Cyberseminar Transcript

Date: July 9, 2018

Series: VIReC Database and Methods Seminar

Session: Phenotyping Physiologic Measurement of Lung Function in the VA EHR Using Automated Tools

Presenter: Kathleen Akgun, MD, MS

*This is an unedited transcript of this session. As such, it may contain omissions or errors due to sound quality or misinterpretation. For clarification or verification of any points in the transcript, please refer to the audio version posted at* [http://www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm](file:///C:\Users\VHASLCMyersK\AppData\Local\Microsoft\Windows\Temporary%20Internet%20Files\Content.IE5\73RXHSNF\l)

Hira: Hi everyone, and welcome to Database and Methods, a Cyberseminar series hosted by VIReC, the VA Information Resource Center. Thank you to CIDER for providing technical and promotional support for the series. Database and Methods is one of VIReC's core Cyberseminar series, and in it, we really try to focus on helping VA researchers access and use VA databases. This slide shows the series schedule for the year. Sessions are typically held on the first Monday of every month at 1 p.m. Eastern. Most session topics for the series are updated every year. More information about this series and other VIReC Cyberseminars is available on VIReC’s website, and you can view past sessions on HSR&D’s VIReC Cyberseminar archives. A quick reminder to those of you just signing on, slides are available to download. This is a screenshot of a sample e-mail you should have received today, and in it you will find the link to download the slides.

Today’s presentation is titled Phenotyping Physiologic Measurement of Lung Function in the VA EHR Using Automated Tools, and it will be presented by Dr. Kathleen Akgun. Dr. Akgun is a staff physician at VA Connecticut Healthcare System where she serves as director of the medical intensive care unit, co-chair of the Clinical Ethics Committee, and VISN 1 co-champion and chair for the Life-Sustaining Treatment Initiative Task Force. She is also assistant professor of medicine at the Yale School of Medicine. Her research interests include pulmonary and critical care medicine. Most recently she worked to develop novel informatics tools to harness the richness of clinical notes to improve measurement of patient-centered healthcare domains. Thank you for joining us today, Kathleen.

Dr. Kathleen Akgun: Thanks so much for having me here and Heidi for guiding me through this process. I am looking forward to having this next hour or so to discuss some of the work we’ve done here at VA Connecticut to enhance our ability to phenotype lung function in the VA EHR using automated tools.

At the end of the Cyberseminar, participants will be able to describe challenges to extracting pulmonary function tests from VA electronic health record data, understand the benefits and limits of automated tools that can extract these PFT numeric values from the EHR, and name at least two potential implications of automated tools for extracting PFTs from the VA electronic health record.

Our session roadmap will first give us some clinical background for the importance of this work, which is truthfully where most of my activities lie, and thinking about what PFTs are, what their relevance is to thinking about chronic obstructive pulmonary disease, and why we should care for COPD research. We’ll then dive in to talk about the existing data sources for identifying PFT values, what methods are available for using these numeric values, and then provide some of the work that we’ve done using a structured query language tool for extracting PFT data from the VA EHR. I’ll wrap up with a clinical example of what we’ve been able to do with this work so far.

In order to get a sense of who we have participating on the call today, I'd like to introduce a poll question, and I'm asking you to let us know your interest for VA data. Let me know if you’re primarily someone serving as (a) a principal investigator or co-PI; a research staff member such as a project coordinator, data manager, or programmer; a clinical staff member; operations staff; or some other function that you could potentially let us know about. And we’re going to open that poll now.

Molly: And the poll is open. Responses are coming in. We’ll give everyone just a few more moments to respond before we close it out. And Kathleen, I just want, while people are responding, we got a note in, someone from the audience. If you could speak up just a little bit, people are having just a little trouble hearing you.

Dr. Kathleen Akgun: Sure, thank you.

Molly: Thanks. Okay, it looks like we’ve slowed down on the poll, so I'm going to close that out. And what we’re seeing is 7% of the audience saying principal investigator or co-PI, 52% research staff, 11% clinical staff, 15% operations staff, and 15% other. Thank you everyone.

Dr. Kathleen Akgun: Thanks so much. So, getting into why PFT work is important and what it means for COPD research, I first wanted to orient us to the problem of chronic obstructive pulmonary disease. COPD affects 7 to 12% of adults worldwide and is the third leading cause of death in the United States. It’s identified clinically, as is demonstrated with the cartoon on the bottom right-hand corner of our screen, where the airways in the lungs get narrower and narrower and globbed up with mucus that makes it harder and harder for patients in order to breathe and exhale their breath at the end of a typical breath cycle. In addition, there are emphysematous changes that are demonstrated by the second inset cartoon here where the surface area of the lung, where gas exchange is necessary to happen, gets destroyed from chronic exposures to things such as smoking cigarettes. And that makes it harder for the lungs to be able to then exhale and get out the waste gas at the end of a respiratory cycle.

So in the pulmonary world, we are listening for these types of stories from patients in terms of the symptoms that they have but then rely on pulmonary function tests, or PFTs, to better ascertain whether somebody has obstructive lung disease such as COPD. PFTs are critical to this identification and have three primary areas that are measured. First is spirometry, which is our primary measurement for airflow obstruction. Second is total lung capacity. That lets us get a sense for the overall volume that the lungs can take in. And the third number, or the third domain is diffusion capacity, which lets us know about the thickness at the level of the air cells, between that and the blood vessels where gas exchange needs to take place.

For the purposes of this talk, we’re going to be squarely focused on spirometry, which is what is used for determining whether somebody has airflow obstruction. This is a forced maneuver that’s done with coaching of the patient that’s represented here in the green robe here holding a chamber that’s closed, with a nose clip as well, where they forcefully blow out a breath and are asked to try to sustain that for six seconds at the very least to have a valid study. This data is then collected by the PFT software and throughout is being coached by a pulmonary function technologist or respiratory therapist to assure that this is a valid study. The number that is determined from that chamber is a volume in liters that is then translated into a predicted value based on the patient’s age, race, gender, and height. And the percent of the predicted is what we use in order to determine lung airflow limitation severity.

So in the setting of somebody having symptoms, we are really looking for what’s called this forced expiratory volume in one second, or FEV1, which is a numeric measurement in liters. It’s commonly used and highly reproducible. And once we get the overall number in liters, they are categorized into groupings according to the global initiative for chronic obstructive lung disease, or GOLD, severity stages. GOLD severity stages vary from one, which is the most minimal amount of obstruction, and up to four, which is the most severe level of obstruction that we observe, again using this FEV1 value.

For completeness’ sake, it’s also relevant to just state that there are radiographic evaluations for emphysema or lung disease with something like a chest CAT scan, but this is beyond the scope of what this talk is going to include.

So accurate identification of COPD and its severity are important for epidemiologic and clinical research. For example, being able to accurately ascertain the presence of airflow limitations and its severity can then help us to determine whether patients are receiving guideline-concordant treatment and management, help inform population estimates for prognosis, and potentially could be used to identify potential participants for clinical trials in improving the care for patients with COPD.

PFT data is the reference standard for measuring COPD and its severity and typically comes as a printout for clinicians that looks something like this where we have on the far left here a sort of alphabet soup of things that start with forced volume. In our second column are the units of what these measurements are going to be. Our third is the predicted value. The fourth here, this pre-drug reported, is that raw number of what that exhalation looks like. And then here we have the ratio of the reported value compared with the predicted value, and this is the percent predicted that then is translated into these GOLD categories for severity of airflow obstruction. So these are what the data look like for clinicians, but again, this ends up being scanned into the computer in such a way that it cannot easily be captured for researchers routinely.

And so I wanted to pause here for a second poll question and get a sense for the audience’s familiarity or knowledge for the methods of ascertaining PFT data for Veterans in the VA. A would indicate no knowledge, and five would be on the level of expert for knowledge of PFT data in the VA.

Molly: And responses are coming in. Again, we’ll give everyone a few moments to respond before we close the poll out and go through the results.

Dr. Kathleen Akgun: And please let me know if I should speak up further.

Molly: I haven’t heard from anyone else, so I think we should be okay now.

Dr. Kathleen Akgun: Thanks.

Molly: Okay, it looks like we’ve come to a stop here, so I'm going to close this out. And what we’re seeing is 56% of the audience saying they have no knowledge, 22% rate themselves at a two, 16% rate themselves at a three, 6% at a four, and zero at five. Thank you everyone.

Dr. Kathleen Akgun: Okay, thank you. And so this reassuring that it’s not as though there’s some sort of Holy Grail for how to best get the PFT data out of the VA yet, and this is a problem that as my clinical role it comes up quite frequently. And so we considered how we could improve the ascertainment from using existing data sources from PFT values in the VA EHR.

Our first impulse is to use something that’s very structured such as ICD-9 codes for identifying COPD. However, these don’t have much in terms of severity indicators beyond indicating that someone has an exacerbation. We can improve performance of ICD-9 codes if we consider pharmacy data to identify whether someone is getting maintenance treatment or has recently received antibiotics or specific medications that are typically used for exacerbations. But neither of these approaches are really going to give us a numeric estimate of what that spirometric FEV1 objective value for airflow limitation is. And so we were looking to use the Corporate Data Warehouse, or CDW, for extracting these FEV1 values, and we wanted to poll the question for the audience again to ask how often folks in the audience have used CDW data to extract numeric values. A is yes, both the master and mini file; and B is yes, only the mini file; C, yes, only the master file; and D, I haven’t used the CDW to extract numeric values.

Molly: And responses are coming in. Again, we’ll give everyone a few moments to respond before we close the poll out and go through the results.

[Pause 12:46 to 12:54]

Molly: Okay, it looks like we’ve slowed down here, so I'm going to close this out. And what we’re seeing is 30% of the audience saying yes, both the master and the mini file; 3% saying yes, only the mini file; 7% saying yes, only the master file; and 60% have not used CDW to extract numeric values. Thank you everyone.

Dr. Kathleen Akgun: Thank you. So VA data sources include a comprehensive electronic health record data that has structured fields but also has free-text data that is generated from the clinical work that’s being done throughout facilities in the country. The Corporate Data Warehouse is sort of the back end where the data is stored in both the structured forms as well as unstructured ways. And so we can do things like search for ICD-9 codes, so standard coding groups for disease conditions, but also look for procedures such as pulmonary function tests to determine whether somebody has had this type of testing done, which would almost certainly need to be done before we could ever find an FEV1 for the patient.

In addition, though, the CDW also houses unstructured text from progress notes, radiology reports, and pulmonary function test notes, but that also will require manual input from the clinicians at the facilities throughout the country. The CDW also has some raw PFT data such as the FEV1, but it’s in a number of different forms and isn’t always normalized to the percent predicted. So we first needed to look at the raw numbers in CDW, but also look further to see if the free text could be another source for these types of values. And so we sought to develop tools to extract FEV1 in order to advance our understanding of COPD in large patient populations and found the VA electronic health record to be particularly well-suited for this type of work.

In terms of the methods that we used for doing this, again there was this PFT domain that was available but did require just normal percent predicted computations, but it required very little cleaning. The problem with the CDW PFT domain was that the PFT software had to be linked to the clinical workstation in order for this information to then be stored in CDW. And asymmetric upgrades between the PFT software and the clinical workstations made it so that these data fields were frequently missing and increasingly so with time.

So when we looked for other places to find the numeric values for FEV1, we looked at the Text Integrated Utilities free-text data, which is basically what the progress notes get stored as, and created a structured query language, or SQL-based tool, to do full-text key word extraction using a SQL algorithm. We created the algorithm looking at FEV1 and its variants and also identified key negation terms that needed to be included to have a reasonable instrument to use.

And so I wanted to ask another question of the audience in terms of how familiar folks were related to structured query language tools, or SQL tools. A indicates not familiar at all, and E is extremely familiar.

Molly: And responses are coming in. We’ll give everyone just a few more moments to respond, then we’ll close it out.

[Pause 16:38 to 16:43]

Molly: Okay, it looks like we’ve come to a stop, so I'm going to close this. And what we’re seeing is 30% of the audience saying they’re not familiar at all, 14% rate themselves at a two, 30% at a three, 11% at a four, and 16% rate themselves at a five. Thank you everyone.

Dr. Kathleen Akgun: Okay, so if you remember, I started this talking about the importance of PFT data but also showing you what it looks like for clinicians. And so while it has this sort of structure for clinicians to review and we’re very comfortable how to find our FEV1 values, this is what it ends up looking like in the CDW TIU document IDs where you see just a number of characters indicating likely columns that were set up for the clinician to compare the PFT results with prior studies, but it’s rather cumbersome to look through and try to pull out where the FEV1 values are. Further, what we found is that often there will be more than one value listed in the TIU documents. And so we needed to be able to also consider how best to develop a tool so that we were getting the blue values highlighted here rather than these red values that are representing prior FEV1 values for the patient.

This is another example of what the PFT data looks like in the TIU documents where, again, we can see just the percent predicted according to that ratio of the actual number in the predicted value, which is important and could be useful, but is another consideration that we had to take as we were creating a SQL tool for extracting FEV1 values from the electronic health record.

To do this work, we used the Veterans Aging Cohort Study, which is a large cohort of HIV infected and uninfected Veterans throughout the country matched one to two on age, race, gender, and site of VA care. Data collection began in 1996 and is ongoing, but for the purposes of this project, we included those from 2000 to 2015. Our initial SQL tool development in VACS restricted to patients who had a COPD diagnosis using ICD-9 codes between 2000 and 2015. We then restricted only to those patients with COPD and CPT codes for pulmonary function tests in order to be able to have a first-pass high yield for looking for FEV1 numeric values in the electronic health record.

In order to do the SQL tool, we first had the document corpus and identified a set of documents, big D, that contained the key word using full-text search. Within each document, little di in the corpus, we extracted 20-character snippets, each which began with FEV somewhere in the snippet. We then created a substring of snippets that included FEV with various dashes, equal signs, and ones to, again, start to hone in on the strings that are going to be most fruitful for us to be able to identify a quantifiable FEV1.

Within the snippet substrings, we then extracted numeric values, Vj, and set further definitions for where that number could lie based on clinical plausibility between 0.5 and 5.5 using expected liters for FEV1. We then implemented the algorithm using SQL with full-text search feature that was supported by Microsoft SQL server.

In order to determine whether the tool was helpful, we did chart review by a pulmonologist that was treated as the reference standard and compared those chart review results with the CDW FEV1 structured value for those that had it, but also the SQL FEV1 value that was extracted using this tool.

We were able to identify 5,958 unique patients with approximately 18,000 documents including FEV1 snippets. We found 12,000 patients who had an FEV1 numeric value using structured CDW data, but another 3,800 or so that we were able to extract using the SQL tool. This led to an increase in 31% of the number of quantifiable FEV1 values that we had compared with using CDW alone.

In a subset that we did chart review for, we compared the SQL tool with the chart review numeric value and found a positive predicted value of 89% for the SQL tool compared with the reference of chart review, kappa for correctly identifying an FEV1 entity of 0.66, and a Spearman’s correlation among those that did identify quantifiable FEV1 of 0.99. Based on that information, we continued to modify our rules for the tool and have improved performance since, but that will be for a separate presentation.

And so I'm going to wrap up thinking about a clinical use of the SQL tool for extracting PFT data in VA electronic health records. And for this we are going to talk about lung cancer and its association with COPD. Established risk factors include smoking, age, COPD itself, and several occupational risk factors as well. However, there are inconsistent associations between COPD severity, or FEV1, and lung cancer risk where some groups have shown that more obstruction is associated with more risk for lung cancer, but this is not something that’s uniformly been seen throughout populations. And given the group that we were working in with VACS, the Veterans Aging Cohort Study, we sought to determine whether there were some different interactions between airflow limitation and lung cancer risk between HIV infected and uninfected patients.

Just for some background for the audience, it’s important to know that our group has previously shown that the incidence of both COPD and lung cancer are higher in HIV infected patients, represented by the dark bars here, compared with uninfected patients in the gray bars, and that this association is seen throughout all age strata, which is represented on the X axis. Again, this is an association that’s seen for both COPD and lung cancer where HIV infected patients have a greater incidence of both of these conditions throughout the age groups.

As I mentioned, we asked how severity of airflow limitation impacted the lung cancer risk in COPD and whether there was a differential effect according to HIV status using the Veterans Aging Cohort Study.

We again restricted to patients who had COPD diagnosis based on ICD-9 codes. Our primary outcome was incident lung cancer, and we determined severity of COPD by extracting FEV1 values greater than or equal to six months from the cancer diagnosis using both structured CDW fields and unstructured data. We categorized numeric values for FEV1 to the GOLD stages of one to four, and remembering that higher stage is indicating greater severity of obstruction. But due to small numbers in the GOLD stages three and four, we put that into a combined group.

What we found were 8,612 patients, 27% of whom were HIV infected, who had both COPD and a quantifiable FEV1 that we could do for this work. Age and race were similar between HIV infected and uninfected patients. Our HIV infected patients were more likely to be current smokers compared with uninfected patients. And the GOLD stages were similar between HIV infected and uninfected patients with 36% having GOLD stage one, so least severe disease; 41% having GOLD stage two, which is the next level of severity; and 22% of patients having GOLD stage three or four severity of airflow limitations.

We also found, similar to prior work by Dr. Crothers, that HIV infected patients continue to have a higher incidence of lung cancer at 709 cancers per hundred thousand person years compared with 461 incident lung cancers in uninfected patients per hundred thousand person years.

In multivariable modeling, accounting for age, race, HIV infection, and smoking status, we did find that there was an association between severity of airflow limitations using GOLD criteria, GOLD stages, and risk for lung cancer incidence. As you can see, there was also a tendency to have a greater risk associated with greater severity of airflow limitation, although this was not something that reached statistical significance, and there was no formal interaction between COPD severity and HIV for lung cancer.

So in conclusion, we’ve talked about the difficulties of ascertaining COPD severity in the electronic health record but found that FEV1 values can be extracted using SQL queries with excellent ascertainment and good accuracy and that we can make substantial contributions to the amount of complete FEV1 values that we can use within VA data. We also have begun answering questions related to associations of COPD severity and lung cancer risk in both HIV infected and uninfected patients without an obvious interaction between HIV and COPD severity for developing lung cancer.

Our next steps are to further phenotype COPD in VA populations incorporating FEV1 as a disease severity marker, and also for internal validity to determine whether the SQL tool values have similar predicted values for outcomes as we’ve seen in other COPD studies. We also intend to look at the association between airflow obstruction in all comers, regardless of COPD diagnosis, and in clinical outcomes, and also would like to start to look at using these types of rules for other electronic health record systems such as Epic.

So I want to thank my co-authors, Keith Sigel, Kei Cheung, Farah Kidwai-Khan, Alex Bryant, Cindy Brandt, Amy Justice, and Kristina Crothers, as well as our funding from the funders that are listed to the right, and of course the Department of Veterans Affairs.

There are additional resources available through VIReC for specific questions that are listed here and will be available through the handouts.

There are also some quick links for VA data resources that are included, including to the Cyberseminars, VHA Data Portal, VINCI, the Health Economics Resource Center, more information about CDW itself for those that are interested, and archived Cyberseminars. And for more information, you can contact VIReC directly or feel free to e-mail me at [Kathleen.Akgun@va.gov](mailto:Kathleen.akgun@va.gov). So with that, I'm happy to take any questions.

Hira: All right. Thank you, Kathleen, for your presentation. I have one question here so far from the audience. If anyone else has any additional questions you’d like to ask, please send those in now. We do have several minutes left until the end of the Cyberseminar. All right, the first question: To get numeric FEV data from an externally performed spirometry test, is TIU the most likely or the only source, and what is the rate of coverage of external PFTs in the TIU files?

Dr. Kathleen Akgun: Let me just start with the first one. They’re both very good questions. I want to make sure I clarify the second one.

Hira: Okay.

Dr. Kathleen Akgun: For the first question, in terms of the TIU documents as source of FEV1, that was something that we were able to observe over time as becoming increasingly frequent relative to the CDW PFT domain. And that’s exactly because of the software disconnect that occurred, I think it was eight or nine years ago or so, where no longer were the PFT software able to feed the information back into CDW. So TIU is increasingly going to, I think, be the source for reliably extracting FEV1 data from VA EHR data. So it’s not the only source, but I think it will end up being the more reliable one. When we had looked at it, I think it was down to about 9% of facilities that were able to still have PFT data from the CDW PFT domain. For the second question, if you would help me here just to try to repeat that to me?

Hira: Yeah, sure.

Dr. Kathleen Akgun: It was about the external sources.

Hira: The question was what is the rate of coverage of external PFTs in the TIU files?

Dr. Kathleen Akgun: I'm not sure that I know the answer to that.

Hira: Okay. That’s fine.

Dr. Kathleen Akgun: Is it, I wonder if we’re asking about this as PFT studies that are being done somewhere other than a VA facility, that the information has been transcribed into VA data TIU documents?

Hira: Yes. The person just wrote in yes, other than the VA.

Dr. Kathleen Akgun: Okay. Yeah, that’s a good question, and that’s something that we have not yet determined. But that is something we’ve also been aware of with our best estimates going to be derived from the TIU documents versus CPT codes for the PFT. So that will be another piece that we’ll be looking at. That’s a very good question too.

Hira: All right, thank you. I have a couple more questions here. Do you perform analyses using the lung function FEV1 from CDW only? If so, do the results from CDW only differ from those from CDW and SQL?

Dr. Kathleen Akgun: So that’s also something that is on the horizon for our next step once we felt confident with the tool was to look at those who have CDW FEV1 values only versus those that have CDW or SQL and try to compare the performance of those two. But we have not yet done those analyses.

Hira: All right, thank you. Would it be possible to get a copy of the SQL query?

Dr. Kathleen Akgun: I imagine it would be. You mean the coding that was used to extract the FEV1?

Hira: Yeah, I think that’s what the person is referring to, right.

Dr. Kathleen Akgun: I imagine that’s something we could do, and if you want to e-mail me offline, I'd be happy to just confirm that with my co-authors, but I think that would be something that we’d be happy to share.

Hira: All right, thank you. I have one more question here. What exactly is phenotyping?

Dr. Kathleen Akgun: Oh, so phenotyping is where we’re getting a sort of, I should have defined that sooner, a picture of what the disease condition looks like for the individuals. And so it’s sort of the blueprint for what the condition is for the patient that’s experiencing it. And so given that we don’t have perfect markers for COPD in terms of biomarkers, we have to rely on both clinical history, exacerbation history, burden of symptoms of shortness of breath and mucus production, and triangulate that with what objective measures we have such as FEV1 values to develop what the phenotype for patients with COPD is like. So it’s trying to have a more structured categorical assessment of COPD for these patients.

Hira: All right, thank you. It looks like those are all the questions we have so far. Was there anything else you would like to expand on in your presentation?

Dr. Kathleen Akgun: I think additional next steps that we’re in the process of are also just to, remember I talked about the different domains that are included in PFTs, but there was also other values that we look at such as the forced vital capacity, or FVC, and the total lung capacity and the diffusion capacity. So those are additional areas in the PFT results that we’re also going to be working with once we iron out the details for the FEV1. But we felt this was the first step in order to start to get a better quantification for COPD severity and next plan to look at these other measurements that are obtained from PFTs fairly routinely.

Hira: All right. Thank you so much, Kathleen, for taking the time to present today’s session. To the audience, if you have any additional questions, you can contact Kathleen directly. Her contact information is on the screen. You can also e-mail the VIReC help desk at [virec@va.gov](mailto:virec@va.gov). Kathleen, if you can turn to the last slide. Oh wait, just got another question in.

Dr. Kathleen Akgun: Sure.

Hira: Okay. About how long did it take to extract FEV1 values for the 5,000-plus patients?

Dr. Kathleen Akgun: It was not overly cumbersome, says the non-informatics person. The bigger challenge was just the iterative rule development. And so that probably was overall a three- or four-month process of going back and forth between Kei and Kristina and Keith and I to determine what seemed to be most plausibly FEV1 values and what was something else and what that something else was in order to then modify our rules so that they were able to capture FEV1 more precisely from the TIU documents. So the programming part is not all that difficult. It was more just about the iterative process of developing the rules for the SQL tool.

Hira: Okay, thank you. All right. So everyone, you can tune in for the next session in VIReC’s Database and Methods Cyberseminar series on Monday, September 10th, at 1 p.m. Eastern. Dr. Denise Hynes will be here to present Comorbidity Measures using VA and CMS data. We hope to see you there. Thank you once again for attending.

[ END OF AUDIO ]