we just search for the blaCTX-M-65. So here is the gene and all of the information that it's associated with it. So here the class is beta-lactamase and cephalosporin is the sub class and it’s listed as AMR. and let me just search for it. You can put a star here on the search. So you can put it here. So of this CTX-M genes, you can put it here.

So this is, another one is called the reference gene hierarchy. So you would want to know which resistant gene is under which cluster. So, again, lets search for blaCTX-M-15. You can see this is under blaCTX-M-1 resistant genes so this is where it belongs to. So this will give you a picture of how similar it is to other CTM-X genes.

So this one is the reference HMM catalog. This is actually just using the HHM model to search for the function region to predict the resistant genes and because sometimes, the sequence doesn't match with the reference 100%. So how do you predict the functionality, how do you predict the genotype, this is the search model for the protein functions.

So this is the last one on the isolates with the resistant genotypes. This is all it has. Always the antibiogram information here. So yes, that's it. So the last one, I just wanted to mention, there's a search term and you have to be careful because all of metadata listed here, it can be searched but you can search with a star. You can search with a space and if you want it to exactly match, you put the quote on the term. So that is it. So I will end my talk and say thanks for listening! Thank you! If you have questions, let me know.

>> Thanks! We can wait a second and maybe while people are typing in their questions on it, I know you did cover this submission on it and talked about it a little bit. But I think the key thing is, maybe too, this is really, I can say for this for me personally and for you, this is how we look at the data every day. So the publicly accessible view of the data is the one we're using to sort of look at the short term trends and find potentially emerging resistance of concern and really track that on a daily basis and also, potentially look for links to other types of human animal illness there too. I do not see any questions coming in so if you want to stop sharing here, we'll let Jason get his screen up.

For this section, I'll be covering CDC isolate data, how it gets from state public health labs to NCBI, to PulseNet and then eventually to NARMS. I'm Jason and I lead the applied research on the lab side of CDC NARMS. But I thought first, I would jump in by giving you a brief background into PulseNet. This is the national network of 82 state and local public health food and regulatory agency labs. They detect food-borne disease case clusters that may be outbreaks. They provide real-time molecular surveillance of bacterial foodborne diseases and those include *E. coli*, *Salmonella*, *Shigella*, *Listeria*, *Campylobacter*, and *Vibrio*.

They collaborate closely with foodborne epidemiologists in state and government agencies to investigate these outbreaks and they act as a rapid and effective means and communication between the public health laboratories. This is just a map showing the PulseNet USA participants. This includes public health laboratories in all 50 states, Puerto Rico and food regulatory laboratories within the FDA and USDA. There are 7 PulseNet area labs that are located around the country to provide troubleshooting assistance, training, search capacity. And then just the three basic elements of PulseNet. The first is data acquisition. The second is data analysis. And then the third is data exchange. And this is a really basic schematic of whole genome sequencing and what happens on the wet lab. I won't go into the details at this time but specimens are collected, pure culture is grown, DNA gets extracted. There's some quality control measurements on that DNA. DNA libraries are prepared and those are loaded on the sequencing machines and then we end up with the raw data in which there's more quality control. I think more importantly for this group is exactly how the data analysis workflow works within the PulseNet national data base. So raw sequence data is collected at the public health laboratory. That typically is stored in private sequencing storage such as BaseSpace. At that point, the data gets submitted into the calculation engine for species identification and some contamination check and there's also a quality verification that happens there. Once the organism is identified, then that data is submitted again to the calculation engine but to the organism's specific data bases and there is where it gets allele calls, genotyping results and that include serotyping AR and virulence factors. Quality again, is verified. That then gets uploaded to the national database. And at the same time, it also gets submitted to NCBI as well. Just quickly for submission to NCBI. This gets done at the state lab. They create NCBI submission templates and they submit BioSamples and sequencing data to NCBI. That cold be locally linked data or BaseSpace linked data. They monitor the status,retrieve NCBI accession numbers and they look for troubleshooting errors that can be within bionumerics or on the NCBI viewer software and they retract and replace BioSamples and sequences as needed. One question we get a lot is about metadata and what is the side that goes on the NCBI versus what doesn’t. So PulseNet has a memorandum of understanding and the terms of reference with all of the state labs and sites. And this is the list of minimal metadata that gets released to NCBI immediately upon submission. So that includes the strain unique sample ID, the organism group, the genus and species, the serovar or the predicted type when it's relevant, the isolation type, organism source, so clinical, food, environment, country of origin (which is typically the U.S.) collected by, so that would be the ID for the laboratory that is submitting the sequence, then collection date and year of sample with the isolate. This is just the minimal set of data that gets released. Certainly states are welcome to release additional information if they're able to.

This graph is just showing the number of the isolates that gets submitted to

PulseNet for each of the organisms that we're interested in.

So over to the NARMS side, I won't provide anymore background into NARMS. I think Errol’s already done that. So this is another question we get about how AR fits in this with PulseNet. Essentially, CDC gets AR-specific funding from CARB and a part of that money goes towards whole genome sequencing capacity at the States and other places and it's that sequencing data that is utilized both by PulseNet for a cluster detection and outbreak response but also by NARMS for AR surveillance and also our outbreak response and then identifying any emerging resistance.

So specifically, how that AR data gets into NARMS. So this is just a schematic of our workflow for the whole genome sequencing. So sequences get, isolates get sequenced at the state and public health labs or they get submitted to NCBI or in some cases, we sequence stuff in house. All of that gets submitted to the CDC computational cluster. That cluster also receives periodic resfinder and pointfinder database updates from the CGE’s Bitbucket and then once in the calculation engine, AR analysis is performed either in bionumerics or it gets run in our in-house’s work flow. That identifies all the resistance determinants. Those results that get imported into the NARMS database and that happens every night. And then typically the next morning, I see those imports they get approved and that's where in our database, the predicted resistance is calculated. And all of those results go back into our database.

So this is just showing what data gets in the database. This is the isolate level of view. You can see on the very top, there's a box. So this gets in at leastthis isolate level view. So you can see in this case, there was AST done on this isolate so there is an actual phenotypic pattern and then there is a predictive resistance pattern. And then any of the resistance genes or mutations that were detected.

So at CDC, how do we utilize this WGS, AR data or this AR data I should say? It's used for both outbreak investigations and for detecting emerging resistance and we really get two different sets of data. On the left side, we get the AR determinants and the plasmids. What do we use that for? A little bit of subtyping, a little bit of source attribution that we know specific genes may be common from a certain source. But a lot of that is used for genetic context, understanding what plasmids are there, what mobile elements and really, that's used a lot for detecting any sort of emerging resistance, concerning resistance.

On the phenotypic side, a little bit of historical comparisons to what patterns we have seen in the past. It can be used for some weak subtyping and weak source attribution. The issue there is just because it has a resistance pattern doesn't mean it/s going to be the same resistant genes. But mostly, this is used to rapidly tell whether or not an outbreak cluster, for example, is multi- drug resistant. If there's any sort of clinical resistance there and that really helps drive our epi (epidemiological) response to the outbreaks.

So where does all of that whole genome sequencing data go? It certainly goes back into the database and I showed some of that. It also goes into NARMS Now. And I’ll briefly touch on that and Jared’s going to talk about NARMS Now later in the day more. It goes into SEDRIC which is not publicly available but many of the state and federal scientists have access to that database and can view our data there. And then it goes into any of our public outbreak notifications. That get released.

So in NARMS, we actually have two ways in which we get isolates or isolate data. So you have a sick patient. They submit a sample to a public health laboratory and an isolate or specimen gets created. At that point, there's two ways. We still have our traditional NARMS surveillance system which depends on receiving isolates from those State labs for phenotypic susceptibility testing. So those isolates get physically shipped to the CDC and there's the sampling scheme shown on there. Those get AST performed on it. If they're not going to be sequenced from the state labs, we do those in house but the majority of those isolates do get sequenced at the state lab and those come through PulseNet. And through bionumerics and in that case, all of the sequencing data comes into our system and goes through Resfinder and bionumerics and we get a list of resistance determinants and that's where we get our predicted resistance on the sequenced isolates.

I did mention NARMS results that do get into SEDRIC so there’s a direct linkage to NARMS that syncs every two hours. Phenotypic and predictive genotypic results that get in there. And this might be a little hard to see. But this is showing we do get a resistance pattern and that's searchable. And you can go to each individual isolate. You can actually pull up these resistance cards to look at what resistance is there. I also mentioned that this data gets into NARMS Now. I won't spend time talking about it because as I mentioned, Jared will be talking about it. There's all kinds of aggregate data looking at trends and patterns over time for the sake of this talk since I'm talking about the isolate level data. I just wanted to point out on that page, there is a way of downloading all of the isolate level data. That's what this little button is here. You can either download everything related to your search or you can download it all. There's also a data dictionary.

And all of that really leads to what we consider our WGS enhanced AR detection and response. So currently, we receive AR data for more than 80,000 isolates a year. That's a lot of isolates and a lot of data. We know that WGS accurately predicts antibiotic resistance and really this is about having that data in real-time, really allows us to respond to any sort of outbreaks with concerning resistance or, you know, emerging resistance.

So I can pause there to see if there are any questions. If there's time, I was going to talk briefly about sort of the future of isolate based data. But I can't actually see if there's questions so Errol, you would have to let me know.

I don't see any questions but they say thank you for the presentation, so far. Jason.

Well, then I will briefly talk about the considerations and the future of isolate based data. So I know many of you are aware that bionumerics is set to go away in a couple of years. So you know, that's a challenge for us since we use BioNumerics and PulseNet to get our data. PulseNet currently contains full epi data for more than a million isolates and the current software system doesn't have the ability to really handle all of that data, plus BioNumerics is reaching the end of life in 2024. Right now, there's no other existing commercial software that is going to replace all of the functions in BioNumerics so PulseNet is working with our office of advance molecular detection and the data modernization initiative at CDC to build a new modular open cloud-based platform for molecular surveillance.

So there are some opportunities there. We look to create a one stop shop for the public health labs, develop an open-source, molecular workflow that can be used for multiple genotypic programs, in addition to PulseNet at CDC. We want to take advantage of cloud strengths to improve turnaround time on analysis.

And we hope to have greater flexibility to onboard new analysis workflows and have an easier way to connect other systems already in the cloud like SEDRIC.

And this is just a brief timeline of what we're calling PulseNet 2.0. So in fall of 2021, so around now, PulseNet is gathering technical requirements for the system. We are about to perform some analysis of the alternatives to PulseNet and we hope to start feasibility studies this winter and then develop proposed architecture in the next couple of years.

And this is just a schematic of the proposed infrastructure will look like for PulseNet 2.0. You can see there will be a series of pipelines that will be built and then data from instruments and the data will use an API to get into a cloud based platform. That will contain data and computer management workflows, containerized systems and cloud storage computing and then that will use an API as well to get that data outputted into share common data and then epi visualization systems or platforms, data dashboards and then there will also be a way of getting that data into external data receivers such as NCBI.

So that is PulseNet 2.0. The other thing I want to mention is, since we're talking about isolate level data, you know, up until now these systems have all been based upon receiving an isolate in order to do that sequencing. And we know on the clinical side that those days are probably numbered. With the increased use of culture independent diagnostic tests, we suspect that we will start losing isolates over time and in order to deal with that, we are trying to find ways to move beyond isolate based surveillance and outbreak response so these are two of the graphs showing the increase over time for the use of some of our CIDTs for some of our PulseNet organisms.

So why is that a problem? Traditionally someone gets sick. They submit a sample to their doctor. That goes to the clinical lab where the isolate is created. That gets sequenced and that sequencing data goes to public health, to us at the CDC. With CDITs essentially at the point of the clinical lab, these tests are performed on the primary specimen so no isolate gets created and that information goes directly back to the doctor and it's great for treatment purposes but for us trying to do surveillance and outbreak detection, it really hinders what we can do without that isolate.

So how do we plan on handling that? That is going to be through metagenomics. And specifically, we plan on doing this through something called highly multiplexed amplicon sequencing assays. So in this case we are able to amplify pathogen targets from a primary specimen. So there will be ways of identifying what the organism is, doing the same sort of genotype comparison to identify how these things are related. So identifying what alleles are there and then also identifying antibiotic resistance determinants.

We currently have an antibiotic resistance HMAS panel. There was a version one that contained a bunch of the resistance genes we see in NARMS. We have version two. We're up to 118 different genes covered by 823 amplicons, and then there's a list of future genes that we hope to add as well.

And then we're starting to look at how this is going to work in a public health lab since we realized that you know, we won't be the--CDC won't be the ones doing this heavy lift. It will be done in the state public health labs so there's a pilot that is already begun and that's in two states, Colorado and Minnesota. So they're looking at, you know, DNA extraction and amplicon preparation. Identifying amplicons. The future plans, more testing in different states and expanding the AMR panel and in is just some of the preliminary data looking at that AMR panel and showing that we can successfully detect AMR genes across the range of template concentrations so you can see that the AR that was in the isolate is in green. So that's from an actual isolate versus the AR that’s in this tool is in orange and then the off target is purple.

And I think that is the end but I will take any additional questions.

Hey, Jason. I'll give people a moment to talk but I think I would like to start with the first question, I think we have heard a lot about PulseNet but just maybe one bit of a tangent question for you. With the standardization around whole genome sequencing on it and your reference around AMR and furthering AMR research, has this led to any development, for example, like breaking down some of the silos that exist between, for example, the foodborne pathogens and hospital acquired infections on it? Is the CDC moving in general in this direction for AMR surveillance?

So for enterics, we have PulseNet. For everything else, for a lot of the hospital acquired infections, there's the antibody resistance laboratory network. There's certainly conversations there. My hope is with the DMI initiative at CDC and the ideas that all of these systems will need to be brought together, not only on the lab side but also on the Epi side. And they’re really saying, we're no longer going to have these silos of specific pathogens having their own system. Everyone needs to share a system and I think in some ways, it's a little bit easier for us in the AR side because we know these genes aren't just present in enterics. A lot of these genes are a concern in hospital acquired infections and for me, it's the gene or the determinant that we're interested in. I don't think the pathogen is as important. And then also with the move to metagenomics where you're getting away from having the organism specific workflow anyway.

I think it will make it easier for us to identify those genes across these different systems but then, we'll have a little bit more difficulty with linking that gene specifically to a pathogen and whether that really is important is still a big question.

Thank you, Jason! Time here for questions if anyone has any. Please type them in or questions from the other panelists, feel free to speak up.

And certainly if people have additional questions outside of this, I’ve provided my e- mail address. Feel free to e- mail me any additional questions.

Thank you, Jason and thank you Cong for those great presentations, introductions and overviews on how the data flows throughout the different systems. Not seeing questions come in so I think we have a break scheduled here. Let me get this back up.

We have a break scheduled until 10:45 so we'll give people a few minutes to get their coffee and we will be back at 10:45 to hear from USDA, FSIS. Thank you!

-BREAK-

It’s 10:45 and so I hope everyone had a nice break and welcome back. It’s my pleasure to introduce Mustafa Simmons and James Gallons. You have heard from FDA about how we generate and transfer data from the CDC about the movement of their data and the collaboration with PulseNet and now we can look at food animals and how that data is generated and then shared. With that, the floor is yours, Mustafa.

Whole Genome Sequencing (WGS) Data Usage and Availability at USDA-FSIS – Presenter Dr. Mustafa Simmons and James Gallons
Time- 1:43:45 – 2:08:52

Thanks a lot, Errol. I definitely want to say thank all the presenters before me, Errol, Jason and Cong because you all have done a good job going over what I was going to go over. So it’s really going to be a reinforcement of what you’ve already seen. As Errol said, I'll be covering over the whole genome sequencing usage and my colleague James Gallons is going to be covering the availability of US data.

So just a few take home messages. FSIS sequences 100% of their bacterial isolates from their verification samples as well as a subset of isolates from their exploratory and cecal sampling. We started doing this in fiscal year 2016. The sequences are available in real time and are available to the public at NCBI as we’ve seen from the previous presenters. FSIS utilizes NCBI to track trends of the public health concerns, such as possible illness cluster, emerging antimicrobial resistance and specific virulence subtypes. And FSIS, in addition to NCBI, posts isolate level AMR data from their verification samples and aggregate level AMR data both their cecal and verification sample available on their website. And my colleague Jay Gallons is going to talk about that.

As I said, I'm from FSIS, and just to go over our mission; the Food Safety and Inspection Service is responsible for ensuring that meat, egg, and poultry products are safe and that they are properly labeled and packaged.

One major way we accomplish this is through our inspection programs. In fiscal year 21, which was our last fiscal year, we sampled, we inspected over 165 million head of livestock, 9.6 billion poultry carcasses and 2.8 billion pounds of liquid or frozen eggs. So in addition to inspecting of the carcasses of animals, there's other activities that the inspectors are performing including looking into labeling. So from all of the inspected samples, we do receive some of the samples in our laboratories so FSIS has three field laboratories and of those, samples that were inspected, 130,000 samples will get shipped to laboratories for either microbial analysis, chemical analysis or undergo pathology. And of those 130,000 samples they are going to translate to roughly 800,000 test and of that, 800,000 tests there's going to be approximately 2.9 million data points or results.

So how does NARMS fit into that? I think Errol did a good job going over the history of NARMS and I'm not going to focus on this too much but as far as FSIS is involvement, we started around 1997 and we were working with ARS to get sampling from post intervention samples, which included *Salmonella* and *Campylobacter* to have phenotypic ASE performed on them. More recently, from 2013 to present, we started looking kind of preintervention and that was our

NARMS cecal program. And we expanded the organisms we looked at from not only like, *Salmonella* and *Campylobacter* but also *E. coli* and *Enterococcus*. I think this was important because we're looking at things that are further upstream so it's closer to the actual food animal. And then most recently in 2020, we began the NARMS expansion projects and still monitor the same organisms but we expanded the commodities that we look at and we expanded to veal, sheep, goat, lamb and Siluriformes and some other minor species. But all this data from 2016, undergo whole genome sequencing or WGS and that's where I will focus my talk.

So I know a lot of you have probably heard about whole genome sequencing in the past. I think sometimes when you hear we're doing whole genome sequencing in realtime, you may think we're finishing it in a day or two. Unfortunately, we still require that we have a confirmed bacterial isolate. So like Jason alluded to, we're not doing CIDT. We still have to have a confirmed isolate so that will put us at least 7 days before we can even begin the whole genome sequencing process so that's why our start to finish is going to go anywhere from 14 to 21 days. But moving forward from samples and more towards the sequencing side of things, we're going to begin sequencing similar to day 11 or 13 after sample was collected and approximately 300 isolates are sequenced each week across three FSIS laboratories. We have a laboratory in Athens, Georgia, one in St. Louis, Missouri and another one in Albany, California. And USDA FSIS's full sequencing protocol is available in the microbiology laboratory guidebook or MLG, chapter 42 which can be reached by going through the website and searching or using the QR code on the screen.

So once sequencing is complete, the public health specialist will determine whether the sequences are of sufficient quality that they are good enough for downstream analysis. All of these metrics and criteria that are used are found in the MLG chapter 42 and its corresponding references. I want to point out some of the corresponding references are PulseNet references because like Errol said, we came together to harmonize what we consider to be *inaudible* significant and this was through GenFS or the interagency collaboration of genomics for food and feed safety and the members include CDC, FDA, USDA- FSIS and USDA- ARS and USDA- APHIS and NCBI. Again, you can find chapter 42 though the the QR code on the screen.

So after we have found out that the sequences of usable quality, there are some things that USDA-FSIS will pull a lot of the sequencing data in house. We're going to pull out serotype for *Salmonella* and *E. coli* using SeqSero2 and Serotype finder, and we're going to pull out species for *Campylobacter*, and Enterobacter using BLAST databases using specific species genes. And we're going to pull out antimicrobial resistance genotypes for *Salmonella*, *Campylobacter*. *Listeria*, *E. coli* and *Enterococcus* using resfinder databases using BLAST and we're going to pull out MLST type using BLAST from MLST databases for *Salmonella*, *Campylobacter*, *Listeria*, and *E. coli*. One thing I want to point out is all of these tools are publicly available so we're not using anything that is proprietary to FSIS or GenFS.

So after we determine that our sequences are of good enough quality and we pulled out any relevant data we want in house, we're going to upload it into NCBI. I think Cong did a really great job explaining how NCBI is organized. I do want to share what FSIS shares in terms of their metadata.

So we publicly share the collection year, the collection state and we share some level of subtype, depending on organisms so for *Salmonella*, we're going to share serovar, *Campylobacter* we're going to share species and for *Escherichia* we're going to share Serogroup.. One thing very important to note here, we also have a strain identifier which is going to emphasize what we call an FSIS number this is important because we recently added the FSIS number to the public assignment specific datasets and that is discussed more later but the reason that's important is that's going to allow you to link this data on NCBI to even more metadata that is available on FSIS's websites.

So where is that data located on NCBI? So they did a great job of describing bioprojects and how they're a collection of sequences, it's almost like in a folder. Our Bioprojects are largely divided into cecal or non-cecal and then by organism. Here is the list of our Bioprojects that we use. One thing I want to point out is that, and Errol touched on this, is that we're actually one of the larger single submitters of *Campylobacter* data in the U.S. we do sequence a lot of foodborne pathogens. So once our data is available in NCBI and is there anything that FSIS does with it? Yes, absolutely! One thing we want to be able to do is not look at not only our own data but look at our data in the context of everything else that's been submitted to NCBI. Like I said before, we're going to look for things that are of public health concern such as emerging antimicrobial resistance mechanisms so seeing whether or not drug resistance has shifted from being caused by one gene versus another. And we're going to share that information with our non-partners and they're probably going to be aware of it in realtime as well. We're also going to look for possible repetitive subtypes so we want to look at whether the same *Listeria* subtype in an establishment or throughout a corporation and again, we're going to further discuss that with the FDA for a dual jurisdiction establishment.

And lastly we're going to look at whether or not isolates are closely related to clinical isolates based on their single nucleotyide polymorphism differences. I want to point out this is just the initial screen of our isolates. We would never want to establish a relationship based on genetic difference alone so we're going to pass the information along to our epidemiological staff and they're going to work with CDC, PulseNet and try to figure out exposure data and additionally, we're going to work with the office of field operations that can look into establishment history and trace back and we share it with FDA if there's a dual jurisdiction establishment.

One of the ways we are able to monitor these things in realtime is via NCBI safe searches. So Cong showed us how to use the pathogen detection, rather, to search for various things such as antimicrobial resistance and various serotypes, et cetera. Any searches you can perform there, you can save it and have it sent to you via e- mail on a regular basis. This is just an example on what a safe search looks like. So they're going to show you the search criteria used to generate the search, the list of isolates found that met the criteria and then provide any links to relevant snp clusters. So this is kind of the extent of what we do directly with NCBI. I do want to point out and I don't have it listed on the slide but in addition to uploading to NCBI, everything that we upload to NCBI that is a pathogenic organism, is also going to go into PulseNet via BioNumerics, so we're actively working with CDC PulseNet. We are Pulsenet members so they're getting data in realtime as well. With that, I'm going to turn it over to my colleague, Jay Gallons and he's going to talk about the data that is available on the FSIS website.

Yes, I’m James Gallons, I also go by Jay, so if you see my name either way, I'm the same person. I work with him within FSIS and I work in the office of planning analysis and mismanagement. I'm a data analyst, I'm not really a scientist so if I make any ignorant pronunciation mistakes, don't hold it against me. So this first slide here is not really meant to go through each one of these at all. There's a couple of points I just want to make. And that's all of our partners whether it's the FDA, CDC, APHIS, and FSIS we have our website or links that can get you to specific data points. Some links are similar where you are going to the same spot. Others are specific to whatever agency or department has that website. For example, FDA, well, first let me say, everything that is highlighted in yellow there is sort of similar. So you can see that the integrated summary report that comes out, you can get to that particular report through CDC, FDA, FSIS, et cetera. And also with NARMS now which is an incredible powerful tool that is really cool to dig into when I developed this presentation. I believe the slide in particular, I could dig into all of these agencies and see what they have available. It's pretty eye opening how much data is really out there. And it's definitely worth going into. I think these slides are being shared so this would be a good reference table to jump into the different data spots.

A couple of other things. One in particular is on the FDA NARMS website, we have for FSIS, and I think, Mustafa mentioned it, we have two types of samples we consider NARMS samples or antimicrobial resistant testing samples and they are what is officially called the verification samples. We also call them, I call them product samples and can be a little more descriptive, you can also consider them passive samples. Product samples are part of our verification testing that we normally do in our regular duties here at FSIS.

The other samples that we have are cecal samples. We assign those out and collect them and do antimicrobial resistance for anything that comes back that is positive that we can get isolates from. On the FDA NARMS website, they actually have some data we don't publish on our website but it comes from the samples we just talked about. And they have product data or what we say verification samples from 2006 to 2019 that is isolate level but it also includes the MICs and genes in it. That's a pretty good resource. We were thinking about doing the same and actually posting it more frequently, a little bit more up to date since we have the data. And that is something that we're going to be looking into. The other thing is the cecal data, we currently do not post isolate level cecal data, whereas, FDA does do it. They do not contain any establishment information but it's at the isolate level and it goes from 2013 to 2019. You can say there’s a couple things, you know, each Agency and department has that is specific to what they do. What I'm going it talk about in the next couple of slides is what we have on our website. We have a link to the integrated summary report and link to NARMS Now and then we have our product and verification cecal sampling information. Next slide.

So this is the product sampling, I’m going to say, verification sampling. We do publish the isolate results. You can see that there's a link right there to get you right there. And the antimicrobial resistant information is in the AMR profile. Which I'm sure many of you know what it is but those who don't, it's basically the profile of the antimicrobial drugs, phenotypically tested to which isolates are found to be resistant. Using the NARMS monitoring system panel five. The isolates displaying resistance to multiple antimicrobial drugs tested by the NARMS panel is classified according to the drugs with the highest classification of risk. The resistance profile when it says pansusceptible means that the isolate is not resistant to any of the antimicrobial drugs tested. There’s a link to FDA’s antimicrobial drug classification table.

Next slide. This is an example of what we published. This is one record. I actually pulled it from our website. From the raw chicken carcass sampling datasets and changed the numbers around and names around just to not put specific establishment on notice right here because really they shouldn't be. But this is basically would be one line. I had to break it up into three lines here. But this would be one line of data where we have established an ID. We use establishment ID as a good marker or indicator of the establishment. We tend, when you start using that establishment number which is the next column, it gets a little messy because it has letters and pluses and things like that. So we have an establishment ID that we use to link a lot of it together. You can see that it shows what project the sample came from, the form that the ID that the sample came in and then, our analysis. This in particular, we test for raw chicken carcass and *Salmonella* and *Campylobacter* and it will show you the serotime, the PFGE pattern, if it failed allele code, the FSIS number and then there's the resistance profile. And it will basically do the same thing for each one of the pathogens or indicators that we found as positives. Next slide.

We also publish for our verification/product samples a quarterly summary aggregated data based on whether the pathogen or indicator, I believe this is for *Salmonella* and what it does is, you can see all of the products that we will have data for even at the individual level with individual isolate level and then the total number, how many were susceptible and the percent that were pan- susceptible and then how many were resistant to one or two drug classes. So you can see that percentage and if they're resistant to three or more classes of drugs, you can consider that multi-drug resistant and it gives you an idea, overall and over time, if our multidrug resistance percentages are going up over time for the various commodities. Next slide.

We do this thing. What we don't do and I mentioned it before, we don't have cecal level data at the isolate level currently. We are in the process of looking into doing that. And with the addition of adding the establishment information to that isolate level. But we do have the same type of thing as we do for the product/verification sample and that's the quarterly summary and that basically does the same thing. It breaks it down by animal and ensures presented pan- susceptible percentage, one or two drug classes and then also, the multi- drug resistance. So that's really what we have data wise. We're planning on doing more with the isolate level and cecal samples. Next slide.

So I will conclude and make sure we're staying on time here. But FSIS developed a high throughput real-time sample characterization system and this includes bacterial characterization and the AST and WGS. For public access, we upload these genomic sequences to the NCBI in real-time. FSIS utilizes the NCBI to track trends of public health concern such as illness clusters, emerging AMR and specific virulence subtypes, et cetera. We also post the isolate level AMR data from all of the verification and product samples and we do an aggregated AMR data for both cecal and verification samples on our website. As I said before, in the future we intend to make NARMS data available to the public in a granular manner and usable format. Possibly adding some tableau boards and doing some more visual type stuff on the website but all of the web sites that we have out there is a treasure- trove of information. A lot of it we can link together and so yes, it's kind of an exciting time with all of this new data. So with that, I'll conclude and see if there's any questions that Mustafa can answer.

I think we'll get started here. Mustafa and Jay, you're safe. Mustafa, this is a question for you. I saw in your slide deck that you're using it now using SeqSero as potentially, replacement for classic *Salmonella* serotyping, looking for the future of WGS at FSIS, do you think it will extend to other types of characterizations and typing of bugs, I'm thinking down the road for potential pathogens. Do you have any of that in the works?

I would say yes in the sense we are doing that currently for the species identification so we switched for *Campylobacter* and we switched over to doing that using whole genome sequencing related results. The only reason why we don't do it for the majority of our results is because as you saw, the timeline for sequencing is bit longer so we want have results out as soon as possible. So for example, if our definition for an isolate or something is dependent on being a specific species, or serotype we need to know it well before the WGS is completed because the establishment probably already has the product on hold. We would like to be able to do that but we need the technology to be able to determine those things up front using whole genome sequencing and the current state of whole genome sequencing that we're using, we're unable to do that.

That's a great point! And I think your timeline is something that is easy to forget when working with this. If you can stop sharing, Mustafa. Let me get her slide deck ready here. So we can see your screen. So I will just say, we heard this morning about the classical monitoring if you will, for food, animals and humans. For NARMS and before lunch, we can hear from APHIS on the health expansion on that. Take it away Christine.

Antimicrobial Resistance Monitoring in Select Pathogens Causing Illness in Food-Producing and Companion Animals, The NALHN AMR Pilot Project –Presenter Dr. Christine Foxx
Time- 2:09:15 – 2:28:01

Awesome. Hi, everyone! Thank you for allowing me to participate in the NARMS technical workshop today. I would like to just introduce myself real quick. My name is Christine Foxx and I'm an ORISE postdoctoral fellow, mentored by Dr. Beth Harris who is one of the two associate coordinators for the National Animal Health Laboratory Network embedded within USDA APHIS and you'll be hearing from her in two days' time on some of the roundtable discussions around the NARMS goals. I just want to discuss how the national health animal laboratory network conducts our antimicrobial resistance monitoring. It's going to sound really repetitive given the excellent talks that everyone at CDC, FDA and particularly within FSIS have already given but just emphasizing the AMR pilot project that we have here. And looking at select pathogens causing illness and food producing and companion animals through the pilot. I'm happy to report also before I get started, that this has been listed, this pilot program has been listed as a USDA- APHIS priority goal so it's slated to actually become a permanent surveillance program.

Okay. So the NAHLN and participating veterinary diagnostic labs have been monitoring AMR in bacterial pathogens of veterinary interest over the past five years or so in several livestock and companion animal, including swine, cattle, poultry, horses, dogs and cats and we also have recently included *Campylobacter* sequencing. So we include caprine hosts such as sheep and goats now and the main objectives of this pilot project are mainly outlined in combating antimicrobial resistance action plan but we aim to develop standards to track the antimicrobial resistance at a national level and identify trends of interest to the veterinary diagnostic community so this includes laying out methods for determining AMR whether it's in lab using antimicrobial sensitivity testing on the broth micro dilution platforms or by whole genome sequencing. I will talk a little bit about more but we also identify standardization guidelines or areas of need for standardization and an interpretation of these results across labs across the country and then establish reporting mechanisms to share this data with other agencies and stakeholders but ultimately we hope to facilitate AMR stewardship and the judicious use of antimicrobials by clinicians, as we get a better sense of what is going on in sick livestock and companion animals.

So briefly, it's important to talk about how this data is collected from participating VDLs. Isolates must meet three primary criteria, and these are identification of pure cultures to the genus and species level and to the serotype level belonging to *Salmonella* enterica, and they must be associated with the clinical disease or diagnostic finding in the data isolation by the veterinarian diagnostic lab which is further validated by information on the anatomical source of the isolate. We also ask for representativeness in the national survey by isolating only once per year from a single animal source. Whether that animal source is a herd or flock, farm or household or owner and we require that laboratories validate this information by including host animal species and state of origin and assigning a unique identifier for each isolate.

Once isolates have met the requirements for inclusion in the pilot, participating laboratories conduct susceptibility testing to determine the antimicrobial resistance phenotypes for that isolate using broth micro dilution tests. We ask that labs, technically isolate 50 to 500,000CFUs per mill in the fresh or frozen double passage sub -of the isolates and explicitly instruct them to use commercially available Sensititre plates on the Biomic or Swin platforms as indicated by the host animal and bacterial pathogen information shown on the table to the right here.

Additionally, in NBS staff of the national veterinarian services laboratories, particularly within our bacterial reference lab, give an orientation to all new labs on how to read off those inhibitory concentration results from these platforms. They have specific protocols in place for reading trailing endpoints for an antibiotics and gram-positive cocci as shown here. But it's as important to note that laboratory staff and external labs do maintain ISO accreditation for susceptibility testing by participating in the annual proficiency test too, so really, test standardization and data accountability goes both ways.

What I have shown here is just some information on how we share this data each year in our annual reports and these are published on our website but I just wanted to emphasize across all of these host animal and bacterial isolates, we received a significant number of submissions from veterinary diagnostic labs in non-livestock, non-food producing animals such as horses, dogs and cats. And these actually produce novel information across bacterial pathogens that have always been of interest to veterinary diagnostic labs but aren't necessarily surveyed on a regular basis which Olga Ceric will talk about later in the Vet-LIRN program. And we report out longitudinal trends in the antimicrobial resistance as well as sample sources in those particular sample types so that you can see as the data is collected, year by year in the AMR pilot program, you can compare it against other year's data very readily.

We also report out minimum inhibitory concentration data. Shown here is an example from a companion animal species so dogs without urinary tract infection sources in *E. coli* isolates and you can see here that the minimum inhibitory concentration data from those SWIN and biomic platforms are reported out in a standardized fashion by the precise dilution and MIC values that come off of those platforms. They are then interpreted as red being resistant, yellow being intermediate susceptibility and green being susceptible isolates. We also try to be as transparent as possible about our data by providing a slew of footnotes that explicitly lay out how we interpret the data, where our sources are, so principally the CSI Vet01Sstandards for interpreting broth micro dilution data and also some exceptions to those rules that are grayed out because they have no standardized interpretations available quite yet. But instead of waiting each year for the data to be written up in a white paper report and published every year, we do also provide our data publicly. So I just want to preface, we rely heavily on laboratories to provide us with this MIC data against all antimicrobials included on the Sensititre plates, by exporting the data directly from SWIN or Biomic platforms. This data can be messaged by laboratories directly through their messaging systems their LIMS or email to us using a standardized excel macro templates so very low tech and then they -- our staff can then upload it into a secure data base or they can provide it directly to us in Palantir.

We perform data aggregation in Palantir and any necessary transforms which is a processing pipeline that has been fully automated and then send the finalized data for visualization and public display through our tableau dashboard shown below. So, this is the same information for dog non-uti sample sources that have been identified as *E. coli*, but you'll notice here that instead of just the one year's worth of data, this is actually aggregated from 2018 so the inception of the data collecting portion of the project all the way up to very recently here. So that covers our phenotypic antimicrobial susceptibility test data which is necessary for any isolate that we also conduct whole genome sequencing on. Isolates are sent directly to the NVSL, the National Veterinary Services Laboratories or sequenced by participating laboratories directly using one of the four sequencing platforms listed here that have been approved for use and they include Illumina iSeq and MiSeq, the Oxford Nanopore minION and the ThermoFisher Ion Torrent as platforms. Double stranded DNA from the isolate can be prepared for sequencing using manufacturer approved kits or the Quantabio SparQ, the Roche KAPA or SeqWell plexWell kits listed here for any lab that has specifically Illumina platforms. And then raw sequence data provided by labs is uploaded to a secure drive, a cloud storage drive in which our NVSL computational biologist or myself can use it to transfer the data to where it needs to go. Okay. And then raw sequence files submitted by laboratories are usually accompanied by assembled data in fasta file format but since we have incorporated sequences from long read platforms such as the minION into our data stream, we also found the need to incorporate UniCycler, De novo assemblies, as well as spades assemblies into our pipeline. All of those submitted sequences are screened for quality and mean coverage to the reference genome that matches the isolated bacteria listed by submitting laboratories on the chart shown here to the right. But generally speaking these guidelines closely match what we see in surveillance efforts led by other agencies where *E. coli* isolates have slightly higher coverage requirements than other bacterial species included in the pilot. Additionally, we also screen isolates for sequence identity using Kraken which maps sequences to that of reference genomes and curated databases such as NCBI so the data streams really kind of circle around and link up with one another quite nicely. And this way we can verify that the cultures are pure which in this example is not strictly true in the chronograph shown here. After we have screened the sequence data for quality parameters, we then run the assemblies through our internal AMR pipeline, which calls ABRicate or AMR finder against the NCBI plasmid finder and RESFinder databases to detect AMR genes. Those data are primarily used for internal research projects or shared with academic partners but we also upload our data to an umbrella bioproject in the NCBI for data sharing with the public. Sequence data is also mostly publicly available. This screen shot shown here to the right is a little blurred out because this is an example of a sample shared already. I covered up the isolate ID because it contains some information about the geographic sequencing, sort of the isolate at a glance but, each sequence include Bio-Sample information about the isolate, and also technical information about the sequences as well. So in terms of Bio-Sample information, we share metadata such as host animal source, data isolation, the source tissue, and clinical disease or diagnosis associated with the isolate. We also include the names of the raw paired end reads so you can see the platform, the library kit used and sequence quality or identity in the SRA or the Sequence Read Archive and I'm happy to report at the time of writing, the NAHLN has uploaded the data spanning the pilot project up until two years ago so 2018 through 2020. And we're actively uploading more data minus the information that is being analyzed as part of the research projects and publications with external groups over the last two years.

So our goal is to make this data as transparent as possible and publicly available as soon as possible. But to be fair, there are ten bio projects in this umbrella that span all 7 organisms mentioned before as well as the three *Campylobacter* species of interest so *C. jejuni*, *C. coli* and *C. fetus*. And we have about 2,250 paired-end sequences corresponding to individual isolates uploaded to NCBI at this time, so like Mustafa mentioned, it's a try to throw them to you and there's quite a bit!

So this begins the question of, you know, aside from providing the raw sequence files to the public for download, how can the public also use this data to detect AMR? We have already heard a lot about the pathogens detection isolates browser in NCBI that you can search. So in addition to searching by isolate species or by particular antibiogram resistance phenotype, you can also search by bio-projects directly so an example here is again the *E. coli* dataset which spans all of the 7 host animal species that we do survey. And you can search any sequences belonging to bacterial species with small genomes using any other number of search terms as has been covered before. So minus the step of figuring out what AMR genes confer resistance to a particular antimicrobial, really pulling out the bio-project really gives you all of the data you may need to identify AMR in a metagenomic dataset in our survey. I'm going to say a quick thank you to my support of colleagues and mentors in the NAHLN Program Office, NVSL scientists who screened and sequenced isolates from around the country and the NAHLN laboratories who have participated in the pilot program over the last few years, we were 28 strong when I started and we're 31 strong now. Also, before I take any questions, I would just like to invite you to scan the QR code below and learn more about the AMR project and the strides we have made towards AMR stewardship. Learn about the labs that specifically contributed to the data that I have shown you briefly and annual reports on AMR trends and the other bacterial isolates in this study and up to date MIC tables and interpretations that were screen shot here. Thank you very much for your time! I'll be happy to take any questions.

Thank you, Christine and we'll give people a moment to put a question in the chat or Q & A box if you want. If you would be so kind to stop sharing, Christine so Olga can get her slide deck up. Maybe while people are typing, I think it's wonderful you have everybody on board with data sharing and putting your isolates at NCBI and also, it's just fascinating that you were able to work through a range of platforms that you have there. So do you have a rough idea or a breakdown of how many labs are maybe sharing, torrent versus Illumina data and how you work to integrate that? Especially for things like the mobile elements or plasmids. We are just understanding the phase of that AMR gene and whether it's the chromosome or plasmid is of critical importance.

I can only answer part of that question unfortunately since we don't, as of right of this moment, we only have about one laboratory of the 31 submitting data from ion torrent and it is, the data they have submitted has consistently exceeded our expectations for the Illumina standards that we published on the coverage of the sequence identity. That being said, there are about 2 or 3 laboratories that are just rearing to go on submitting the minION data from the Oxford nanopore series of platforms and with those, we have really been looking through the background literature from academic groups on minION versus Illumina sequencing standards so that is actually something we have developed and stood up a working group and the need for standardizing bacterial identification submissions from. I hope that makes sense. But yes, we are actively working on that effort to kind of lead the charge on incorporating data from as many sequencing platforms as possible because we understand not every laboratory has the bandwidth or the funding necessarily to have multiple platforms on hand and they may reach for the one is less pertinent for their field needs or their on-site sequencing needs. So we really want to be as inclusive as possible on that front.

I think it's amazing! And it fits in very well with Mustafa's comments on timing. Certainly for us in the MiSeq platform, it's great if you have 30 to 50 isolates and sometimes to process it but some of the other platforms have dramatically different turnarounds, it's just wonderful to see how you have incorporated it in your system. I do not see any other questions in the chat or in the Q & A box so I think with that, Olga we can see your slide deck.

FDA Vet-LIRN AMR Monitoring Program Dashboards and WGS – Presenter Dr. Olgica Ceric
Time- 2:28:15 – 2:50:27

Okay, thank you! Can you hear me? Okay. Thank you! I will be presenting on the FDA Vet-LIRN AMR monitoring program dashboards and WGS. Vet-LIRN stands for Veterinary Laboratory Investigation and Response network. We are located at the FDA's Center for Veterinary Medicine within the Office of Research. Now, my presentation has two parts. First, I will be talking about integrated AMR data dash boards which is a collaboration between Vet- LIRN, NAHLN and NARMS and then in the second part, I will be focusing on Vet-LIRN AMR program, specifically on the overview of our sequencing. So this diagram highlights the pathogens monitored by Vet- LIRN, NAHLN and the NARMS AMR programs. The NARMS part of this slide is only focused on the animal side of the monitoring for NARMS program. As you can see, each of the programs have pathogens specific to the program and there's a few pathogens that are common between two or all three programs. So for the first part of the presentation, I will be focusing on these pathogens that are common for Vet- LIRN programs as they are a part of the integrated AMR data dashboards. We formed the cross agency collaborating group to develop the centralized data collection and reporting process across participating laboratories from both networks. This group consists of members from FDA's Vet-LIRN and NARMS and USDA's NAHLN. More than 40 laboratories from Vet-LIRN and NAHLN provided antimicrobial susceptibility data on *E. coli*, *Salmonella* species and *Staphylococcus intermedius* group species in dogs, using commercially available testing platform. Laboratory sequenced the subset of the isolates and then submitted the whole genome sequencing data to NCBI. Importantly, both networks followed clinical and laboratory standards institute, the CLSI, antimicrobial susceptibility testing methods.

This is how this basically looks in a flow diagram. This is the process. In the first step, Vet-LIRN and NAHLN laboratories collect clinical isolates from dogs for this specific project. They are select bacterial pathogen, routinely isolated by the veterinary clinics and diagnostic laboratories. The next step-antimicrobial susceptibility testing, is completed using commercially available testing platforms and plates and then the data is reported and standardized. Next, laboratories sequence the subset of the isolates and then upload the data to NCBI where it's publicly available. With regards to the timeline, the first integrated report for 2018 data was released in December of 2020 as a part of NARMS integrated report summary for 2018. This was actually the first time that integrated AMR monitoring data from dogs collected from FDA and USDA network became available in the United States. The report included dashboards with minimum inhibitory concentrations, MIC data, for approximately 2,300 isolates.

Then the second joint report for 2019 data was released in March of 2022 and this one included MIC data for about 4,000 isolates. This time the dashboards also included the resistance mechanisms from genomic data. The third report for 2020 is pending. We are currently working on it as a group and planning to release it soon by the end of this year, and the report will include data for approximately 4,000 isolates. I will show you the demo of the integrated dashboards and they can also be accessed using the provided link or the QR code. I hope you can see this. This is how the first page looks like.

There's the background on both Vet-LIRN and NAHLN AMR monitoring programs.

There is more information about the Vet-LIRN program, with a link to our online page as well as more information about NAHLN and the link to their online page. Next one is the page with participating laboratories. Here in red, the map shows the NAHLN participating labs. The green is for Vet- LIRN laboratories and then the purple color is for Vet-LIRN and NAHLN laboratories and this is because some of the labs participate in both programs, but we make sure that the isolates are not duplicated.

Then the next page is the positive isolates by bacterium and location. Here we can filter and select the bacterium we're interested in. We can choose between *E. coli*, *Staph pseudintermedius* and *Salmonella*. I am going to click on *E. coli* as an example. Here, we can see that in 2019, the total number of *E. coli* isolates for both programs was 1,775. And then in the graph below, we can observe the laboratory locations and the number of isolates collected by NAHLN or Vet- LIRN or both programs. The next page is with percent resistance with gene and MIC distribution. Here, we can also make our selections. We can look at the data for both programs or NAHLN or Vet- LIRN individually. I will click on all, then we can select the bacterium here and how we grouped it is, we can select *E. coli* or *Staph. pseudintermedius* isolates from the urinary tract infections or *E. coli* or *Staph pseudintermedius* from all other sites in the body. So for this presentation, I will select *E*. *coli* from UTIs and the first graph here shows resistance from UTI infections in dogs and we're showing the antimicrobials where we had the data from at least 100 isolates. So if we hover over each of the antibiotics, you can see the information about the resistance to that specific drug. And the specific break point for the drug.

So I will select enrofloxacin as the example. And here we can see that the enrofloxacin resistance was 12% in *E. coli* from UTI infection and you can also see the breakpoint. In the graph below, we can take a look at the resistance gene distribution of enrofloxacin and then in this graph, all the way to the bottom, we can take a look at the MIC distribution of Enrofloxacin with the red indicating resistance. Now, there's a few antibiotics with a footnote. This is really telling us that the break points have not been established for dogs for those antibiotics and that we used the breakpoints for humans to determine resistance. So for example, let's take a look at the tetracycline. Here, the resistance was 30% [corrected: actual resistance was 12%, not 30%] in isolates from UTIs and there's this breakpoint that is actually the human break point. Again, we have the resistance gene distribution and the MIC distribution with red indicating the resistance.

And then the final page is, the reference page where we are providing the break points for dogs or for humans. As I mentioned, where the dog break points have not been established. So we now have the data for 2018 and 2019 in the dashboard. By the end of the year, we will have the data for 2020 which will really enable us to start looking at the trends. We can monitor the data for trends in AMR phenotypes and genotypes to identify new or emerging resistance profiles to help monitor the continued efficacy of antibiotics over time and to provide information to all our stakeholders regarding these trends. So that's our ultimate goal with the dashboards. So for this second part of the presentation, I will shift the focus on Vet- LIRN AMR program and specifically focusing on the whole genome sequencing results. You will remember this slide from the beginning of the presentation when the focus was on integrated data dashboards and the pathogens common for Vet- LIRN and NAHLN programs and now we're looking to take a closer look at Vet-LIRN AMR program with a few selected highlights.

These are selected bacterial pathogens which are a part of the Vet- LIRN AMR program. As you can see we are very focused on the companion animal side. For example, *E. coli* and *Staphylococcus pseudintermedius* isolates are being collected only from dogs and the Enterobacter cloacae complex from dogs and cats. We also have a group of pathogens that we're collecting from other animal hosts. So of course, *Salmonella* enterica, we are collecting isolates from any animal host and the same is with *Klebsiella pneumoniae*, *Pseudomonas aeruginosa* and *Enterococcus faecalis* or *faecium*. We are also collecting the *Campylobacter coli*, *jejuni* and *fetus* in swine, poultry, cattle, small ruminants, dogs and cats and I also want to mention that we have a group of several laboratories collecting the AMR data from fish pathogens. At this point, we do not have any specific pathogen. We're just collecting any bacterial pathogen from the clinical cases.

Looking at the total number of isolates with data since 2017 when we started our program, we have more than be 15,000 isolates collected with the susceptibility data. More than 4,000 of those were sequenced. We have pathogen projects available on NCBI. So all of our data is uploaded to NCBI and can be viewed there. And here we provided the project numbers for *Salmonella*, *E. coli* and *Staph. pseudintermedius* isolates. We have six whole genome sequencing labs and each of them has their own project for each of these pathogens so they can be seen here. And then we also provided the link to Vet- LIRN umbrella project on NCBI where we have several additional projects for other pathogens that we're sequencing outside of our AMR project.

We wanted to show the overview of the sequenced isolates from *Salmonella* in our animal hosts. As we mentioned, we are collecting the *Salmonella* isolates from all animal hosts and out of a little over 1600 isolates that were sequenced the majority as expected came from cows, horses, pigs and poultry. However, we also have this interesting group which is basically labeled here as the other. But to expand on that, this is the group of very different wild and exotic animals, wild birds and interestingly, many of these isolates actually very often foster closely either between the group with the other animals or even with humans, within 20 SNPs, so that's interesting to mention. Some of the highlights for the project that we currently have in progress: We are working on characterization of our *E. coli* and *Staph. pseudintermedius* pan genome in dogs from our and NAHLN AMR project in comparison with humans and analysis of genetic components of isolates from dogs, correlated with the AMR. And then we are collaborating with various groups in FDA, CDC and USDA on multiple clusters related to *Salmonella* isolates.

We were able to sequence some of our aquatic pathogens. Currently all of 2019 were sequenced and at that time, we had 60. These are the project numbers for those isolates. And then we are currently sequencing our 2020 isolates where we will have more, slightly over 100. Now, interestingly when we first sequenced our aquatic pathogens and our data was uploaded to NCBI, we realized that many of those pathogens weren't the part of the pathogen browser, however, NCBI was very helpful, and they added those pathogens in the pathogen browser so they can be found there more easily. And then we just wanted to show you the most common AMR genes that we saw with our aquatic pathogens and there's really nothing interesting here except that when we first looked at the data, we were surprised to see that the great number of our isolates had this cphA gene which confers the resistance to CARBAPENEM but then we realized that about 135 *Aeromonas* hydrophila isolates were currently in NCBI browser and then only 5 of those isolates lack the gene, so it's a pretty common one. However, it's important to know for clinicians that production of the cphA enzyme is not able to confer the carbapenem resistant phenotype.

Now, we're also excited that we were able to sequence our *Klebsiella* isolates from various animal hosts and that data has been uploaded to NCBI and approximately 200 animal clinical isolates come from Vet-LIRN. We were able to sequence these with huge help from NARMS. 90 isolates come from urine. Here is the project number for those isolates. And then we were curious just to learn how many isolates, *Klebsiella* isolates were sequenced in the United States in total. At the time that we checked, there was slightly over 10,000 isolates from clinical, human isolates and environmental. Now, within environmental group, there were slightly over 500 isolates, and this group really includes the isolates from animal, feed, food, hospital waste- water, sink water, retail meat, and then also animal and clinical isolates. So in this group of slightly over 500 environmental, there were 263 animal clinical isolates and out of these, 200 is from Vet- LIRN. So we're very happy we were able to contribute to the collection of *Klebsiella* sequencing data on NCBI in this way.

This is the end of my presentation. I would really like to acknowledge all of the Vet- LIRN and NAHLN participating laboratories as well as the members of our integrated data working group from NAHLN: Beth Harris, Jennifer Rodriguez, Christina Loiacono, and Christine Foxx. From NARMS: Amy Merrill, Claudine

Kabera, and Gordon Martin and from Vet- LIRN: Sarah Peloquin, Jake Guag, and Gregory Tyson. That's all.

Okay, thank you, Olga for your presentation. We'll give people a moment to once again, type in a question to the Q & A or to the chat if you would like. If it's okay, Olga, I would like to start and maybe just, this is a question in here but Vet- LIRN has been a leader in the dashboards in releasing the information and the sequencing here for years. I think my question for you maybe as we start to share this data, you know, and you built that initial set of dashboards on it, were there views or options or functions that kind of evolved over time? Did we start off with the simple view and people wanted to see, you know, things differently? Track different resistance? Have things at the front of your dashboard site? I know you have heard and seen it all so I'm wondering if you have seen that change over time.

Yes! When we started in 2017 with only Vet- LIRN data, it was pretty simple and at that time, we were only showing the MIC data. Then when we integrated the data with NAHLN for our 2018 report, at that time, we were also only showing the MIC data. Now, we are showing for 2019 and we'll be doing it in the future, also, the predicted resistance data and then the future goal and future development will obviously also be focused on trends. We couldn't do it until now because we didn't have enough data but that's one component that we will be adding to our dashboards and we're always looking for our ways to improve them. Luckily, we have great example from NARMS and especially NARMS now. So that is something that we often check, you know, on how that looks like and how can we use some of that information and presentation for our dashboards. So you know, we hope that we will be able to develop them more and make it more simple for use, for general public, and really provide as much information as we can.

Thank you so much, Olga! Well, I do not see any questions coming in the chat or the Q & A but once again, feel free to put those in over lunch. If you would like, I would like to thank all of our speakers from the different agencies. Again, this morning on it. It's everyone's favorite time now. We're going to take an hour break for lunch. I would say, please, please stay in the meeting. Please join us afterwards. We showed you sort of the data points if you will, this morning, the WGS data flow, how the data is being generated. This afternoon, we'll get into really aggregating those data points, looking at trends on it and if you saw a couple of times today, talking about especially with the WGS, what does AMR look like when we apply these AMR tools outside of NARMS. We have the resistome tracker and AMR finder and some of the other data bases and tools for using it. So once again, please enjoy your lunch and we'll join back up around 1 p.m. So thank you!

-LUNCH-

Errol Strain is speaking. Hello, everyone. Welcome back to the NARMS technical workshop. Today, Amy, if you would like to get started screen sharing, I'll introduce you a little bit here. Thank you, Amy. Welcome back, everybody. While Amy is getting the screen and audio set up, I will say, this afternoon, as I mentioned earlier, we talked about the individual data points this morning, what goes into NARMS. I think this afternoon's session will help address some of the questions we had. We had questions around how current is the data on NARMS Now, and Amy will address that. And other things around some of the mapping potentially of the AMR related genes to the actual antimicrobial or species. The nice part about things like NARMS Now is the mapping is already done, you don't actually need to go back in and pull the data from Resfinder or AMR Finder. Amy has done that for you. I think with that Amy, we can see your screen. So, if your audio is working, the floor is yours.

FDA NARMS Tableau Dashboards
NARMS Now: Integrated Data and Strain Explorer –Presenter: Amy Merrill
Time- 3:59:20 – 4:28:46

Like Errol said, I am Amy. I'm a mathematical statistician for the FDA NARMS program. I work a lot on the data visualizations. Today I'm going to show you two of our tableau dashboards, the NARMS Now: Integrated Data and the Strain Explorer.

I'm going to start with NARMS Now, and just a little bit of background. NARMS replaced annual report tables with dashboards where users can explore the data in an interactive way. They did this a couple of years ago. NARMS Now is a collection of visuals that allows users to compare different antimicrobial resistance trends in bacteria from animals, retail meats, and humans. More recently NARMS Now started taking advantage of whole genome sequencing to predict antimicrobial resistance, allowing for users to have access to results in close to real time.

So, I'm going to start by breaking down the data that goes into NARMS Now a bit. NARMS Now is mostly comprised of different graphs and each of these graphs is broken into two parts. In this gray portion is the phenotypic resistance data. Phenotypic resistance is determined using epidemiological cutoff values or clinical breakpoints to interpret antimicrobial susceptibility test data. This data is updated after the closeout of that testing year for each of the agencies. There are Excel files available with that data on the integrated reports and summaries page on FDA.gov. This has the raw data, so it’s the isolate level data, not aggregated like it is in NARMS Now. So, moving on to the blue portion, which is where the genotypic resistance data is. Genotypic resistance is the presence of acquired genes and mutations known to enable a bacterium to grow in the presence of higher antimicrobial concentrations. This morning you heard about a lot of the partners submitting their genotypic data to NCBI. I go in and download this data from the pathogen detection browser on a weekly basis, usually on Mondays. And after I download the isolate level data, I link the resistance genes to the individual antimicrobial agents using a carefully curated reference gene catalog. Right now, we're continuing to verify the genomic results using in vitro susceptibility testing methods. So, after we finish AST and verification of the data for that testing year, the break in the graph gets moved over. We started doing this in 2018. This break in the graph used to be here in 2018 and then after we finalized the 2019 data, the break got moved over. So, everything in the blue portion is kind of considered to be preliminary. As we sequence more isolates, or re-sequence isolates, this data could change.

So, this is just a visualization of the flow of the data that goes into NARMS Now. This is from my perspective. There are other things going on as this is going on, but this is kind of how I view it. If we follow it in this direction, you can see how the genotypic resistance data gets into NARMS Now. Our partner laboratories collect the samples, then they isolate the bacteria and then they perform whole genome sequencing. They submit the sequencing data to NCBI. As our sites are doing that, CDC and USDA sites are also sequencing and submitting their data to NCBI. Once a week I download the data using some R code and I clean it up and join it with the reference gene catalog that I was talking about so we can link the genes to the antimicrobial agents. After I create this data set, I shoot it through the tableau prep, which you can guess from the name, preps the data a bit more. It cleans it up and makes it a bit more compatible with tableau desktop. It also creates an extract which is much smaller than what this file would have been, so NARMS Now works a bit faster.

Then following it in the other direction, you'll see how phenotypic resistance data gets added to NARMS Now. Again, the sites collect the samples and isolate the bacteria. They send those physical isolates to us at CVM and our lab team performs AST. This is happening on a rolling basis, so as more things come in more testing is done. We don't finalize the data until everything has been tested and retested for that year. So, after that, we clean it up and we put it in this Microsoft Access database. Again, as we're going through our process, CDC and USDA are going through their own. When they're all done finalizing their data for that year, they send it to us, and we add it to this Microsoft Access database. Again, I send it through tableau prep to get an extract and I put it into NARMS Now.

Now I'm just going to bring up the real NARMS Now to go over some additional features. So right now, I'm on the integrated reports and summary page. This is where you would be able to find NARMS Now, it's right here. On this page you can also find a tutorial that I made, which goes over how to use NARMS Now in a little bit more detail than I'm going over today. Then this is where you would find those Excel files I was also talking about.

So, I'm going to bring up NARMS Now. And this is what you would see when you navigate to that page. This is a -- basically a table of contents. You would need to click on one of these icons to navigate to the different visuals. So, there's a couple of different ones. There's antimicrobial resistance by year which compares resistance trends of antimicrobial agents for each bacterium. There’s resistance to multiple antimicrobial agents which compares resistance trends for any combination of antimicrobial agents. There's multidrug resistance by number of antimicrobial classes, which you can guess shows you multidrug resistance trends. There's also a map of resistance which shows the resistance percentages across the U.S. for a particular year. Then we also have some additional information. So, we have this first one which is the number of resistant isolates and isolates tested which is basically just the table version of the resistance data. And then we have this genes by antimicrobial agent and class, which is essentially that reference gene catalog I was talking about. So, you'll be able to see exactly what we're using to map those genes to the antimicrobial agent. Lastly there is this references page which just has some definitions and some additional resources.

So now I'm going to navigate to one of these so we can go over the features on each of the dashboards. So, I'm going to click on resistance to multiple antimicrobial agents. And it's going to bring me to that page. And all of the dashboards in NARMS Now are set up similarly. It will have this information icon at the top where you can learn a little bit more about the dashboard and how to use it. Then all of them have this series of filters at the top and that’s what you will use to view data of interest. Going through these on this page, I wanted to change the bacterium to something like *Campylobacter*, I would click on *Campylobacter* here. And now that has changed. If I wanted to look at coli or jejuni isolates, I could use this filter. And then the antimicrobial agents filter is where you select your combination of antimicrobial agents. You are capable of just looking at one, so if you want to see tetracycline resistance you would just keep that one selected and hit apply. But let's say I want to see nalidixic acid and tetracycline combined I'll hit both of those and hit apply. And now the graph has changed again. And right now, it's showing you isolates from chicken. I'm going to switch this, using this last filter, to look at cattle. For *Campylobacter*, we have beef cecal content and dairy cecal content isolates. I'm going to click on both of those and hit apply. Now you can see what that looks like for beef cecal and dairy cecal.

If you hover over a data point, you'll be able to see a little more information. So, this pop- up will appear, it says in 2019, 29.1 percent of isolates from beef cecal contents were resistant to at least nalidixic acid and tetracycline. As I was hovering, you may have also noticed that distribution of resistance genes was also changing below. Now that I'm hovering over 2022, the distribution has changed to show you the resistance genes in 2022. So, looking at that, it looks like for both beef cecal contents and dairy cecal contents, all of the resistant isolates contain tet(o) and gyrA_T86I.

Now I'm going to go through one more example of NARMS Now, looking at a different visual. Let's say I'm interested in looking at multidrug resistance in *Salmonella* from chickens. I would navigate to the multidrug resistance page. And then use these filters at the top to view what I'm interested in.

So, I said *Salmonella*, so I'll switch the bacterium. And right now, I'm curious at looking at all the serotypes, so I'll keep this as is. For the number of antimicrobial classes, I'm interested in using the definition that we use for MDR which is greater than or equal to 3 classes. I'm going to keep this as well. For sources, I said chickens. So, I'm going to unselect the sources I'm not interested in and select the ones I am. So, for chickens we have retail chickens, chickens the product verification samples and the cecal content samples. I'm going to hit apply. And now the graph is showing me what I wanted to see. I can look at the trends. It looks like around 2015, 2016, we started to see this increase in MDR *Salmonella* in chickens. If you have looked at our reports or other analyses, you'll know we attribute this increase to multidrug Infantis. So, let's look at that real quick. I’m going to unselect the other serotypes and reselect Infantis. Hit apply. And now we can see what MDR Infantis has been doing. It looks like from 2015 to 2019, you see a pretty large jump in MDR Infantis. It looks like as of 2022, it's still pretty high. Looks like we're above 90%. But it does look like it's going down slightly. But again, you have to keep in mind that we're still sequencing isolates and resequencing isolates so this could change.

So that is NARMS Now. Once you know how to use one of the dashboards, it's pretty easy to use the rest of them. So, I'm going to move on to the second dashboard, which is the Strain Explorer.

So, the Strain Explorer is a new data tool that allows users to explore emerging resistances. This tool is not available on our public website just yet. We are still working on fine-tuning it and figuring out the best vehicle for publication. The Strain Explorer uses the same data as the genotypic resistance data in NARMS Now, and it's also updated on a weekly basis. So, where NARMS Now is good for looking at trends, Strain Explorer is good for highlighting strains that we are kind of keeping our eye on right now. So, what it does is it looks at the different SNP clusters of related isolates from NCBI and pulls the ones where there are isolates from at least two different NARMS sources that exhibited one or more resistance of interest. An isolate is considered to exhibit a resistance of interest if it contains genes that confer resistance to one of the clinically important antimicrobial agents. Right now, it only includes *Salmonella* and *Campylobacter*. The clinically important antimicrobials for *Salmonella* are azithromycin, ciprofloxacin, ceftriaxone, meropenem and colistin. For *Campylobacter*, there's azithromycin, erythromycin, ciprofloxacin and meropenem.

Before I bring up the Strain Explorer, I want to go over the features a little bit. There are two pages. The main page includes a table of SNP clusters that meet the criteria I just spoke about. It also includes a map that's broken into the ten HHS regions that allows users to see where in the U.S. the isolates were discovered. Finally, under the map there is also a graph that users can toggle between showing the number of isolates submitted over time or the running sum of those isolates. There's also a second page which is just a table of isolate level details for one of the selected SNP clusters and also some links to NCBI.

Now let me bring up Strain Explorer. So, this is what it looks like right now. It looks pretty similar to NARMS Now where it has this information icon. If it will load. Okay. So, it has this information icon that will tell you a little bit about the dashboard. And then you just need to use these series of filters to view data you're interested in. The first filter is bacterium, you can look at just *Campylobacter* or just *Salmonella*. I'm going to click on *Salmonella*.

And then the second filter is where you would choose the resistance of interest. You can come at that two ways, you can look at the antimicrobial agent. So, the dropdown will include the clinically important antimicrobial agents. If you switch this to genes, the dropdown will change from showing the antimicrobial agents to showing the genes that confer resistance to those antimicrobial agents. I'm going to continue going this way. Let's say I'm interested in looking at isolates that are resistant to ceftriaxone and exhibit decreased susceptibility to ciprofloxacin. I'm going to add DSC and hit apply. And now the table has changed to show you SNP clusters that contain isolates resistant to ceftriaxone and exhibit DSC.

This is where you can select a time frame from where the last isolate with this resistance was submitted. Right now, it's set to the last three months. That means for all the SNP clusters we're seeing in this table, at least one isolate was submitted to NCBI that was resistant to ceftriaxone and exhibited DSC within the last three months. You can change this time frame with the dropdown, select a measure of time and type in the number. If I wanted to go back six months instead of three, I would just type in 6 and hit enter. And now this table goes back six months instead of three.

These next filters are all pretty optional. So, if there was a serotype that you were interested in looking at, you could use this one. Or if you were just interested in looking at a specific source, you could use this one. And the SNP clusters, if there was an SNP cluster you've already been keeping your eye on and you wanted to see if it met the criterion, there's this one. Finally, there’s this minimum percentage of isolates. Let's say you only want to look at SNP clusters where at least half of the isolates from NARMS sources show the resistance. You could use this and drag it to 50 or you could type in 50, hit enter, and now for these two remaining SNP clusters, at least half the isolates are resistant to ceftriaxone and exhibit DSC. I'm going to set this back just to go through the table in a little more detail.

So, looking at the table, it tells you what the SNP clusters are, and this next column is the total number of isolates. When I say total or all within Strain Explorer, I'm limiting that to isolates from NARMS sources. So, there's 1,300 isolates from NARMS sources in this cluster. The last time an isolate was submitted with resistance to ceftriaxone and DSC was submitted was the 9th of September. The first time was December of 2015. Looks like this is a Dublin strain with some humans and some cattle isolates. So, looking at some of the other SNP clusters, this Infantis one is the Infantis strain we’ve been keeping our eye on for the last few years. It is the one that's causing the increase in MDR that I looked at in NARMS Now.

Then there's this. This is what we wrote the first NARMS interim data update on. I'm going to use this as my example to look at the other features. So, what I'm going to do is, I'm going to click on the SNP cluster. And that changed the map and the graph below. So now I'm just going to go through time by using this right here. I'm going to click on the arrow. I'm going day by day and seeing where the isolates were found around the U.S. In the graph I'm also seeing cumulative total of isolates. It's adding what isolates appeared on this day to the previous ones. If I wanted to see what happened most recently, I would just scroll to the end. So as of the 10th of August, you can see there have been isolates all around the U.S. Looks like region 10 and region 2 are the only ones we haven't seen any isolates. Then looking at the graph, you can see that as of the 10th of August, there's been 138 isolates from humans that are showing this resistance within this cluster.

I'm going day by day; I find this more useful when you have a smaller cluster or one that is newer. So, this one's been around for a couple of years. We've seen this resistance since 2015, I think it said. Yes. So, what I would do is I would change this to year, so you can kind of get the bigger picture a little quicker. I'm also just going to switch this toggle so we can see the number of new isolates without adding them together. So now I'm going to follow this through again. And we can see in 2015 is when we were first seeing human isolates with this resistance. Looks like in 2016 we were seeing some market swine isolates. 2019, we're seeing some from swine, some clinical samples, these fall under Vet-LIRN. There is also a retail pork isolate which is what the NARMS interim data update was about. Following it through 2022, looks like 2020 was when we saw the most isolates from humans with this resistance and it looks like it's been going down since then. So hopefully we continue in the right direction.

So, another feature is when I clicked on this SNP cluster, there was also this pop- up. And it says, click here to see more details about isolates. So, I'm going to click here. And it's going to bring me to that second page. So now I'm seeing a table that has all of the biosamples that were included in the counts in the previous table, and it has some more information. It has all of the genes that those isolates contain, those are mapped to the different antimicrobial agents. If I wanted to see the details from all of the isolates from NARMS sources within the SNP cluster, regardless of the resistance, I could click this, and the table will change to include all of the isolates from NARMS.

And then finally, there's one more feature. If you click on the SNP cluster -- hopefully it will load -- well, hopefully it will, but if it doesn't, you would click on the SNP cluster, another pop- up would appear and it would say, click here to learn more about the SNP cluster at NCBI. You would click on that, and it would bring up the isolates browser that people showed you earlier today. In the search it would already have the SNP cluster filled in. Then you would see all the isolates that -- all the isolates that fall under the SNP cluster, including the ones that fall outside of NARMS, and you would be able to do more analysis from there using the isolate browser.

Unfortunately, it doesn't look like it's cooperating. That's what it would do. So that's it for the Strain Explorer.

Here I have the QR codes that will bring you to the integrated reports and summaries page, which is where you'll find the NARMS Now, the tutorial, Excel files, and the second QR code would bring you to the Strain Explorer. These you can also kind of Google "FDA NARMS" and you'll end up there. You can't find Strain Explorer, you'll need this direct link, because it's private on our tableau account. You can't just search and find it. If you missed scanning this QR code, feel free to email me. I put my email here. Also feel free to email me if you have any questions or you want to provide some feedback. I want to make these tableau dashboards as useful as possible, so I always welcome feedback. I quickly want to say thank you to everyone who helped put these together, it was definitely a group effort. I think that is it for me. If there are any questions.

Thank you, Amy. A wonderful presentation. We have one question coming in from the chat here. Is Strain Explorer available for chicken products?

So the product verification samples are not included yet.

I have to work with USDA. So hopefully that will be added soon before we make it publicly available.

Thank you, Amy. One question from me that is not in the Q&A. What's the level of difficulty with maybe adding another SNP cluster or potentially extending this to *E. coli* in the future now that we have the framework in place?

I don't think it would be that hard to add another organism. I think we just need to come up with the criteria so the clinically important drugs we want to include for *E. coli* but because the framework is all laid out, I think it would be pretty easy.

CDC Public Dashboards
NARMS Now: Human Data and BEAM – Presenter Jared Reynolds
Time- 4:26:39 – 4:59:28

Thanks everyone for having me. Good afternoon, my name is Jared Reynolds, I'm an epidemiologist and currently the acting lead for NARMS epi team here at CDC. I'm going to preview two public dashboards that CDC has. Most of the presentation will focus on the NARMS Now, the human data version. This complements the integrated version that Amy just showed. I'm also going to talk briefly about a new dashboard that was released earlier this year called the BEAM dashboard.

I'm going to start with a background on CDC NARMS routine surveillance testing. You heard Errol and Jason give a better overall description of this. But just to remind you, if you missed the morning sessions, it's going to feed into our live demo. So, we collect physical isolates that are shipped to us from all 50 state health departments including four metro area health departments, New York City, L.A. County, Houston, Texas, and Washington, DC. All those health department laboratories do a sampling of various enteric pathogens and ship those to NARMS, our lab here, for antimicrobial susceptibility testing.

Here's a little more detail on our routine surveillance testing. So Nontyphoidal *Salmonella*, *Shigella*, and STEC, specifically *E coli* serotype 0157. We ask the state labs forward us a sampling of one in twenty. The first isolate they receive in their public health lab followed by every 20th thereafter. That results in 5% of the total that reaches their lab gets forwarded to CDC for antimicrobial susceptibility testing. We also ask for typhoidal *Salmonella*, so serotypes typhi and paratyphi A and C, we ask for all of those. *Salmonella* paratyphi B, for historical reasons of difficulty distinguishing them, has been included in nontyphoidal *Salmonella* sampling. We also ask for Vibrio species other than *Vibrio Cholerae*, we ask for all of those. For *Campylobacter* we receive those, instead of all fifty-four NARMS sites, we only get those from ten sites that are in the food borne diseases active surveillance network or food net. That sampling varies by site based on the burden of *Campylobacter* in those states. That involves convenient sampling of participating clinical laboratories.

And so, for *Salmonella*, to give you kind of -- to put it into perspective of how many isolates our laboratory normally gets in a non-- Covid year, before Covid, just for nontyphoidal *Salmonella* our lab would receive 5 to 6,000 isolates per year for testing. That represents about 5% of the -- usually around 50 to 60,000 laboratory confirmed salmonellosis that reach our public health, state public health system.

And so, this routine surveillance data is what feeds one aspect of the NARMS human data displays. Then as both Jason and Errol talked a little bit about earlier, you've heard throughout today's discussions, whole genome sequencing is really enhancing our surveillance of antibiotic resistance among enteric pathogens. We were closely working with the PulseNet group, where whole genome sequencing is happening at all participating PulseNet labs around the country. Over two years ago, 2019, they fitfully transitioned from pulsed-field gel electrophoresis to whole genome sequencing as the standard for isolate characterization. That of course also helps the detection of outbreaks. And for us, it gives us the ability to screen the genomes for resistance determinants that allows us to make a predicted resistance designation.

And so, this is just a schematic, Jason showed this earlier, how sequencing has enhanced our surveillance. Now in addition to the 5% sample that we -- for example, for *Salmonella* we get at CDC for phenotypic testing, sequencing is happening in the states. It varies by pathogen. It's furthest along for *Salmonella*, it's over 80% of the isolates reaching the public health lab are getting sequenced. It's not as robust for some of our other pathogens. We'll talk a little bit about that later. And so, these isolates being sequenced in the state public health lab, we work with Pulse Net to get records of those into our database. They're screened through ResFinder, through the bionumerics Pulse Net database, the gene calls are imported into our database where our laboratory every morning approves the results to produce a predicted resistance pattern that can be compared to our traditional phenotypic resistance pattern.

Okay. Now I'm going to jump into the live demo, so I'll pull over our NARMS Now human data site. When you first go to the site, it's by default selected on a certain pathogen/antibiotic combination. In this case it is *Salmonella* Typhi, the antibiotic here, one of our -- one caveat to our antibiotics that's shown here, we normally just list single antibiotics. When you select these, it shows resistance to them. But for ciprofloxacin, because we see significant percentage of isolates that have decreased susceptibility to ciprofloxacin, which for *Salmonella* includes both the intermediate and the resistant designations, we do specifically show this in our displays and here you can see a tooltip indicating what this MIC range correlates to. But otherwise, if you select antibiotics, this dropdown will show resistance to that antibiotic.

But backing up a little bit, starting on your left in the search options, we have test method. And we have AST versus whole genome sequencing. AST is, again, based on the phenotypic testing done here at CDC on that routine sampling of isolates. Then whole genome sequencing is what we've added just earlier this year. And that will change these displays to update to predicted resistance. I'm going to start by showing some examples based off of our phenotypic AST testing. These are the bacteria that we test here at CDC, *Salmonella*, *Shigella*, *E. coli* O157, *Campylobacter*, and *Vibrio*. You'll notice a few differences between our site and the FDA's integrated site we just looked at. NARMS CDC does contribute data to the integrated site for S*almonella* and *Campylobacter* that we as an integrated group all perform surveillance on. CDC also looks at *Shigella*, *E. coli* O157, and *Vibrio*. Specifically for *Salmonella* Typhi and *Shigella* those are serotypes found in human reservoirs and are not found in animals besides primates for *Shigella*. But we -- so we do have some differences and some need to have our own site. You may be asking why the NARMS Now human data site even exists. That kind of grew up, as do many things in the government, in silos. So, we built this in 2015, before any of the other groups had an integrated interactive display. So, we've maintained this and improved this over the years. Then based on what bacteria you select; we do have different options for subcategories. If you pick *Salmonella*, we'll have different serotype options, including typhi, that’s the default, and the top nontyphoidal serotypes and some groupings of all nontyphoidal *Salmonella* or all typhoidal *Salmonella*. We have the antibiotic dropdowns. As of last year, these are able to be multiselected. You can choose multiple drugs; you can also choose preset patterns. If you click by pattern, the option will change, and you'll get some common multidrug resistance patterns with tooltips that indicate what these symbols stand for. And then we also have antibiotics that have clinical relevancy. So, there's a description on what drugs are included in these patterns. When you select these by patterns, this indicates resistance to at least whatever is listed, and the rest is a wild card. For example, in this case, if we were to select this option, it would be resistance to at least ampicillin and amoxicillin-clavulanic acid and ceftriaxone and then anything else. The data years that are available represent our surveillance. And so, for *Salmonella* Typhi, we began surveillance in

1999. If I change this to nontyphoidal *Salmonella*, it will update dynamically to 1996, because that’s when NARMS surveillance overall started. If I change this bacterium to *Shigella*, for example, we actually get a different option for our subcategory. We then look at specific species or all species as opposed to serotypes, which is a characteristic of *Salmonella*. The year range, again, can change based on when surveillance started for that pathogen. For *Shigella*, it started in 1999. I want to go ahead and switch off the antibiotic by pattern and look at it by individual. I'm going to uncheck all and select ciprofloxacin. We’re seeing increases in some clinically relevant antibiotics among *Shigella*. Especially among MSM, or men who have sex with men outbreaks, some of the drug resistance we're seeing is clinically relevant to ciprofloxacin, azithromycin, and even more recently ceftriaxone. So, I’m going to click on ciprofloxacin and show you how that all looks and run that.

When you run that, these below displays will dynamically update. And it's important to just realize when you go back up here to maybe change something, these don't change automatically. So, what you're previously looking at, you know, don't get confused if you've changed the selection up top. And so, our first panel is resistance by state. And so, this can be -- you can hover over each individual state to see the aggregate number of isolates that are resistant, in this case to ciprofloxacin. It shows both the numerator and denominator and the percentage resistance. These data are aggregated by state. In this default you they're actually aggregated by time as well. So, this is all the isolates that were collected, for example, in California during '99 through 2002. You can hit play on this to animate the map, it will change dynamically. In the earlier years, not every state was participating in NARMS but since 2003 we've been nationwide. This can be paused at any point to look at closer data at the state level. You can also scrub this to look for particular years. And the asterisk up here indicates we have preliminary data. In this case, we've marked our preliminary -- the preliminary data after 2019, we're still working on closing out a few stragglers from those years. We are pretty robust in those years, actually. We're actually, for 2021 is where we're really preliminary because we're still actually collecting isolates due to a backlog we've had since Covid hit. So, this map, the display can be changed to a table. So, you can get the individual year/state combinations, total number of isolates tested and resistant percentage. And this is not exportable directly, but this is very easily copyable. You can copy this and paste it into an Excel file.

The next panel shows resistance by year overall. Again, this is for the entire nation, as we've selected all states. There is an option to select individual states if you wanted to focus in on a particular state outside of the map. And so, this shows the percentage resistance. The gray shows the confidence interval around those. And this confidence interval will show up when you have a single drug selected. There are tool tips that are the same tool tip by year. It shows the year, number resistant, total tested, percentage resistant.

It also shows the lower and upper confidence interval of percentage resistance. You can see this obviously gets wider for 2021. We are still collecting isolates. Our sample size is a little bit lower than previous years. That results in the wider confidence interval. And we do as well have the designations to indicate what years were still considered preliminary. This display also has a table view. So, you can get just the data behind those graphs in tabular form. I'm going to switch that back to graph and select a second antibiotic to show you how that looks. Again, whenever you make a change in your search options, you need to go ahead and hit search again. So, when you select two antibiotics, there's going to be a second option on the resistance by year chart. By default, you're looking at the combination. So, you're looking at the percentage of isolates that are resistant to both azithromycin and ciprofloxacin. You can see how this resistance has really emerged over the past several years. The map is only going to show you the resistance to both of these drugs. There's not an option to look at those individually. If you want to, you need to switch back to single selection. And then this, in this graph, though, there is an option to look at these in tandem. And so, you can switch to individual to see both the drugs side by side. And these can be toggled on and off by clicking on the legend. And again, because of the possibility of multiple lines, we decided to not show the confidence interval here. But the tool tips are here to show you if you wanted to drill down to the data. The table then as well shows the data by year and drug combination with an individual view. And then the combination, it's the resistance to the numerical values of resistance to both.

The next panel shows the MIC distribution, so these are the primary data produced in our lab during antimicrobial susceptibility testing that are then interpreted to tell us if an isolate is resistant, susceptible, or intermediate. This display was launched earlier this year. This shows one drug at a time. It represents our 96 plate or our standard panel of what we actually test and at what dilutions. For azithromycin it shows we test MICs anywhere from 0.25 to 32. These are the CLSI breakpoints that are indicated by a green solid bar. Anything below that is susceptible. A red dashed bar indicates resistant. Anything to the right of the dashed bar is resistant. Anything in between if applicable is intermediate. Like the map, this is an animated graph. One of the benefits of this is to see how MICs have changed over time. For azithromycin, because of this emerging resistance, you can see how the MICs have creeped up over the years. I'll let that run. You can see the green bar is getting smaller and the red bar is getting taller. And then depending on what options you've chosen at the top, you can switch this to an individual drug, to the other individual drug, in this case it's ciprofloxacin. Again, this dynamically updates to show the different range of MICs that were tested as well as the different break points. And I should also mention there are tool tips. If you hover any of these bars, you'll get the MIC, the percentage within that MIC, the numbers, and the interpretation. And again, this, as well, has a pretty striking MIC shift over the years. So, like the other figures, this chart does have a table view. So, this is just -- in case you wanted to, again, easily copy and paste out of here into an Excel file, if you wanted to do an aggregate analysis of these data. The chart also does have -- this particular chart was built more recently with the different -- we had an export feature here. These two charts, if you wanted a picture of these, you could take a screen grab of them.

Then these are the quick stats. This just kind of shows the aggregate in total, whatever your selection was. In this case, it's showing the total number of isolates that were tested from '99 to 2022 for *Shigella* over the years and then the totals that are resistant to both azithromycin and ciprofloxacin. As Jason mentioned, we do have some download features for these data, and this represents the isolate level download. One caveat, as we talked a little bit about earlier, the metadata that we display in these downloads matches the metadata that had originally gone on the NCBI through the PulseNet agreements. We piggyback off those Pulse Net agreements that the states have signed, memorandums of understanding, to both -- that dictates both metadata we show in the download as well as what data actually feed into it. So, we are getting closer, but not every state has signed off on the latest PulseNet agreements. We're down to only three holdouts. We're working on getting those. If you download these data, you'll notice the total in the download may be smaller than the total shown in the quick stats, because these are aggregate data and they represent our entire surveillance, whereas the download is missing a few states to have not signed up, because the download shows isolate level data in some more detail.

I'm going to go ahead and download data related to my search. There's going to be a pop- up that says what you're actually downloading. There is an option to include preliminary data if you would like. There are other download features. This one says download all NARMS Now data. This is agnostic of your search. This is going to be a pretty big download, that's why there's a note telling me it will take a few minutes to run. There's an option to download NARMS Now WGS data. We'll talk about that. When you change your test method to whole genome sequencing to look at predicted resistance, that download feature will only include isolates that have been sequenced here at CDC or through Pulse Net that have been analyzed for resistance genes. Some isolates may include isolates that have phenotypic testing, sometimes those are one and the same. It will give you the total from the PulseNet sites that have signed the agreements. We'll go ahead and just quickly download this search that we've done.

While that's running, I'll actually switch the WGS method and show that -- switch the test method, excuse me, to WGS. And when we do that, these displays will update the data as well as description and it will change the predicted resistance.

We now see the predicted resistance by state and by year.

You can see these data are limited back to 2016. That's when we really started ramping up our sequencing here at CDC. Again, the displays are very similar in that they're showing percentage resistance. But in this case the isolates that are predicted resistant by the presence of a resistance determinant. The MIC distribution chart is disabled in this view.

I'm going to go ahead and open that downloaded view to show the variables available to us. We have our specimen I.D. We have the NCBI accession number if available. I'll sort on that to show you. There are ways to look those up in NCBI. The WGS, which is either produced by the NARMS laboratory, if we did the sequencing, which would be characteristic of older isolates, or now more recent isolates are those that have been sequenced through PulseNet.

We have -- whether the isolate is based off AST, WGS, or both. And the information on genus and species or serotype. The year in which it was collected. The region of the country it came from. These are based on HHS regions. Age grouping. So, we don't actually show the granular age. But we do show the age grouping. Specimen source. The resistance patterns. This is based on phenotypic testing. And then the resistance determinants, the listing of genes is that we found. The predictive resistance pattern based off of those genes. We have a variable that indicates whether we think the isolate lost its resistance plasmid that might account for discrepancies we see between genotype and phenotype. Then we have the actual individual drug MIC. These variables work together to form in this case the MIC is equal to 4, meaning sensitive. And then the conclusion that we found through predicted resistance. NPR means no predicted resistance.

So, the rest of the file just fills out all the drugs that we accounted for.

I'll quickly switch to the tabular view. This view is going to be really just a more interactive version of the data download. So, the total number of isolates that are shown in the tabular view, because this is isolate level, only includes those states with signed agreements. You might see this number to be slightly smaller than the number you see in aggregate displays. It's kind of just an excerpt of the Excel download. And then on the third tab are interactive reports that we've built and launched last year. If you all are familiar with the CDC NARMS reports, they used to be really long PDFs of just table after table after table. We stopped doing those reports and are working on a revamped, shorter summer that we're going to put out on the web. To replace what used to go on those reports, we've made these tables that you may have seen if you're familiar with our reports. It shows the resistance by agent. And so, this shows just an aggregate view in a nice summary table of individual drug resistance by year. And again, this can be based on phenotype or genotype. Resistance by preset MDR patterns. And again, by pattern and by year. And then all these figures are also exportable and in case you want to look at your own analyses. Then we also have the MIC distribution. And so, this is actually one year at a time. But it's drug by MIC.

So, I'm going to go back to the power point and talk about some of our future improvements that we're working on. So, one of the things you may have noticed is we don't have an interactive display summary of our resistance determinants. Those data are located in the download. But we want to build a chart, and we're working on that now, this is a demo of one we have in dev, a summary of the different determinants that account for the predicted resistance we're seeing.

In this case, the *Salmonella* ampicillin, these are the different genes that we're observing in our sequence population. We're also working on building alternatively views for the reports we just looked at that are more familiar to clinicians who are used to looking at antibiograms that focus on clinically relevant antibiotics, more recent years only, and percentage susceptible. So, antibiotics that clinicians are often looking at are percentage susceptible as opposed to what we typically report as the percentage resistant that they're looking for options to treat. We're hoping our data can help inform what is going on nationally in terms of clinically relevant resistance for infections that may need treatment. We're also looking at increasing our filtering capabilities in our search options. We do have data that we can slice up more at an aggregate level by sex, age group, specimen source, even strain type. Maybe later today and definitely at the next couple of days of the public meeting, you'll hear about REP or recurring and emerging and persisting strains. That's another way we can examine our data.

In the last few minutes, I want to briefly touch on another dashboard that our division here at CDC has created. And that was -- is called the BEAM dashboard that was released earlier this year. So, BEAM stands for bacteria, enterics, amoeba, and mycotics. We talked a little bit earlier about SEDRIC and how PulseNet data currently feed into SEDRIC. NARMS data as well goes into SEDRIC which is used by our Federal and state partners when investigating outbreaks. The current and first version of this BEAM dashboard focuses on *Salmonella* bacteria from human specimens that have been uploaded to Pulse Net. Future iterations of the BEAM dashboard will include additional pathogens, microbial resistance data, and epidemiological data from outbreak investigations as well as data from FDA to include nonhuman or animal or food isolates. This dashboard is currently updated quarterly in March, June, September, and December. There was recently an update this month. Future iterations will increase update frequency to near real time. I'm going to go ahead and pull over the beam dashboard. And so, this can really be thought of as the first public display of Pulse Net data. And so, for us, for NARMS, this does not have any resistance information. But we hope to add high level resistance information into this dashboard soon. And then with link outs for those who really want to drill down into the NARMS Now human data or NARMS Now integrated data. There's just a number of displays these isolates -- these data go back to 2017. So actually, during the earlier years, this does include data that has -- was PFGE'd because these are all PulseNet isolates right now. But in the later years, especially since 2019, these have been whole genome sequenced. There are nice summaries in terms of the specimen source, the number of outbreaks that were detected by PulseNet and investigated by our Outbreak Response Prevention Branch. There is a graph that shows number of isolates by month. AS well as what is of interest to many partners, which is the overall serotype distribution we're seeing. Then there's an outbreak detected by states figure.

There's a second page to this dashboard that shows the quarterly report. This really focuses in on the previous year. And it shows the number -- the total number of isolates in the most recent quarter and compares that to past averages. Again, there are serotype distributions, different ways of looking at that, different ways of comparing it to the past. Okay. So that's all I had.

If you had questions about the BEAM dashboard, there is an FAQ document you can Google, you can either use this link or Google "BEAM FAQ." We have a group who built the BEAM dashboard outside of NARMS, it's called the SIMSO group. If you have further questions for the NARMS Now human data dashboard, feel free to contact me at uvz6@cdc.gov.

That’s great Jared. We have two questions. The first one is just Is CDC NARMS Now linked from the CDC homepage.

It is. We actually update our data every night. Depending on if it's in the morning or at night, it's going to be reflective of either that day or the previous day's latest data.

Second question from me, so we're working on that data- sharing relationship with the different states. Will that be retroactive, will we get the historical data, or only data going forward?

We've been treating that retrospectively; so once we get the agreement we will just flip the switch on all the data.

As it should be. Thank you, Jared.

Resistome Tracker –Presenter Dr. Heather Tate
Time- 4:59:53 – 5:31:42

Sorry. I lost my Zoom screen. All right. So, I'm going to talk about resistome tracker which is our global resistome tool. It was created a while back actually. We released resistome tracker when we released the first iteration of NARMS Now back in 2017. It originally showcased just the *Salmonella* data but since then we’ve made some updates. In fall of 2020 we added more bacteria to resistome tracker and then last year we added other gene classes so we're not just tracking resistomes, we’re also tracking virilomes and stress genes.

So, this is a little bit different from the NARMS now tools that you just learned about from Amy and Jared because it does not incorporate susceptibility testing data it's built solely on the backs of WGS. So, resistome tracker is based on the premise that there's this high concordance between the presence or absence of acquired resistance genes and the corresponding phenotypes. The papers that are shown on the slide are the NARMS papers. NARMS was one of the first groups to demonstrate this in the four enteric bacteria that we studied: *Salmonella*, *E. coli*, *Campylobacter* and *Enterococcus*. Resistome tracker is a publicly accessible epidemiological tool that allows users to track the appearance and spread of resistance genes in enteric bacteria from different sources around the world. It's really meant to complement NCBI's pathogen detection, pathogen detection is a lot of data and several folks have shown you the pathogen detection isolates browser. What this does is really fits on top of that. Some data dashboards that can help you visually track and analyze the data and identify trends or patterns and then present those findings to users.

So resistome tracker can help you identify potential reservoirs for the dissemination of resistant bacteria. It can also help you speed the investigation of resistant outbreaks if you know what those reservoirs are and you might be able to generate hypothesis for research.

All right. So, the data that go into resistome tracker are pulled from NCBI's curated detection system which as they describe on the website is a centralized system that integrates sequence data for bacterial pathogens obtained from ongoing surveillance and research efforts whose sources include food, the environment, and human patients. So, that make this tool inherently one health because it is covering all of those three aspects of one health.

A number of U.S. state and federal public health agencies contribute to the pathogen detection database. In fact, about 65% of the sequence data that we pull into the *Salmonella* portion of resistome tracker from the U.S., a lot of the international isolates are coming from public health England and then the remaining international isolates are contributed through international genome tracker partners also DTU and Senasica and also some research labs.

So, this is our current process. It's changed since we first created resistome tracker. Amy downloads the analysis results from pathogen detection on a weekly basis. She actually downloads the full pathogen detection dataset every week, not just new isolates that come up in that week time period. That’s so that we can capture any changes that have been made like to the isolate name or the to the strain ID or computed stereotype.

And then she matches the genes to phenotype using the reference gene catalog which NARMS helped contribute. Currently we run this file through a program called lex mapper that cleans and normalizes the isolate source text, which is free text and really messy. So, after it cleans that text up, it matches the clean words to ontology terms and the ontology terms group the sources into broader categories so that when we compare resistance genes across categories, now we are more likely to see patterns arise from the data. Then we create all of our data dashboards in tableau which is the business analytic software that Amy talked about earlier.

Okay. And so hopefully you can see this because the fields are tiny, I know. These are the fields that we pull into resistome tracker, organism group that is especially important now that we show more than just *Salmonella* genomes in resistance tracker. The collection date, strain ID, bioproject, AMR genotypes. Again, now since last summer we've incorporated the stress and virulence genotypes, the computed serotypes as well as the isolate ID, the creation date which we call in resistome tracker the NCBI release date. Also the location, the geographical location that this sample originated from. The isolation source, the isolation type, SNP cluster and biosample. All are all included in resistome tracker. So now I'm going to exit from the PowerPoint and hopefully now you can see the resistome tracker page. And so I'm going to walk you through the resistome tracker. Now, I have it appearing right now, I believe, I'm in Firefox. But you can view it in chrome or Microsoft Edge or Firefox all the same. We've had problems with certain browsers in the past, but all appear to display properly now. And this is the main page of resistome tracker. There are five main activities that you can do from this page. You can customize your resistome tracker results based on predicted resistance to specific drug classes. You can compare the distribution of resistance genes across source groups. You can discover genes that are new to the database and have been published in the last six months. And then also, if you have a specific gene you're interested in learning more about, you can explore where, when, and in what source that gene was first identified in the database and then you can track the spread from there. And the fifth thing you can do on this page is you can look at these preset alerts which flag the appearance of specific resistance genes uploaded in the last 60 days and identifies resistance genes that are reemerging or appearing the first time in a year or more. The numbers here in these alerts change every week.

The time stamp here will always tell you when the data were last updated. And again, Amy updates this every Monday. And the time stamp now coincides with the content current as of detail appear. For a very long time actually, it did not coincide with that. So, I know people were confused. They'd come to this page, and even though it was being updated frequently the content current of, had a date that was a year prior. So, people didn't think these data were being updated but they are. So when you start resistome tracker, it's best if you start with an idea of which organism and/or genotype or predicted phenotype type you want to investigate. That's because the first thing that you need to do is you need to select the genus. If you hit the carrot you see the drop-down menu of all the 4 genera that are included in the resistome tracker. These other drop-down menus will update depending on what you pick for your genus. So, for instance, if we select Nontyphoidal *Salmonella*, if you go over here to serovar/species, you'll see all of these are serovars but if we change the genus to *Enterococcus*. Sorry, I had to update, change it to *Enterococcus*. Then you'll see that all of these are *Enterococcus* species. The gene type indicates whether you want to use resistome tracker to look at AMR genes, the stress genes or virulence genes and then the gene sub type field is really most useful if you want to look at the stress genes. Because then you'd be able to see the different types of stress genes and you'd be able to filter on those whether they're biocides, heat, metal and actually if we were to have selected *Salmonella*, you would also see acid as a selection here as well.

So, I'm going to go back to non-typhoidal *Salmonella* and the first thing I'm going to show you on this page is the alerts. I’m actually going to go back to AMR. Notice alerts didn’t update because we don’t have alerts for stress genes. So, if I go back to AMR, you’ll see that we do have some alerts. And I click on that. Okay. So, alerts are mostly for the AMR genes right now, but we also do have some alerts for shiga toxin genes and *E. coli*. And the really nice thing about going to alerts which is what I would do if I were a layperson in the field of antimicrobial resistance because really this page spells it out for you. It tells you what is most important for us to be looking at. We're most interested in transmissible colistin resistance genes and macrolide resistance genes, transmissible fluoroquinolone resistance genes et cetera, et cetera. And we're actually, in the process of updating this list so that it reflects updated information on new and emerging genes in NARMS that I believe Jason might present to you a little bit later when he talks again. Ideally what we would like to do is have users create their own alert but that’s not possible with tableau. We are also looking at other applications like R shiny and other software programs where that might be feasible. So anyway, this top portion of the alerts page shows you the alert-able genes and the number of biosample records that contain them. You can also see if this is the first time that these genes are appearing in the U.S. That's based on the color. So green means no. If there were a gene that was appearing for the first time in the U.S. it would appear as orange and I can tell you that we almost never, if ever see the color orange. If you scroll down, you will see a map where all the genes of interest have been located in the past 60 days and then further down, you will see the genes that haven't been seen or sorry, the first time the genes that we're seeing for the first time in over a year. Also, you'll see the total number of biosamples in the database that have this gene. This will really give you an idea of what type of gene might be reemerging. So, all the data on that page are isolates that have been uploaded to NCBI within the last 60 days. I do want to mention also, even though I selected this page based on all nontyphoidal *Salmonella*, this can be filtered by serotype by clicking on the serotype drop down or species if I had selected another genus that has a species.

All right. So now, let's go onto compare. This compare page is really useful if you want to look at the relative distribution of resistance genes among sources and if you want to see whether certain sources are more likely to have certain resistance genes. So here I want to look at quinolone resistance genes and you'll see the distribution of genes in each source category. The source categories again are dictated using the lex mapper ontology. I just want to make sure that all the other fields are set to where I want them. When you're changing a field it's very important that you hit, well, where it's applicable it's very important that you hit apply because, if you close this out, your changes will not apply and sometimes it's hard to realize that. Okay. So now, looking at the source categories, I would prefer, when I look at this, I look at source categories that have similar number of biosample records that have the particular gene class that I'm interested in. So like here, I would look at cow and pork. These numbers down here really provide the sort of contextual information because it allows you to see, okay, how many isolates do we have that have a quinolone resistance gene? Something where you only have one isolate that has a quinolone resistance gene and so it's appearing as 100% of the isolates have a qnrB19. That might not be as reliable as a source category that has more isolates or more biosamples. So anyway, looking at a cow a pork and let’s say turkey, if we look at cow, we can see that the majority of quinolone resistance is conferred by mutations in the gyrA gene. Very few of the isolates have the qnrB19 gene but when we go to pork we see there are quite a few isolates that have qnrB19 and now it’s reversed, very few have mutations in the gyrA gene. When we go to turkey, the majority of these isolates have the mutation in the gyrA gene. So, then what that tells me is that we're more likely to see transmissible fluoroquinolone resistance in pigs than we are in cattle or turkey. One thing to note is, these lex mapper assignments can be a little faulty. If you go up here and click on the question mark, it will tell you how the sources are being mapped. And so, like if we click on beef, you'll see all of the sources that make up that beef category. Same thing with cow, you'll see all the sources that make up the cow category. Some of this again can be a little bit fault so if you see something that doesn't make sense in the way it's categorized, feel free to let us know. This lex mapper software can improve if we provide corrections and continue to update it so it would be good to have your feedback on that sort of thing.

So now I'm going to go back to the main page and we're going to go to customize. Customize is a good option if, actually customize and explorer are good options if you have specific resistance genes in mind. For customize, I want to use the example of ST198 *Salmonella* Kentucky. So ST198 *Salmonella* Kentucky, this is a group of Kentucky it's shown in the purple right here. They contain triple point mutations in the quinolone resistance determining region of the gyrA and parC genes along with other resistances including cephalosporinases and sometimes carbapenemases. This particular group of *Salmonella* Kentucky is a global issue, originating in Africa but these Kentucky have also been seen in Europe, the Middle East, and most recently in human clinical isolates in the U.S. And there are current investigations going on now to discover the potential sources of these infections but we might be able to use resistome tracker to find some clues. Let me go back to resistome tracker. I'm interested in *Salmonella* Kentucky so I’m going to go over to the serovar tab and type in Kentucky and apply. Notice how everything updates. So now these alerts would be only for *Salmonella* Kentucky and when I click on customize, this would only reflect *Salmonella* Kentucky. Now because I'm interested in multidrug resistant ST198 that has the mutations in the quinolone resistance determining region of gyrA and parC I'm going to select quinolone resistance and then on this page I see a histogram. The histogram will show me how many records are in in each of these boxes, these colored boxes indicate the geographic area that the source originated from. So, these sources in this box will all be from the U.S. Sources in the orange box are not from the U.S. and then gray we don't know what country the source originated from. So, I'm going to click on gene because I'm interested in the isolates that have the triple mutation. So I'm going to click on isolates that have gyrA mutations and parC mutations, or parse mutations. Because unfortunately, this is not additive when we click on them. So we apply and then, let's see. I'm just going to click on again I want to see what are some potential sources for the human infections in the U.S. so that's why I clicked on just USA. And I see that now we have some things that appeared under this blue bar. We have a bar chart that shows us the sample distribution by source category. We also have a spreadsheet that gives us quite a bit of information. It shows us all of the genes of the ones I selected which ones are appearing in these isolates, where they came from. So again, I selected USA but now we can get more information on state if that's available. T

The source category, the source if that's available and then serovar, year. If we hover over these dots, we'll get the full genotype for that biosample. We'll also get some additional isolate identification information and we’ll get the, rather than the year we will get the month that this information was released by NCBI. If we click on the dot, we can actually be taken to other pages at NCBI, which will show us more about the biosample. We can learn more about that particular gene and we can also get links to the NCBI pathogen detection isolate browser where then we’d be able to actually go look at the SNP tree. Trying to answer my question, I'm seeing a lot of clinical and human isolates which we knew there is increasing numbers of cases of humans that have the ST198 which is indicated by the presence of these three mutations. But then we also see these other sources which I had looked this up earlier. These are actually owls and another bird species. What this tells me is that maybe, you know, if I am interviewing a case that has a ST198 *Salmonella* Kentucky infection, I might want to include, have you had contact with owls in my questionnaire. So that's just one, you know, one tool that we can use to get a potential source for an outbreak. These bar charts are also clickable, too. If you click on them, they'll give you some additional information about the gene. But really a lot of the information is contained in the spreadsheet here. And the spreadsheet can be filtered by source and also geographic region. I’m going to go back up to the home and now I’ll show you explore which is the final part of this.

For explorer, I really want to use the mphA positive *Salmonella* Newport as an example. So let me see. All right. So CDC recently released guidance telling travelers to watch what they eat if traveling to Mexico because recent travelers there have been infected with multidrug resistant Newport harboring the mphA gene which confers resistance to Azithromycin. So maybe after reading this what I might want to know where else have we seen an mphA containing *Salmonella*

Newport and what has been its geographic pattern over time? So I'm going to go back to resistome tracker and this time I'm going to select Newport. And go to explorer. And nothing's appearing right now because the default is this aminoglycoside resistance gene which because it has parenthesis around it that tells you it's apparently not in *Salmonella* Newport. I can either scroll down to or I can type in mphA. And when I do that, I see a number of things.

First are some KPIs, key performance indicators that show me the most relevant information. We can see that the isolate was first seen in 2002. It's been seen in 24 locations in 7 countries from 13 different sources. Then down here we have a map, and we also have a year by year play with some color indicators that tell us what the food source is that these *Salmonella* Newport genomes were found in. And if you see a star that's because that's first time we found it in that particular food source. If you see a circle, as you see here, it's the second time or at least the second time that we saw it in a particular food source. So, as you can see, we first saw it in 2002 in the U.S. in a veterinary or clinical research isolate. And then again in 2003 in that same type of source. And again in 2006 and 2007. And then in 2011, let me zoom out, we saw it for the first time in a fish source from Vietnam. And then in 2014 from an unknown source in Canada. And then in 2015 from a human isolate in the UK. So, you can keep scrolling through, and you'll see, you know, where this mphA gene has been found in *Salmonella* Newport over the years.

On the right side, you'll see a chart that doesn't change but it's the first appearance of the source category. I'm sorry, first appearance of the gene by the course category. And then below is a country collection timeline which not only shows you the first appearance by the source category but also shows you that information by country. So, like in the U.S. again it appeared for the first time in 2002. It appeared in beef for the first time in 2016, in humans for the first time in 2017. In cow in 2018 and you can see what has happened since. In Mexico there was an isolate, an environmental water isolate of *Salmonella* Newport with MPHA that was found in 2019. And cow in 2020. And first in beef in 2021. So, I mean it doesn't, you know, this is not a SNP tree. There is nothing about the relatedness of these isolates in resistome tracker but it does give you some clues as to points to where you might want to point your questionnaires. Looking at this I'm really interested in the fact that it has also been found in fish and in chicken. And so, I would be interested in seeing, you know, how related these isolates are to the beef in the clinical isolates. Is their potential connection with some imported fish or chicken product or feed coming from those products? And if there's some sort of epidemiologic link there. So that is essentially resistome tracker in a nutshell I hope I've given you some ideas about how you can use resistome tracker for your work. Let me just return to the PowerPoint so you have the QR code. So that you have the QR code. And if you have any questions or any feedback about what's been presented to you today, please feel free to email me or Amy or Errol or send an email to NARMS@FDA.HHS.gov. I should note that the QR code, you can use this to bring it up on your phone but ideally resistome tracker should be viewed on a desktop because it has not been modified for use on cellphone. SO maybe after you get to this webpage you can send it to yourself in an email so you can view it on a desktop. All right, that's it.

Great, Heather. Wonderful talk. We have one question in the CHAT. So where does resistome tracker show the timeframe for the results? And can users change the timeframe? I think this may be around the customer data exploration aspects of resistome tracker?

Yeah. So um, right. So, you can do some, look at timeframes in the customization, actually in customization in compare and explore. You can look at different timeframes. For instance, with customization, you'll see the different timeframes here of all the data that have been collected. So, you can select it that way. If you were to do some comparisons, you can also filter by different years. So I had it on 2018 when I was giving the presentation but you can select other years as well.

Great. Thank you, Heather. One question or comment from me. I know you mentioned sometimes the lex mapper and ontology mappings may not always work perfectly. But just, I'm not sure people realize but the isolation source at NCBI is free text. And so even with, I think maybe some errors occasionally introduced by lex mapper this is a massive upgrade if you want to look at the data which is part of the reason it exists.

Exactly. Exactly. The free text was very messy. So the whole lex mapper project was a huge undertaking. And I believe that the lex mapper outputs are now included in pathogen detection isolates browser as the IFSAC category, under the IFSAC category field. And Michael Feldgarden can correct me on that one what he talks. But yeah, much, a big improvement.

Thank you so much. Okay, we will take a short break here on it until 2:45 when we'll hear from Mike Feldgarden about AMRfinder and NCBI. So, I'll see everyone back in just a few minutes

-BREAK-

I have 2:45 here today. It's a real pleasure to introduce Mike Feldgarden. You've heard a lot about NCBI. You've heard a lot about pathogen detection, AMRfinder, AMRfinderplus. Mike has been one of the key collaborators for NARMS helping us move WGS forward as it relates to AMR and the public release and sharing of this data. Mike the floor is yours.

Accessing AMR Data at NCBI –Presenter Dr. Michael Feldgarden
Time- 5:43:53 – 6:14:05

Today what I want to focus on is how to get to a lot of our AMR data in NCBI and how it's generated. At this point I have this slide, I have this introductory slide about what NCBI is. The National Center for Biotechnology Information but given so many talks, have talked about the various things NCBI does I'll just basically skip this and say a lot of the tools you might be familiar with, Pubmed, the sequence read archive. They all have come from NCBI at the National Library of Medicine, before I get into the talk, the other thing to note down at the bottom, I have an email address (pd-help@ncbi.nlm.nih.gov). If you have questions, or if you want a copy of the slides, a lot of the demo stuff I'll be showing today, there are links in the slide so if you want to try to figure out some of the links yourself you can get a copy of the talk from us.

So, this, the AMR work we've done has really stemmed in part from the NCBI pathogen detection pipeline. At this point I'm repeating what other speakers have said about our pipeline but essentially we bring in a variety of data from various data generators, U.S. government agencies, international agencies such as public health England, basic researchers as well as clinical sources. We have a couple collaborations with the U.S. hospitals. And the data come in and get the metadata, for instance, in biosamples or what epidemiologists call the data. We also have sequence data in the SRA and GenBank. And that's all run through the pathogen detection pipeline where we do clustering tree construction to assist in epidemiological uses but what I want to talk about today is AMR finder plus. And this is just to give you a slight overview of sort of how the pipeline works. Like I said data come in from genomic data come from two sources either the sequence read archive or GenBank if they're deposit assemblies, if they’re reads, we assemble them then they're clustered and trees are built and people can do various epidemiological analyses determining what is related to what. But what I want to focus again on today is once you have these assemblies you can also annotate them and then run them through a tool called AMR finder plus, that identifies AMR genes, virulence genes and stress response genes.

And so, the basic question I'll be focusing on today is what is the antimicrobial resistance stress response and virulence gene repertoire of isolates and that's what AMR finder plus is designed to do. First, I want to briefly describe how, why we built AMR finder plus sort of what the goal of AMR finder plus is. Then I want to give a very brief description for people who want to use it on the command line it can be run as a standalone software. We strongly encourage data submission to public repositories so other people can look at it as you’ve seen in multiple talks today. But we realize some people might not be able to submit data or do so right away. I'll give a very brief introduction to NCBI AMR web resources and multiple speakers have described these so I'll just try to highlight some of the key points focusing on the isolates browser as well as a newer tool we have called

MicroBigg-E and how to use both tools together. And finally, I want to end with sort of a coming attraction. It does exist but how to access the data in the Cloud using Big Query. For people who are really starting to ramp up to large scale searches, we now have the ability for you to do that.

So just to start in on AMR finder plus. The saying “the beginning of wisdom is the ability to call things by their right names” was really the motivating reason for us developing AMR finder plus. That’s because when you have a lot of genomes, large scale requires concise information. We have hundreds of genomes a day. And we need a concise discreet signifier that conveys appropriate information about genotype and hopefully phenotype. And one of the key things we want to do in all of this is incorporate the ambiguity. We want accurate specification, but we also want to incorporate ambiguity and just to give you an example of what I mean if we think about the KPC family carbapenemases, so this a group of beta lactamases that confer resistance to just about all betalactam antibiotics, so just about everything that ends with -cilin, starts with -ceph or ends with -penam. There are multiple alleles in this family that is unique protein sequences. So KPC-2 is sort of the granddaddy of them all, it’s a carbapenemase but there's other KPC alleles we've seen such as KPC33. This is actually an inhibitor-resistant cephalosporinase. It’s one amino acid change from KPC-2 and in fact one nucleotide change from KPC-2.

So, while it loses the carbapenemase function it’s now inhibitor resistant. Often when patients have a carbapenem resistant isolate we use ceftazidime avibactam a cephalosporinase inhibitor combination, and that would fail. Likewise, with KPC-8, it's sort of a Reese's peanut buttercup of horribleness in that its inhibitor resistant and it’s a carbapenemase. It's a small number of changes from KPC-2. The reason I bring this up is by way of background. If we discover a novel KPC enzyme we want to convey the users when we lack information. We don't want to say the closest hit is KPC-2. We just want you to have a very clear indication that this is a KPC but we don’t really know what it is. We haven’t seen it before.

To describe how to do this, in AMR finder plus I think what makes the tool a bit unique, the special sauce is really it has a hierarchal structure. If we start from the top and work our way down if we have a protein and I should say we work in the protein space. The AMR finder plus tool can take in nucleotide data and just translate it over to protein. If there's something that’s 100% identical. We blast it. It's KPC-2. We know what this is, there are dozens of papers published on KPC-2 we have a crystal structure and so on. But if there's something that’s a little more distant let's say it’s 98% identical we say it's part of the KPC family. It's probably resistant to carbapenems although as I showed on the previous slide it might not be but we at least have a pretty good idea of what it is and what it does. If it's more distant, let's say 75% identical, we might only be able to say, it's part of this larger family of class A beta lactamases, all of which can degrade carbapenems or at least most of the proteins in those families can. But it's probably a carbapenemase but we really don’t know exactly what it does. Even if it’s more distant we might only be able to say it's a class A beta lactamase and another key point is that we can also say it's not a beta lactamase at all.

And this hierarchy is really critical because it allows us again not to over-specify the statements we're making. The other thing I should point out is, um, AMR finder plus we report blast output. We have a set of reference sequences. When we identify a resistance gene, we do blast and report the blast statistics. But what we do to sort of do this binning it's really hidden markov models. And these are multiple sequence alignments that are subjected to hidden markov algorithms. And then that allows us to identify certain features in each protein. The alignments are manually curated. The cut off scores are manually curated by biologists. And then we are able to then apply that in AMR finder plus across what now over a million genomes and be able to bin various resistance genes into this hierarchy.

Just to briefly talk how to use AMR finder plus locally, the optimal use is with the nucleotide sequence, a protein sequence and what’s known as a .gff file which describes the location of sequences. We’d also point out that PGAP, the reference gene catalog, the AMR finder plus database curation is lined up to NCBI’s PGAP annotation tool. Things like protein will typically be called the correct length is you use PGAP, but other tools can work pretty well, too. AMR finder plus can detect species specific point mutations and genes. It can also optionally, if you choose to detect virulent genes and stress response genes. It’s also relatively easy to install using Bioconda. If you’re starting to get into bioinformatics, I think Bioconda is actually a good tool to familiarize yourself with. It’s easy to download you don't have to download eight other programs that’s pretty much taken care of for you with Bioconda. Although again, I should point out we really do like to see data when possible, made publicly available so the entire community can look at it. I should just point out that all this information is subscribed on our GitHub site. If you Google AMR finder plus and throw in GitHub in there perhaps, you'll find it.

This scary looking slide, at least for people who don't typically use command line programming. This is really just to show you in fact, believe it or not, actually simple to use. If we start in the third row below example, these are just flags. The first thing is just saying here's where the nucleotide file is and give me the output in a text delimited table format in this location. And you can do more things, more complex queries shown on the bottom, you add your genome sequence a set of annotation, annotated proteins. The .gff file, describing gene location where to put the output. You can also put an organism flag so you say this is an *E. coli*, look for *E. coli* features. As well as another flag which we call scope and that allows detection of virulence and stress response genes.

And so, I'm going to show you firsts the screen shot and then we'll move to the browser. This is a NARMS isolate actually, a *Salmonella* isolate and just run AMR finder plus on it. Like I said you get tab delimited text file. And so, you can pop it open in Excel. And I've done it here so you can see it more easily. Some of the data that are included, each line is a gene or a point mutation that's been identified. We give you the protein accession so you can look this up in GenBank. And we give you the contig ID moving left to right here, the location on the genome, the most key thing here is the gene symbol shown here. There's various genes. We also give you things like the sequence name as well as that scope I talked about is this core, this, you know, perhaps an optional gene you might not be so interested in. And again, this is just sort of sliding over. We also have fields like class and subclass which give you more information. For instance, on the third row, the general classes, aminoglycosides but it's actually a gentamicin resistance gene. And then we also have this method column that describes how it was identified. Is there an exact hit to something in the database? Is it close hit or only identified by hidden Markov model? There is other information in this column I'll show you as well, there are some other columns I haven't shown you just the result of the blast output, the coverage of the reference gene, the percent identity to the reference gene and the accession for the HMM.

What I want to move on to now are the data interfaces. And again, because people have talked about this today that makes my job a lot simpler. Instead, I'll just try to hit some of the highlights. On the left these we have three browsers that allow people to access the databases we use, the first on the top is the reference gene catalog. That's where every row is a reference gene or point mutation. You can download those reference sequences, it's searchable it's browsable. There's, it uses many of the same tools and techniques that I'll describe when I talk about some of our other tools. Likewise, there's a reference gene hierarchy. Here this is a graphical interactive version of the hierarchy or of our entire gene hierarchy like I showed you with KPC-2, every row is a node in our hierarchy. That just allows you to see sort of OK how these things are related. Finally, we have the reference HMM catalog where every row is an HMM. This ultimately has links out to our protein family database where you can learn how was the, what genes went into the HMM cut offs, you can download the HMM for use in programs like hammer 3, et cetera. But what I really want to focus on are the two databases on the right. The first is the pathogen detection isolate browser. You can think of this as a table where every row is an assembled isolate. It may have AMR, virulence and stress response genes as well as an antibiotic susceptibility testing data. You can not only download this metadata but you can also download genomes at this site and protein files and the .gff file. The bottom right is what we call MicroBIGG-E. This is relatively new it's the microbial browser for identification of genetic and genomic elements. This table gives you a different perspective. Every row is an AMR, virulence or stress response gene with method and describes the method that was used to identify it, additional information such as class and subclass as well as the supporting evidence used to identify that gene. And here you can download genes, you can download flanking regions of genes and can also download the contigs.

So, let's move into the isolates browser. I had to change the slide this morning because we used to have 1.1 million genomes as of yesterday and we ticked over to 1.2 genomes. This is probably familiar to people who have been at this workshop. But again, the isolates browser gives us a whole bunch of information, not just AMR information, data collection, where you got it from, details about the isolation source, isolate identifiers and so on including things how it was sequenced and so on. But the columns I want to focus on are these three here today, third, fourth and fifth and just to sort of blow this up. AMR genotypes, there are two different AMR genotype columns. One is sort of a broader genotpes. We have some things that are genes that we don’t really consider core like chromosomal or ampC that don't confer any phenotype or efflux pumps. Our collaborators thought they would be interesting. We sort of have a broader net under AMR genotypes and then we have the genes where it’s been pretty well established that there should be some kind of phenotypic effect under AMR genotypes core. I apologize for that being cut off. Then there's also this AST phenotypes column. I just want to use this to point out that we do accept the antibiotic susceptibility testing data. NARMS has contributed a lot of this data and it’s actually quite useful for people trying to figure out how to improve prediction and how to find new mechanisms.

So now I want to attempt to toggle between my slides and the isolates browser website. I'm going to cover a bunch of topics. Some of these I'll just go through quickly. Like I said you can get the slides and the links and there are hot links in the slide so you can click out. Just to start, how to get to our pathogen detection sources. If you Google pathogen detection and perhaps add NCBI after that, you will get to our landing page. We are at the top of Google. We didn't arrange that. It seems to work this way. Two key things. We have our data resources so if you want to go to the isolates browser, MicroBIGG-E or any of these other things you can just click out here. Two things I do want to point out we have help documentation. This is very thorough and will probably contain your answer. No one in their right mind will sit and read the entire help document at once. You can go in there and usually find your answer, find examples of how to do something. We have another link which is how to, and this gives you some basic links of I want to be able to do this. And we tried to mockup some scenarios and there are PowerPoint that you can just go through, and say OK I have to click this button, I have to type this and so on.

So, moving to the isolates browser, this is just, there are again, I imagine most people are very familiar with this. At least some people are familiar with this. This is just searching for that *Salmonella* isolate. In fact, this is an easy thing to do. You can go in and add in a whole bunch of in this case I'm using biosample accessions but there are other terms and we'll just return everything you have. So, if I just hit return, now I've got two isolates here, these two *Salmonella*. Some of the features, I'll just go through some of the basic features so you can sort of kick all the tires and push all the buttons. If you want to expand everything so you can see the list of phenotypes or list of genes, you can hit that and collapse it. To download the data, again, the isolates browser you can download all of these metadata just through the table and it comes out as tab delimited. You have the ability to download various genome related files. The nucleotide sequence, proteins, .gff and so on. The other thing I would want to point out is default setting doesn't display everything. To answer Heather's question, we do include IFSAC category. I'll move this over to here now for display purposes and these two isolates unfortunately don't have an IFSAC category but that's where it would be displayed. Just so this is again just if you knew an isolate ID, you can just pop it in there and it’s a standard search.

To move onto other things. This is the same search that I showed you. These two isolates. But suppose I want to go in and this is sort of a toy example that I’ve already queued up. I can go to the filters bar, open this up and there's a whole bunch of fields I can go in and use. I don't have to sit there and type a long complex string. In this case I've asked, OK I want to go to AMR genotypes core and find the isolates, in this case out of two that have the TEM-1 gene. There is one isolate and if you notice there is this little blue one and that actually is the count of isolates that have that feature. And again, if we go here, if scroll down, you now see one isolate and it has TEM-1. These filters are very useful. The other thing is if you can figure out how to do it in the filter it will also give you the syntax so then you could just go cut-and-paste that syntax for future use.

This is just one more basic thing. And this is to remind me if you mouse over each of these table, column headings, you can see the syntax used. So here I've just done AMR genotypes. And it’s basically saying can I find this string, blaKPC-2, in this column. And in fact, you can. And there are a whole bunch of isolates that have this. One other thing just to point out so we can see this. If we, well no, I'll skip that in time we'll get to that. One very sort thing here, again we're still in the isolates browser. You can use standard boolean terms. So, if you just have two things you want to string together it's just “AND” and there’s also this sort of wild card. If you say well, I want to look for all KPC genes, not just KPC-2. You can go here and use that. And that will return any KPC gene not just KPC-2, and in fact now we see there are over 25,000 of these isolates.

Finally, just one thing to end with the isolates browser because I do to mention antibiotic susceptibility testing. Again, here we can look at phenotypes. It's this command “AST_phenotypes” and here we put a wild card followed by “penem=R”. that just is telling us look for any antibiotic in this case anything that ends with penem that happens to be resistant. And that will give you a list of isolates. The filters, if we look at filters and go to AST phenotypes we can type in penem, there we go. And you can actually see like I said with these counts you can see how many are resistant to each of these various carbapenem and carbapenem combinations and if you were to click them, you're limited to anything what happens to be if the penem resistant. This is one more thing again, I believe Mustafa mentioned this in his talk but there's also the capability, this is just a search I put together of looking for something that is *Salmonella*, is an environmental isolate. This is a field that's derived from some of the other data from biosample and has the CMY cephalosporinase and depressingly we see 6200 isolates have this. But what you can do is go in and save this as a search as long as you have an NCBI free account. And then it will be saved and there will be an email associated with that account and you’ll just get an email every time there's an update, every time something comes through and hits this particular queries.

So, I’ve covered, you know, the isolates browser as well as saved searches. I want to move onto microBIGG-E. Again, just to remind you this is the perspective from the genetic element vantage point. In this case, each row is either a gene or a point mutation just again to jump into the microBIGG-E. This here is a search for, let me go back to the sides for one second. I'll just use an example of searching for anything that has mcr-1. I've done this as a wild card, so it’ll be anything with the M string, mcr-1. The element symbol, you can mouse over and get the search terms. Just to walk you through this in the output I've added the HMM accession. So, this is a hotlink you can go look at the HMM. Likewise with protein you can get the protein link. We do pull in some things from the isolate browser but not everything so we'll give you the scientific name, you get the biosample, the isolate identifier as well as a few other fields that are available and choose columns such as the isolation type or epi type field that I showed earlier. Other things like collection data and so on. But not everything in the isolates browser is available. We give you the position information of where that element is. So contig start and stop. Importantly we give you the element symbol, so this is mcr-1.1. Because of this wildcard we're also pulling an mcr-10.1. We give you things like the element name. So, this is basically the AMR finder plus output. But just in a graphical user interface. Things like percent coverage, percent identity, the method used.

If we look at the filters again there are filters and it's the same kind of method. If we click method for instance, this gives you counts. And you can see that, not everything is identified using the same tool. Many of these are identified using the protein sequence and its exact match to a known sequence but there are things that are a little different. This BLASTP, there are also some partial proteins.

We distinguish between partials in the middle of a contig and partials at the end of the contig. The partials at the end of the contig, the line with 38, those may be functional genes that just have assembly problems but the partials in the middle of the contig those might not really be functional. And we can also identify things such as if the gene has an internal stop as well. As you see second from the bottom with five. And you can click any of these and to limit your search. The other thing that I think is notable about microBIGG-E, is that you can identify contigs that have multiple genes on them. Here I'm identifying a contig with TEM, that all the contigs that have TEM-1 beta lactamase and the KPC-4 beta lactamase. And again you can go through and filter, I should point out in terms of download you cannot only download the table but importantly you can download the elements. This would download all the genes on the contigs or you can limit it through the filters. You can download the contigs although this is limited to 1000, as well as you can download the genes of interest with flanking regions or proteins.

I'll skip this in interest the time and just move back to this example of I've identified all the genes on contig, all the contigs that have TEM-1 and KPC-4. What I want to do is look at them in the isolates browser. You don't want to sit there and have to paste all the biosamples and then do that. There's a very simple thing you can do, just hit cross browser selection, show isolates and now basically you have the typical isolates browser view. But this is only consisting of those isolates with KPC-4 and TEM-1 on the same contig. And you can go in and do the various downloads.

The other search that I think will be interesting, another sample search is I identified a large *Salmonella* cluster with over 16,000 isolates. Let's say I want to look for azithromycin resistance. If I did that in the isolates browser I’d have to know every potential azithromycin resistance gene and stick it in the search bar but with cross browser selection I can just go to microBIGG-E, use the filters, go to sub class in this case and start typing azithromycin and just click the box and now I'm limited to four. All of them actually have this resistance mutation which is actually showing up with typhoid Salmonella. If I then said I want to look at just these four with isolates browser you can ping pong back and forth and go to the isolates browser and it will display the four isolates.

So, that's a whirlwind tour of some of our GUIs. I just want to end with one thing before I get to the credits we've also put a lot of this in the Cloud. So, this is through a tool called BigQuery using the Google cloud platform. You do have to know a bit of SQL programming language. Currently we have the microBIGG-E in there, the AMR finder plus output, the contig sequence and protein sequence as well as the isolates browser metadata. The two key things are is it allows access to large datasets. There are people who want all the *Salmonella* microBIGG-E data. We're not going to let you download a table with 35 columns and 6.5 million rows but you can get it through a BigQuery or also something like 100, tens of thousands of contigs that we cap that, but you can actually do that through BigQuery. The other thing is you can join the isolates browser in microBIGG-E tables, the data, so you can do something like easily find isolates that are carbapenem resistant but lack any known carbapenem resistance mechanisms. That is something that can be easily done in SQL. I sort of gave an example in the top right. I just want to point out that we are doing a conference, we are hosting a workshop the day before the ASM NGS conference on this. And we have a lot more information about this at the website shown here.

And I just want to conclude that this is, I'm just one person on this project, pathogen detection project held by Bill Klimke. There are a bunch of people that worked on the AMR project as well as a whole bunch of collaborators that we’ve dealt with over the years. And I'll end by leaving up some NCBI resources. If you have questions, contact us at pd-help@ncbi.nlm.nih.gov. And if there's still time, I'll take questions. Otherwise, thank you for listening.

Thank you, Mike. We have one question in the chat. I think it's related to the AMR finder predictions. Are these predicted resistances from whole genome sequencing or based on MICs from AST. So, I think this may be in reference to some of the reference gene catalogs you showed.

Um, so just looking at the question. So, the, if you're talking about what the column that is the AST phenotype, those are not predicted. Those are user deposited through biosample. That is experimental data done with usually we have a whole bunch of fields that describe how it was done and what methods and so on. So that's how we're doing it. It looks like that was in fact the question.

Thank you so much, Mike. And so, if you will stop sharing, we'll let Jason start. Jason, you are back. We can see PowerPoint. We can see the presentation view.

Emerging Antimicrobial Resistance (EAR) Alerts –Presenter Dr. Jason Folster
Time- 6:14:34 – 6:38:53

Okay, so last talk of the day. I realize that it's been a pretty long day so I'll keep this short. But I did just want to mention something that we’ve been working on this last year. So, this was NARMS, all the different partner agencies including public health agency Canada. And that was you're trying to establish an emerging antimicrobial resistance alert system. So just a quick little background. I think we realize, you know we have all this great data and we’re certainly looking at that data and you've seen all the different ways that we make that data public. But I think internally, we are all sort of doing our own things when it came to looking for emerging resistance and there is tons of overlap and also some specific things that different people are looking at. So, we already had established the molecular epi working group that met once a month and they actually had a section there for discussing any sort of emerging or concerning resistance that we're seeing but we realized that wasn’t really being done on a routine basis. Sometimes there's a one-off search. Also there really wasn't any sort of harmonization. One group might be looking a little different than the other groups. We decided to go ahead and establish a new working group. That was the emerging antimicrobial resistance working group or EAR-WG.

This is just a list of the current EAR-WG membership. I think the takeaway here is just to show that we had really good buy in from all the different groups in NARMS including public health agency Canada, included people from the lab side and the epi side which I think is really important to this sort of work. The first thing EAR-WG did was to establish some different goals. The first was to identify what our current alert systems were in all the different agencies. How do we actually identify emerging concerning resistance? Whether that's internal alerts like for example the CDC radar system, database queries we're running in our internal databases or evens external alerts a lot of us still have. NCBI pathogen alerts set up concerning resistance genes. And the next question was really what do we alert on? Certainly, the easy ones, you know carbapenemases, ESBLs, mcr genes but also recognizing that we do have some differences in what organisms we're looking at. We have human specific pathogens at CDC and public health agency Canada and also just the idea that, you know, we may see something on the clinical side that really is not concerning to us but maybe concerning to USDA or to FDA and vice versa. So, recognizing that alerts may not be specific to the group that actual identifies them first.

So, then the second goal was to set out to try and harmonize our alerts wherever that was possible. So, what are we alerting on, you know, the specific gene, the mutation, what are the targets. Are we looking genotypically alone, are we including phenotype in that and where we're including phenotype are we just looking at resistance versus elevated MICs? For example, for carbapenemases. We explored whether or not we have any methods or workflows that we can share. We knew that for PulseNet/Bionumerics, that's used for clinical isolates coming into CDC but FSIS actually does a fantastic job of uploading their isolates as well into PulseNet. It allows us to see that data and alert on it as well. And then like I said, most of the groups already had some NCBI pathogen detection alert set up and sharing those and deciding if there was any sort of harmonization we can do there.

And then the last goal was to really establish this NARMS alert system. You know, we realized from the first two goals that, you know, the groups really had separate systems already in place and that everyone was sort of already looking for emerging resistance but there was no real good way of sharing those systems.

They were in place. We certainly didn't want to make more work for people. So, we just landed on a very low tech solution that it created a shared Excel template for alert submission. It's a single template that everyone was able to submit onto and that gets collected. And then it goes out for discussion. We established a monthly EAR-WG meeting. This happens the first week of every month. So in the first couple of days of the month, people look back at the previous month. They either run those alerts or if they have those alerts already in place or they're getting things from NCBI, they go ahead and put those onto the shared Excel file and then that just gets collated. And we have that monthly meeting where we sit down, we go through each of the alerts, discuss any sort of next steps on the epi side or on the lab side and the following week is when we have our molecular epi working group and it’s a standing update on the agenda from the EAR-WG group. Really that's to go ahead and get that same information out to NARMS leadership and to everyone who participates in that meeting which is a lot of the members of NARMS. Now, I'm going to try and switch to, can you see that file?

Yes, we can see the Excel file.

This is the Excel file. It's just trying to take screen shots of this just really didn't work. So hopefully people can see it. It's not really important seeing the details. This is an example, this is actually the one from July of this year. So, what you can see at the very top, we have the alert name, so the agency that's running it, what organism, what sort of tracking system whether it's being done in our database or NCBI. Phenotype, whether we're looking at phenotype or genotype. The alert period, so typically it's the month. So, in this case it's July. The number of isolates that are either clinical or nonclinical, species serotype, source if we know the source, genetic mechanisms if that’s important. We have a column for WGS IDs so people can go and look at these themselves and any sort of additional notes. This is color coded by agency. So, the first section here in orange is for CDC and then FSIS has two additional alerts because as I mentioned we're able to alert on those isolates through our system. And then FDA has some specific alerts. And then public health agency Canada. And then just going through them, I mean, a lot of these are overlap between all the different agencies, carbapenemases, MCR genes, pan resistance, acrB mutations which we know are important for macrolide resistance, tetracycline. We have some specific alerts at CDC and public health agency Canada that are human pathogens so typhi. We have a couple ones for typhi. And then I'll just mention we also have two sort of specialty alerts that aren't resistance genes. In this case we're actually alerting on a REP strain. And so, this is *Salmonella* Kentucky. So, this is considered the European strain. And I'll come back to that in a minute as a specific example. And then this one is actually looking for a plasmid, so this is the, what people call the pESI plasmid from *Salmonella* Infantis. But in this case, we're looking for that plasmid outside of Infantis so we see this plasmid all the time in Infantis and it wouldn't really do us a lot of good to alert on hundreds of these that we would see a month. So, instead we're really looking at the plasmid outside of that serotype.

Okay and then here is just the timeline of EAR-WG. So, you can see EAR-WG was formed we had our first meeting the beginning of February. That's where we established our purpose and our goals. In early March we tackled goals one and two. Early April goal 3 and then late April is where we finalized that alert process, created that template and EAR alerts went live May 1. So, we had our first sharing of the results with the molecular epi group in mid-May.

Okay, so I mentioned I'll come back to REP strains. I think there's going to be some other talk about REP strains at the public meeting in the next few days. I think in some of the working groups but just in case I just want to give a quick intro to REP strains. These are reoccurring, emerging or persistent set of bacteria related by whole genome sequencing that continue to cause illness over time. What escalates a strain to be a REP strain, a lot of times these are existing clusters or outbreaks. Sometimes they have concerning resistance patterns, sometimes they may have nonclinical isolates. It's really about genetic relatedness over time. It really allows us to characterize new sources of disease and try to develop novel prevention approaches.

So, as I mentioned there are three main categories. You have reoccurring strains. These that periodically cause illnesses. Typically, they're in outbreaks but they’re separated by periods where they don't. Here's a good example there. We have emerging strains these are ones that increase in frequency or have the potential to increase in frequency over time. And then we have some persistent strains. These are strains that cause illnesses consistently over time and may or may not rise to a level of outbreaks over time but they're always sort of there.

And here's a good example of where a strain can go into different categories over time. So, this is looking at *Salmonella* Reading associated with turkey products. So, this started as an emerging strain, you can see the epi curve here. At which point we saw increased numbers to trigger an outbreak investigation. And then after that outbreak investigation it really settled into a persisting strain of concern.

So why are these REP strains important? We know that most illnesses reported through PulseNet are not linked to a source. In fact, 90% of isolates don't have a cluster code and really only a fraction of those gets solved. To really drive down the incidents of enteric illnesses we need to better understand the seemingly sporadic illnesses, we think that REP strains probably represent a larger fraction of illnesses than traditional outbreaks.

So, I mentioned in the alerts that we do alert on this MDR *Salmonella* Kentucky strain REPJGP01. So, what do we know about *Salmonella* Kentucky? This serotype is commonly isolated from chicken in the U.S. but it's much less commonly isolated from humans, so in LEDS about 100 cases per year in the U.S. puts it way down to 54th most common serotype.

And so, you know, why do we care about *Salmonella* Kentucky? Well in the U.S. there's a strain of *Salmonella* that's called ST152 that's common in the U.S. It's in poultry but it seems to be much less likely to cause human infection. However, in Europe and other parts of the country there is a concerning strain of *Salmonella* that's MDR, it's called ST198 and that is circulating other in parts of the world. So, what we're really asking here is what can we do to try to prevent that from emerging in the U.S.?

What's known about ST198, it was first recognized in northern Africa in the mid-2000s. They think it emerged around 1989. It now circulates widely in Northern Africa, Europe and Southern Asia. It's commonly MDR there's a variety of resistance profiles. And just to note we do have a related but a distinct lineage of ST198 in the U.S. that's been recovered from dairy farms in the U.S. But so far it doesn't seem to cause as much problems with disease.

So, you know, why should we be following this strain? Why make this a REP strain? The concern here is really that this strain could emerge and expand in the U.S. Right now, cases tend to be due to international travel and those are really less concerning for us. It should be noted that the strain has been found in imported U.S. products or imported products to the U.S., for example, spices. What we're looking for is non-travel cases could signal the emergence of this strain in U.S. food production like poultry. So, we currently track the strain in EAR. We see about one new U.S. isolate in every one to two weeks. So far, they've all been clinical. Most have international travel but not all. Again, suggesting they may be imported products. As to date we haven't seen any animal or retail meat isolates with this REP strain.

And then moving onto that MDR Infantis plasmid example. A little bit of background on Infantis. This strain hit our radar in July of 2015. The FDA is the one who first notified us of a concerning resistance gene, in this case the CTX-M-65. That was isolated from a retail chicken meat in December of 2014. Way back in 2014 we were still looking at PFGE patterns. This started as a very rare PFGE pattern first appeared in 2012 in PulseNet and we really focused on that pattern at that time. But then that pattern grew and changed over time and expanded to other PFGE patterns.

Characteristics of this MDR *Salmonella* Infantis. It does have this IncF1B-like plasmid, it has a number of resistance genes, most concerning is the CTX-M-65 confers resistance to a bunch of different drugs including 3rd generation cephalosporins. It also has a gyrA mutation which will give a reduced susceptibility to Cipro. At the time it differed by 2-47 SNPs in the initial investigation. It did isolate closely to a chicken from the FDA.

So, at that time, we looked back at where we felt Infantis was coming from. This is an Asian phylogeographic analysis, what we call Bayesian analysis. It looks at relatedness over time. And on the little tree here, the ones in blue are travel associated or come from Peru. The ones in black from the U.S. What we really saw is that most likely the U.S. cases were coming from a common ancestor that started in Peru in 2008. That's this fork in the tree. After that you can see multiple introductions to the United States due to travel, black boxes represent travel. And then later on in the temporally you see at the very top a whole number of FSIS isolates from chicken and human isolates were highly related to those as well. I'll say at that point, it seemed that this strain really spread across poultry production in the U.S. So, it started off as probably travel associated in humans but then it really spread to many of the different production facilities.

And then in 2018, we had our first outbreak investigation of this Infantis strain, and it was linked to raw chicken products. We had some difficulty in finding this and that was due to diversity of PFGE patterns. Before that, 2018 time you can see at that time the initial PFGE pattern only 14 percent had that initial pattern. That was really the difficulty in detecting it.

Since then, I think most of you are aware that this has really become a persistent strain, the *Salmonella* Infantis. And what we know is that this strain tends to carry this plasmid. So, for us the concern really is within that plasmid itself. So, what's so concerning about this plasmid? Well, it has a number of AR genes and that’s obviously concerning on the clinical side. It also has several different fimbriae. It has a Yersiniabactin iron acquisition operon and those two things are thought to maybe give it an advantage in spreading over other strains, fitness-wise. It also has some heavy metal resistance genes and some antiseptic resistance genes. Both of those may be important in it surviving and spreading in production facilities.

Our question here is really what happens if this plasmid moves into another serotype? Will that then serotype have this sort of advantage. What if that’s in a source that's not chicken then is that going to be the next sort of problem. And through EAR and the work of others just recently, FDA released this notification where it looks like we do have that plasmid now in *Salmonella* Senftenberg and at the time I think we had two or three cases in humans as well that got sick with this. And I think those are the two examples I wanted to mention. And I will stop there and see if their questions.

Thank you, Jason. And I'll give people a moment to type their questions into the chat or the Q&A if they’d like. I would just like to, I know you ended on the fitness note there Jason. I wonder if you can expound a little bit how that may be playing into, for example, the EAR-WG work whereas I understand the potential emergence or transmission of this AMR gene or to a specific background is a reason for an alert. Now when you're talking about emergence and persistence, are there kind of more fitness elements that are going to be playing into that down the road?

Yeah, we're certainly now that we've established the alerts, I think we're all now comfortable with them. I think most of us have relatively easy ways of doing that on a monthly basis. Just the last meeting we did extend the idea that, certainly if there are other things that people in the group want to alert on, we're more than happy to add those in as well. And I think the value is that the different agencies have these different systems or sometimes it might not easier for them to detect a plasmid or a specific antiseptic gene that maybe CDC is not looking for because it's not in our ResFinder database.

Thank you, Jason. Okay, well, I think with that, uh, here today, we're close to ending. I would like to thank everyone from the different agencies who took the time to speak today. I personally got a lot out of it because I feel like you know a little part. I know my part at the FDA, retail NARMS work but it's great just to hear about everything going on across the different agencies. I want to thank Jason for talking twice today on it, the begin and end. We really appreciate that and I'd also like to say, hopefully this will lead into a better understanding of the data for the next two days at the public meeting. And I know also all of the presenters today, you know, emphasized the fact that, if you have questions, if you have ideas to reach out to us. You know it's people who build dashboards, data scientists and musicians we all like an audience so if we know people are looking at it and they have questions, other features, other views you'd like to see we'd love to hear from everybody. With that I would just like to thank everybody. I will give the panelist a brief moment to open their mics and say anything they'd like before we end the day.

I just want to thank you Errol and others for organizing this workshop. I think it was great

Thank you, Jason. And I'm glad to see people are saving their energy for the next few days. I look forward to seeing you all online tomorrow and Thursday. Thanks, everyone.