Cyber Seminar Transcript  
Date: 09/17/2015

Series: VIReC Good Data Practices  
Session: Mind the Gap: Using administrative and claims data to answer your research questions

Presenter: Todd Wagner  
 *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* [*www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm*](http://www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm)*.*

Unidentified Female: Hello and welcome to VIReC's Cyberseminar Series, Good Data Practices. The purpose of this series is to present and describe issues related to good data practices and discuss examples from the work of VA researchers. Before we begin, I want to take a moment to acknowledge and thank some of those who have contributed to the series. We have an active and valuable advisory group that guides our choice of topics and concepts to be covered in these sessions each year. This year's advisory group included Matt Maciejewski, Peter Groeneveld, Jennifer Garvin, and James Burgess – Jim Burgess.

This year's good data practices Cyberseminar series includes four sessions presented on Tuesdays and Thursdays, this week and next. On this past Thursday, Sarah Krein in this session on Planning for Data Early, Often, and Ongoing reviewed the importance of planning for data early in the research process. If you were not able to attend the presentation, it will be available on the HSR&D website – Cyberseminar web page soon.

I am sure Molly can give us the information about that at the end of the session today. Today, Todd Wagner in a session entitled Mind the Gap using administrative and claims data to answer your research questions will discuss approaches he used to handle data limitations and data linkages.

Next Tuesday, the 22nd, Denise Hynes will present the session on Selecting Methods and Tools for Data Analysis. In our fourth and last session on September 24th, Stephen Deppen will tell us about how the VA's instance of the REDCap Data Collection and data management application can be used for studies that combine primary and secondary data. With that said, Hera, would you like to take over?

Unidentified Female: Thanks Linda. Today's speaker is Dr. Todd Wagner. Todd currently directs HERC, the VA Health Economics Resource Center. He is the Associate Director for the Center for Innovation to Implementation of the Palo Alto VA. Over the past 20 years, his research has focused on how health information affects consumers and providers; and how to effect the value of care. He has developed the wealth of experience analyzing healthcare costs leading to over 100 publications.

Most of his work involves estimating costs or analyzing costs for large administrative databases like VA Medicare. He is also involved with the local HSR&D fellowship program. He helped to develop a new training program in big data for oncology research. Any questions you have, please send them through the chat box. I will monitor them during the talk. I will also present them to Todd at the end of the session. At the end of the session, a brief evaluation questionnaire will pop up. If possible, please stay until the very end and take a few moments to complete it. Without further ado, I am pleased to welcome Dr. Todd Wagner.

Todd Wagner: Thanks Hera. Can you see my slides?

Unidentified Female: Yeah.

Todd Wagner: Great. I think invited me to give this presentation because they knew that I had made so many mistakes over the years with data. I have hopefully learned my lesson and can be a little bit more clear and helpful for other people. Alright, I am going to give you a walk through and talk about data in the VA and also in Medicare. I am going to do it within the context of a big study that we just finished. It is called the intended and unintended effects of large scale adverse event notifications.

You can see the paper there in all of its glory. It has to do with medical errors. I just want to be very careful that my opinions and my views do not reflect those of the VA. Also, I have no conflicts of interest to report. This study was funded by HSR&D. We were going to do a poll. I was not sure if we had… They keep changing the platforms on us.

Unidentified Female: No problem. I got you covered. Alright.

Todd Wagner: Awesome, Molly, thank you.

Unidentified Female: Yeah. We do have the poll question up. We are trying to get an idea of what your role in research is. Are you a research investigator, a data manager, or analyst, a project coordinator, or other? If your answer is other, go ahead and type your response in directly to the question section of the question box that is in your GoToWebinar dashboard.

It looks like we have a nice responsive audience; 70 percent have already voted. A couple of people are writing in. But somebody's position is a statistician. Another person is a librarian. It looks like we have got a good indication of our audience. I am going to close the poll and share those results now. Todd, do you want to talk through them? Or, would you like me to?

Todd Wagner: I do not see them yet. Do you want to –?

Unidentified Female: That is okay. They are behind your slides. A third of our audience are research investigators; 47 percent data managers, or analysts; 12 percent project coordinator; and 9 percent responded as other. Thank you to our respondents. I think we have one more poll to get an idea of how many years of experience do you have working with the VA data? The answer options are less than one, one to three, three to six, or seven or more. We have already had two-thirds of our audience vote. The answers are still streaming in. We will give people a little more time.

A nice spread across the board here, okay. We are about 80 percent response rates. I am going to go ahead and close the poll, and share those results. Almost half of our audience, 43 percent have less than one year experience with the VA data. About a quarter of our audience have one to three years’ experience; 15 percent, three to six years’ experience; and 17 percent have seven or more years of experience. Thank you very much. We are back on your slides, Todd.

Todd Wagner: Thank you so much. That is an amazing spread. I am really excited to give this talk. I will hopefully be able to hit both the experienced and the less experienced VA data users. I am going to talk about really four things in this study. I will use these little red stars you will see up here on some of the slides. That is going to be slides that I am wanting to emphasize. But in this talk, I really want to identify clinical administrative claims that may be relevant for research and ways to think about those data. Each of the data sets that you can use have strength and limitations.

I just want to want sure people are aware of those limitations with their data sources. I want to highlight the data linkage methods. There is a reason for that. I think a lot of people when they work in merged data that is where errors come into play. That really is to my personal experience. That is where if anything happens, that is what happens. I want to sort of stress that issue. Then I am going to talk about methods for good data management practices. Again, this is all within the context of a study. Here is the context of the study. But we recognize that errors happen in medicine unfortunately.

There are types of errors known as large-scale adverse events. These are errors that expose more than one person to an infectious disease. An example might be improper cleaning of medical equipment. I know what you are thinking. You are like that is gross. That should just never happen. Let us talk about stainless steel. When you realize that the complexity here is incredibly diverse; and it is not always easy to recognize. You realize there is a lot of challenges here. There are a number of types of stainless steel.

There are three major types that are used in medical applications. Type 304, is things like medical trays and equipment. One of these types of scalpels. I forget which one is a 304 or one of the clamps is a 304. There is 316 stainless steel. This is things like needles, syringes, catheters, piercings. If you have body piercings. You have probably got 316 stainless steel. The final one is what they think of as surgical stainless steel, 444, 420. All of these differ in their metallic properties.

The key here is that the cleaning protocol differ depending on the type of stainless steel. If you auto clip; you draw all of your stainless steel together in the autoclave. Autoclave it for the exact amount of time, that is perfect for some of the stainless steel and not ideal for other stainless steels. It leads to small increased risk of things like micro pitting in the stainless steel and the ways that it could infect people.

I know that we think of errors as being these terrible things. But you have to think about my goodness. If it is really hard to track all of the different types of stainless steel let alone everything else that happens in a medicine surgical suite. I am getting some feedback on the line. Just so if you are not out there; if you could mute, that would be great. Thanks.

What we do not know is how patients respond to being notified of one of these large-scale adverse events. I am just going to talk to you a little bit about what happens in the VA is that the VA sends out a certified letter. If you are one of these receiving patients, it is not going to be all of that enjoyable to receive one of these letters. You have to open up a couple of the certification across it. \_\_\_\_\_ [00:10:00] tells you that you have been exposed. It tells you what you should do next.

Communicating about errors is one of the things that physicians really struggle with or clinicians really struggled with because there is not only telling about happens. The implications and the processes of what went wrong. But then you will need to engender trust and guide them to appropriate action. That can be things like we need you to get tested for infectious disease; HIV, HPB, HCV. You need to be able to do that in a way that is informative, caring, and instructive. Often the clinician who does that is not the patient. It is the clinician who normally sees that patient. Some of these exposures and when a facility finds out, they bring in a clinician to review all of the cases. Then there is a panel that is sort of contacting the patient. Those are not the regular doctors that the patient sees.

Here are some of the challenges that you could think about when it comes to these types of notifications. The risk of infectious disease from the exposure is not always well known. The absolute risk is very small. If you are a researcher, and if you know that it is very hard to communicate risk to people. Patients do not even understand risk when it is things that use daily, helmets for bicycle riding; all sorts of other things, drinking and so forth. When we are talking about extremely small risks, those are also hard to communicate about. Then the clinical risk and the perceived risk may vary considerably especially when the patients get these letters. By statute, VA must provide timely and accurate information.

Then most clinicians like I said who are involved with notifying are not – do not have a prior relationship with the patient. This is the day of Twitter. This is the day of websites. We know that things go viral. Bad communication is something that is spread like wildfire. It is not just the VA. This is true whether it is a private hospital. There have been evidence and examples in private hospitals where this has happened too.

What makes VA a little bit unique in this scenario is that VA has to identify it and report these, if it is known. It is not like you are a private hospital. You are going to well, we think it is so minor. We are not actually going to report it. The VA actually has to report it. Because they reported it, it gives us this chance to analyze patient behavior. Figure out, one, are the patients who get these notifications doing what they were told to do?

That is largely get tested for infectious disease. Then two, we get to figure out. We think of that as an opportunity, really. If they are not getting tested \_\_\_\_\_ [00:12:44]. Then they are not getting tested, are there certain subgroups of the patients who are not getting tested? The second thing is are there unintended affects? A lot of what we hear about from hospitals; CDC has a whole reporting system on this as well, if they help facilities walk through it – is facilities are really nervous about what is going to happen to them and their reputation should one of these things become known.

It is a challenge. They want to be open and transparent. They want to get ahead of it before the media gets ahead of it. It gets a lot of internal communication challenge within itself. There is a lot of potential for unintended effects here. When we think about unintended effects, we are going to get very specific about patient behavior. That is going to be things like do the patients leave the VA system and go get care elsewhere? Like I said, VA mandates reporting of LSAEs. We can observe these events. You see and there has been reports of LSAEs in commercial hospitals. But not all LSAEs are reported. There has been discussions about doing so.

The VA really provides a unique opportunity to analyze patient behavior. I want to be very careful here. Because we studied this in the VA, it does not mean that LSAEs are more common in the VA. It is just that they are mandated reporting in the VA. It is one of those things that people have heralded the VA for taking a major step forward in being a leader in adverse event reporting, and notification of errors and so forth. But it is not to say that the VA is more likely to have these types of errors. Please do not take from this that we are studying it in the VA because their errors are so much more common. That is not the case.

Before I get into the data, let us just talk about the specific aims. When you start a study, every study is going to have specific. Specific aims have to be objective and measurable. The specific aim one is infectious disease testing. Do exposed patients obtain follow-up testing for HIV, HBV, and HCV, so hepatitis B and C? We are then curious about do patients when they are exposed; and let us just call that – they were exposed for X. Do they come back for X at the same facility in the future?

You can imagine. Let us just say you went into a facility. You had dental exposure. You might say well I really do not want to go back to that facility for that same type of dental care. That is the kind of returning to the same kind of care in the future. There might be a broader issue, which is objective three here; which is this idea of switching and trust, which might mean maybe this event sort of caused you to really worry about just your trust in that system. You wanted to get your care elsewhere. When Medicare eligible Veterans are notified about an exposure, do they switch to a non-VA "Medicare provider"?

Now, let me get into the methods. I really beefed up this part of the talk to talk about all of the data linkages. This is where you are going to see those asterisks, those stars and so forth throughout. One of the keys here is that we had every LSAE that goes through VA is reviewed by the Office of Inspector General. Like I said, for some of these events – I will – there is a dental one that we will talk about. What happens is there is a clinician that reviews each dental case to identify whether that person was exposed or not. They might observe in the medical record that this particular tool or whatever it might be did not expose and was not used for the patient. The patient was not exposed. They were moved.

We needed a method, if wanted control. We needed a method that we could apply equally to cases and controls. There is no list of patients out there – there is nothing that says these patients were on sent leave, a notification letter. Here you could think of that synthesizing a control group. We came up with a method that would allow us to identify whether we think we are hitting the cases. But also a standardized method that could allow us to identify the controls. Let me just walk you through that. We are going to use six VA Medical Centers. If the controls are generally patients who receive the same types of services as the effected VA prior to exposure.

This was at a time when the exposure was not happening. Where it is also going to look at control sites. We are going to take the same pre-imposed set of control sites. We are also a little bit careful to think about this idea if there was an exposure and things went viral, what is the media market? We were trying to get controls outside of that media market.

What we thought of is that often used in economics is this difference and differences methodology. Let me just see if I can walk you through this. You can see on the rows of the site, you have the exposure site. You have the control site. Now before the exposure happened, you have the timing, which is the prior to exposure. Here is a control group that was at the exposure site but it was prior to the exposure.

Then you are going to have the control site. It is at the time of the timing. But it is just at a non-exposure site. Then at the exposure period, these are the people who were getting exposed. Then you have the controls. It is contemporaneous. If there was uniform improvements in safety over time, you would have a contemporaneous control. It is this upper left-hand box. You will have to remind me, Molly. Do I have – I was going to say, do I have a notation? I do have it.

Unidentified Female: Yes. You have an annotation tool right there.

Todd Wagner: Yes. Very good, yes, thank you, a pen. This is purple on my screen. That is the case group. You will see as I go through, statistically it is relatively easy to analyze these data. The statistical test basically says is this group different from all of the other groups? It is in an interaction effect, which we will get to. It is quite easy to do a statistic. But hopefully the notation and the method here, it is straightforward. Whenever I take a study, I have both the proposal that was written to get the money. Then I develop a data analysis plan. The data analysis plan is a living document. I tend to have these weekly meetings with our teams.

Every week we come up and we review parts of what we are putting together. That data analysis plan has notes of decisions made. It has tables for variables. But here is a highlight for you of what this data analysis plan grew – it eventually grew to be a 22 page data analysis plan. It includes a study overview. You will see things like here is a data set. This is the events file. The SE file, this is a VA outpatient care file. You get to see that we did not want every variable on the file. We just wanted a handful of files.

Just for the simplicity of this talk, I cut out a column here which was we also use the MCA, the Managerial Cost Accounting system laboratory file. You will see down at the bottom, there are some variables that are not on this list over here. But they were on the DSS or the MCA list. I apologize ahead of time. About a year and a half ago, DSS changed its name from Decision Support System to Managerial Cost Accounting system. But if you hear those two, they are the same. Today's current technology, the MCA is the preferred name. This data analysis plan, I cannot live without them. They go through. They are very specific. But they also – they do two things. They keep me as the PI in the loop. But they also do not make me micromanage.

When we get into studies, it is at the clinical trial, for example. It is 2,000 patients. There might be ways that we pull data that are very different. If it is 800,000 patients and you are pulling all of their SE data for 800,000 patients .I will be very honest that I used to love programming. I used to do it mostly in Stata. The technology is so different nowadays that I am not a good SQL programmer or even a good efficient SAS programmer. My goal is to hire great programmers. We work together as teams. But I tell them. I look at them and say here is where we are headed. Help us build this plan. You guys are the ones who are to figure the efficient use of how to use the resources to get there.

What we can then figure out; and I will show you is we build in data safety checks. There is not a concern did we mis-merged data. Or did we somehow do something that \_\_\_\_\_ [00:21:42] something that we were expecting to \_\_\_\_\_ [00:21:44]. That is part of this data analysis plan. But this is a really key document for me. The other thing about the data analysis plan is it becomes your methods part of the paper. It is really nice in that regard. Let me get back into this case identification. Like I said, there are these Offices of Inspector General. There is no publishing list that says, here are all of the patients who were sent the letter. We were really interested in developing this method using exposure dates, exposure clinics. We even knew according to the OIG reports the problematic procedures.

We could then translate those into procedure codes, CPT codes. Now this method does not exactly match the case method that OIG used when they went through to determine whether to send the patient this letter. I will get into that exactly. You will see how that effects our data. But then we were able to test at. We were also able to use this method for the control identification. All six LSAEs were in ambulatory care. We were able to use the SE data for all of them. Here you get to see, for example, there were two using endoscopy. There were two using colonoscopes. There were some in dental – two in dental.

Here were the effected CPT codes. We were able to say okay, here is what we think in these clinic stops. The CPT codes and bates; and we were able to pull what we think were the cases. We were able to use the same method for identifying the controls. Again, it is the same if you go back a slide. It is the same clinic stop and the same CPT codes. Now, we are just changing the dates and the station. That pulls our control group. If we decide to go back through and figure out what is a good control station; and then figure out what is a good control time period. For the station, we were very interested in geographical proximity.

We were trying to figure out, okay. If there is clinical practices who differ by region of the country. We wanted to get as close as we could geographically, but not be in the same media market. We tested out a number of control sites for that. We also recognized that clinical practices are different depending on the clinic side. If you are a huge volume site and you have an exposure, we would not want to compare you to a very small volume clinic. We are trying to get similar patient volume there.

Then like I said, we were trying to get people outside of the control and media market. For example, if you had an exposure. I am trying to think of you said Boston was an exposure. It is not. If you Boston was an exposure, you might not want to take Providence as the control. You might say that is going to get too close to media markets. But for example, you might say well, if it was San Diego with the exposure; and again, it is not. You might say San Francisco would be a great control.

Those are quite different media markets. For those on the East Coast, that is about a nine hour drive between the two even though it is the same state. Now, of course, media markets are different. Actually there is like a limitation there with today's digital age. There is no guarantee that some sense of the nation and the world is a media market today. But we tried to focus on what you might see on the local news, for example. We did methods qualitatively. We worked with Ronnie Ellie and her team. Then we are doing qualitative interviews on these news blurbs to try to figure out the good control sites.

Here is a table showing you the tracking of events. Here is an exposure. For each exposure, we chose two control sites. You have these – you notice the, in red there, is the notification date. That was the date that the letters were sent out according to the OIG reports. We wanted to go back in time to identify cases that would be people who were exposed. I am circling that. Then, of course, you are going to have to follow those, and that cohort after the notification to figure out did they do what they were supposed do? Or, did they do other stuff? Here is that outcome period that we are tracking.

You would have to go back. What we did is you would go back in time to do the case selection and the controls as well. This was following that forward. We did this for each set. For ease and simplicity here I am just showing you one of the exposure of the six. It was too much to show all of the six. Each exposure had different dates associated with it and making it a little bit more challenging.

Here are the sites. You would get to see it. I am not going to specifically say which types they are. But you get to see here are the six. Like I said there are two that were endoscopes. There were two that were colonoscopy related and two that were dental related. Now in the perfect world – and one of the things that I want to pull out here is the distinction between what we think of as the cases that we identified. What the OIG reported as exposed and who sent the letter. Here is the OIG report table. What they listed. There was one site, site 2 where the OIG did not actually specify a number.

We ended 378 cases in that site. You get to see most of the time the OIG was a little bit more generous. In site 1, we identified 675 people who we thought would be exposed. OIG said 1,069. Now we got one site 4 here in red. You will see that the difference is both large relative and large in magnitude, absolute magnitude. We said according to our methods that we have guessed that 6,100 people were sent the letter. OIG said no, only 3,200 were sent the letter. Boy, that is a huge misclassification both in absolute and relative.

We decided after discussions we were going to exclude that entire site in the control groups from that site. That is a big exposure site that is one of the biggest. We did not feel comfortable making claims about what those cases and controls were doing with 50 percent misclassification of 3,000 patients. But for the rest of the study, site 4 has been excluded from the analysis. One of the things that by developing this method is you can just sort of push this method and test this method; and see if it works for you. Then you end up with your control groups.

Now depending on; and you will see a couple of other things up here. It is probably worth just quickly talking about. Depending on the exposure period, we have some not availables. We have to work through that analytically. That is just because these were new clinics that were coming online typically. That is when they realized they were not quite doing things right. But there is no pre-exposure period to them. Here is the final analytical sample; 5,000 or 54,912 cases are people – cases, we have 9,638. This is – and let me just circle that. These were people who were exposed and would not – according to our data been sent a letter saying they needed to get testing.

We can go back in time. You have these three different control groups. Our method assumes and just be clear. That if there was no exposure that these folks would be the same. It would be sort of parallel trends over time. What were – the first goal that we are looking at and one is did these folks go back and report what we thought of as appropriate action which was getting tested HCV, HIV, and HVP testing based on CPT count; an also, using information from the DSS laboratory data.

Now just like we have these controls; we can use not only the data and the testing. We can information on timing of testing and so forth. For the dental care, that is a really interesting study. Because for most people, if they get dental care through VA, they do not have dental care otherwise. Unless you have a jaw problem, you do not get dental care routinely through Medicare. It is all private pay. You might, if you have it through your employer. Because that is because of your employer based health insurance plan and not through Medicare. Dental really provides this unique opportunity to look at further down and name two outcomes. Then the VA medical utilization for \_\_\_\_\_ [00:30:17] three outcomes.

Tracking end points for aim 1; so, here are the CPT codes, if you are interested in the gory detail of the CPT codes that are tracked. Did they get the test? Now you could imagine a clinician assigning one of these specific tests. Who says, I would really like to test you for HCV, or HIV, or HBV, or do that individually. A lot of times what happens and this is very common is they will do a panel of tests. You can see a panel of CPT codes. Here is where I also want to take a small segway. If you are a researcher in VA, it is really helpful to sit down with a clinician and walk through. Have them walk you through CPRS. It is really helpful to understand how clinicians are coding the data.

This is not like a commercial enterprise where clinicians are paid by the VA based on the CPT codes. In the commercial world, Part A, Part B, and Medicare payers are reimbursed based on the care they provided as tracked in the things like the CPT code. There is a strong financial incentive to not only report the codes and to be accurate in the codes. In managed care plans and capitated systems like VA, physicians here and there are employees. But they are not reimbursed depending on how they are coding. That is a weak incentive to be accurate in the coding. Yes, I have to understand those methods. You have figure out, okay. How are the data getting in the system?

If you are new to the system, I would really encourage you to sit down with the clinician and walk through how they are putting information like procedures in and diagnoses in the system. It is very helpful. For Aim 2, this is the probability of returning for the same type of care. Again, we are just going to focus on dental. We expand the timing of the dental a little bit because dental is often done in these six, 12 or 18 month periods. Think about yourself and when you go in for your dental cleaning. Many of you are going to go in every roughly six months. But it is not exactly six months. It might be six and a half months or 12 and a half months. This allows us to look a little bit say okay, is it six and a half? Is it 12? But by also using the 18 month period, we sort of catch up on maybe it was not 12 months. But it was 14 months.

What other thing that is really cool here was we can look at the type of care. You can imagine the situation looks different. You have been exposed dentally. You might say well, when it comes up for my next cleaning, I am not really sure what I want to do. I am going to hold off on that. That is a very different story if you have got a cavity. That is a restorative care. You might say well, I am in serious pain. I really want to get this cavity taken care of. Even though I was exposed eight months ago, I want to go back and get this taken care of. I do not like the pain.

Then, we sort of looked at all other cares, which can be things like root canals and what not. We did not use any Medicare data in this regard. Because unless you have something like facial surgery, Medicare does not cover just general preventive care for dental care. Aim 3, it gets a little bit more challenging. Here we are looking at this idea that if you have been sent a notification that you are going to switch your provider. You can imagine this affecting sort of the probability of use. You could also think of it as affecting the volume of use. We really focused here on people who are over 65 because we know that they are covered. Or most of them, 99 percent of them have Medicare insurance.

It is not true that no one under 65 has Medicare. People who are disabled also do. But we do not have easy flags on that to say here was somebody who was exposed. Who was under 65 but had disability insurance; and would have Medicare. It was a shortcut for us to say we knew that these people were dually eligible. Part of the trick here also was coming up with categories of care that make sense both on the VA side and on the Medicare side. Now you will see, for example, and I will just highlight. In Medicare we had more information on things like any inpatient short stay, long stay, and skilled nursing. We can keep those categories.

In many regards what you are going to see on the next slide is that you want to keep all of the data and categorize it somehow. The reason for this is you want to start off and say we saw 17 million visits and whatever it might be. I don't know the exact total. Let us just say it is 17 million visits. If you start focusing in on a population, you will never get back to that 17 million total. But if you keep that total at the beginning and do all your merging, your matching, and so forth, you should be still able to get back to 17 million of \_\_\_\_\_ [00:35:08]. Now you might use all 17 million. You might say I am really interested in this subpopulation. But that is a data check for you to make sure that you have done it. That you are not losing people in your merging – or your matching. Or if you are doing many too many merging, you could be making people even worse.

Here is another thing that I want you to think about. I have talked about this a little bit. For point 2, which is get into the CPRs, you need to understand the data generating process. That is really critical for data. Do not just assume that data or data – and you could just use them willy-nilly. There is a reason why the data exists and how they get in there. If it is a clinician and entering all of the data that comes into the National Patient Care database which feeds the SE data are clinician entered.

You have to understand a little bit about clinician behavior. That explains one of the reasons why people are interested in natural language processing with the text notes. But very different things than if you are focused on inpatient care in the patient treatment file, which is professionally coated. You also have to understand for Medicare that there is a relationship between coating and billing like I talked about. I will just \_\_\_\_\_ [00:36:13] one of my slides that I thought I had caught. My apologies for that. VA types of care – here is how we broke this out.

There are things called evaluation and management care. These are office visits. You can think about internal medicine, and coming back for E&M visits and not in internal medicine. You can use stop codes. There are many ways to categorize VA care in the data. We also use preventive services. The more specific you get though, the more careful you understand your data generating process. If you are looking at one very specific CPT code and one specific clinic, you have to be very careful to make sure that clinicians are using those codes reliably. If they have shortcuts or are figuring other things out. Or, are relying on other data systems to capture the data, you might be missing it.

We looked at the Emergency Room visits and urgent care stops and so forth. Like I said you want to figure out a way to build in data checks with your analysis. In my experience, almost all of the errors happen when you reshape data. Now reshaping data is just a term for taking different data sets. You can say well, I am collapsing the data. Let us say you were looking at patient events. You want to say I want to summarize that into number of visits per a six months period. You are collapsing the data. You could also say well I am merging it. I am going to take Medicare data. I am going to put it to that right of VA data. It is a person's record. You have variables for VA. Then there are variables for their Medicare use and merging data. If you are merging many to many accidentally, you could end up with huge problems. Transposing data, so this is where you take a data set. You reorient from landscape to portrait or vice versa. These are all common ways that errors enter the data. I find it much less common for people to crunch or crash their data when they are just creating new variables. Now, it does happen especially if there is messiness in your data. But just to keep in mind that you need to build in data check to make sure that you are doing it. The one variable that you could say well, let us just look at averages and make sure that averages hold. Sums actually work really well. You would still want to say something like what is the total number of days of care? Or, total number of visits? It is very hard to maintain that sum if you have errors in the way you have reshaped your class and so forth. Whereas you can imagine; if you averages you could keeping almost the same average with errors and putting errors into the data. Hopefully that helps.

For example, I am going to focus in aim 3 on patients who are 65 and older. I can throw out the patients who are under age 65. We kept those in at the data check. The data sources that we used for this study; I talked a little bit about the SE data, which is the VA data for outpatient care. The laboratory data which is the MCA data. We also were very interested in purchase care. For example, if you were a rural Veteran and you were told by this letter to get HIV tested; you might say well, I am going to go to my closest facility and get tested. I want to get tested quickly. Can I do so and just have the VA pay for it, and on the fee basis plan? Sometimes people say it is a non-VA purchase plan or purchase care. I will just call it fee basis in this talk.

Then we also pulled in Medicare. For those of you not familiar or have not worked with the Medicare files, there are a lot of files. VIReC is your friend here. There is also a group out there called ResDAC at the University of Minnesota that are very good at helping researchers understand the Medicare data. But I would say VIReC is probably your best source of information. Because they have a lot of information that is VA specific. In this study – and I have used these three files time and time again. You almost use information on Medpar. If you are not familiar with Medpar, it is an inpatient stay. It is the facility claim for inpatient's stay. If you go back to the 1960s when Medicare was created, we had this really weird system in the U.S. where facilities are paid by Medicare. Physicians are paid separately.

In that first payments of the facility is called Part A. the physician payment is called Part B. now, if you want the provider information and the Part B; which is important. It is 15 to 20 percent of cost and sometimes a little bit more. That is in the different file called the carrier file. That Part B is things that are both the inpatient side as well as the outpatient side. There is a separate outpatient file but it does not complete everything. You also need the carrier file. The outpatient file is outpatient care that was also provided in a facility. You can imagine going to your local hospital. Most people think of a hospital as this inpatient facility for the olden days. If you are old like me, it is this old hospital, a big building. You can get inpatient care there.

Nowadays you can get outpatient care there. That is exactly what happens at the VA in Palo Alto where I am. If you are at a facility like that and you are providing outpatient care, that code – that information is going to be captured in the outpatient file. That is a facility claim. That is important to know because there are different payments depending on whether you are in a big facility or just in a clinic's office. Data linkage methods, we are going to use scrambled Social Security Number. I am so old nowadays. They are coming out with all these other ones, SID and NICNs. I still just like the scrambled Social Security Number. It makes sense to me.

Dates of service are very important. I am showing you the tables once through the dates of service. Then in this case, we are using this station. Because we are using station with specific clinics, I am not so worried about whether it is sta3n or sta6a. You can imagine, if you are not familiar with those different codes. There are different ways to identifying a station. Here in Palo Alto, it is the Palo Alto healthcare system. That is station 640. Now there are satellite campuses. I am actually at the Menlo Park campus, which has this still 640. But if you were to look at the sta6a, it has a modifier there. In this case, there is no endoscopes provided at the My Menlo Park. We can double check that with the data. We are here. We were able to \_\_\_\_\_ [00:43:01] load the station information for that, both our controls and our sites.

There was a question about how do I organize the data? I do this thing. Someone typed into…. I was just reading the data. I have this routine where I like to pull the data sets allow. I do this typically. Imagine a clinical trial. This is a relatively small study. It works in this study, too. But imagine a clinical trial where you have 2,000 patients. Typically what is the first thing I do in pulling my data is I go across all of the data sets that I use. I start with the SE data. I might take the PCS data. I will take the Medicare data or the fee-basis data. I will say pull all of my patient records using the scrambled Social Security Number.

I often give a very wide time frame. If it was a clinical trial that started in 2004, I might say it is to give me everything from 2003 to the present. Those are the claims. I will put those in files. Then I will start cleaning the data in what I think of as their native form. I might say okay, then I will have a secondary data set that drops out variables that I am not interested in. Things about are there variables that I want to clean up and generate new variables? Then the third step is I will start to merge these files together. That is just – and if I need to back up a step, it is very easy for me to do so in that format. If I need to sort of dig down and say, okay, but why is that happening? I can always go back to the source data very easily in that format.

You need a little bit of care when you are working on a huge study. We are doing a study right now that is on specialty or access. One of our cohorts is 300,000. You would not want to say you have 300,000 patients and pull all of their SE data for it. Your local server or VINCI is going to freak out on you to pull every variable for a huge time period. You have to be a little bit more careful and think about the variables in your initial pull. You sort of skip that first step or are pulling right to a second step, if you will. In the analysis, I always start with unadjusted analysis. Difference and differences is very straightforward. You can look at mean effects. It should be very intuitive to think about what these might be and whether they are probability or sort of count of visits. There is a nice analog for moving for those unadjusted analyses into whether it is a logistic regression or into linear regression.

I often – it sounds crazy, but I will often even if it is a zero one variable, I will often use linear regression and multiple regression. The reason for that is that the interactions sometime get weird. There is a great paper by Ai & Norton that talks about interaction effects for nonlinear models including logistic regression. Many times, I will also just sort of start with unadjusted. I will move multivariate linear regression. Then I will get out at some sort of the fringes whether it is my GLM models or my other types of modes that I might need for the final analysis.

This was true here, too because we were looking at interaction. One of the questions that we had was this LSAE effect, was it different for people who are African-American or Black? About 50 percent of our sample was in that category. You can imagine the sort of biological plausibility if you will is that there is a less – there is greater concerns about trust in that community with the VA system. This might have a harder impact – a bigger impact, if you will on that community. I am going to just show you the results as we walk through it. These are all published. These are not unknown.

Here you get to see the control sites and the exposure sites. Now, all of these differences are statistically significant. This is a randomized trial where you exogenously place people into the cases and controls. All of these differences. We can control for all of them in the regression model. But they are significant. I am not trying to hide anything on you. It just means in my way of thinking is that one needs to be careful what you are controlling for. It would also have you pause if you were worried about the exogeneity of a large scale adverse event. This is one of those situations where if it was a situation or a choice variable; and you were saying well I am really interested in comparing smokers to nonsmokers.

You see all of these differences. You would say I am not sure we could ever control for the differences. There might be unobserved differences as well. You are right. There might be unobserved differences here that are affecting us. I do not think any facility is going out there and saying, hey, we are trying to have an error. Everybody is trying to be perfect without errors. You could imagine a situation that places – or people are not choosing their systems, or their care, or their providers. Because they think that there is going to be an error. Although, you could have the countervalent be true. Here is your unadjusted notification effect. Let me just walk you through this. The exposures that you remember. These are people that we do not have the exact list of people who were sent the letter.

You will see that on average, HCD in the control sites about seven percent of the population at any given time was getting this test done. HIV, that is about 2.3 percent of the time. These are tests that are done in normal clinical practice. We are not expecting those to be zero. In the pre and exposure sites, it is a little bit higher. It is a little bit higher, a little bit higher. But here, and look at this difference. Clearly, we have tapped what we believe is the exposure effect. A lot of these people were getting the HIV, HCV, HBV testing. Those were all statistically a clinically large and meaningful.

If you were to look at the difference and difference, you would say well the difference being is that about three-quarters of the population got tested in the year after notification. Now, you can go and do the timing of that. In the paper, we talked about who got tested in the first 30 days. Who got tested in the first 60 days? Clinically, and implementation wise, those might be really important. You might say we are not doing enough. We need a second notification to go out for people who did not get follow-up. You can do the \_\_\_\_\_ [00:49:39] of the adjusted notification effect. That was unadjusted. These are odds ratios; again and very similar things here. I actually find in this case, the unadjusted to be almost more telling because that magnitude is so large. But you see it is associated with very large odds ratios, too.

What is interesting here is when we talked about this, this subgroup effect is that it did seem to hit African-American and Blacks harder. It had a bigger effect on them. We are talking about a five percentage point difference in screenings for a follow-up testing. For HCV, about a 14 percentage point difference for HIV and about a 6.6 percentage difference, percentage point difference for HBV. You could imagine these situations where you would say the VA needs to figure out a way where it is the wording of the letter or better outreach to make sure that we are reaching out to people who are not responding. This could be one subpopulation. Sorry, I am running out of time. I will try to speed up a little bit.

The next question is going to be receipt of dental care. Remember this is the 18 month period. The slide that I am going to give you here. This is preventive care. You see that the effects are in odds ratios – are slightly bigger for preventative care than for restorative care. For things like root canals, they are not different. There is no difference. The other thing is you follow this out. The effects generally are significant in the first year for preventive care. In this case, we have got significance in the second period. But all of them rebound. My 18 month people are coming back to the VA. It clearly had an effect on their use of dental services. We do not believe that they are getting services elsewhere. But they came back to the VA after about 12 to 18 months.

Here is your – in the year following – and this is the idea of the third aim of switching providers. This is just those people over age 55. It is about 20,000 people. You get to see the effects. Generally speaking and what we have. There is a nuance here, which is they could be because of coding differences or not coding differences; they could be coming back from the HIV, HCV, and HBV testing, and the CPT code. We could be accidentally capturing that in the things like emergency care or outpatient medicine. They could be in there even though we tried to curve that out. You see some of these effect the difference.

One of the ones that I find very interesting intuitively although there is not a huge signal here is this one down here in outpatient surgery. It seemed to have a large effect on outpatient surgery especially the first quarter. A corresponding bump in Medicare use. You can post-hoc come up with explanations. Let me – these are exploratory in some sense about what is driving this. In many of these cases, these are things where it is elective surgery. Patients have choices, and time, and luxury of time. There is also the risk of exposure. Typically when you are getting surgery, you are getting some sort of procedure where there is blood exposure.

Here where it might be a situation and unlike emergency care or urgent care where you are thinking well, if I have a choice, maybe I will exercise the choice here. We coined this. I am not sure if it made it into the final paper, which was trust sensitive care. The reviewers might have told us to take that out. I cannot remember. But it is an interesting question about if we continued to follow surgery; and outpatient surgery specifically, would this be a way of identifying trust in a system? It seems that out here, at least exploratory, there is a corresponding offsets in the outpatient and inpatient – for outpatient and Medicare. Just to summarize and try to wrap up before I get kicked off here.

Receipt of notification – we saw a large effect, a 72 to 76 percentage point increase in testing. We seem to think that we caught all of the cases. We are not going to be 100 percent on the cases. That is going to be a limitation that we are just going to have to live with. Fifty-six percent were tested in the first 30 days. Think about yourself. You get this certified letter. How fast are you going to back to the facility? The facilities are all required to make extra staffing and availability for these folks to get testing. It is not a fact that you would think there are barriers to getting testing. The facilities are specifically told to make the testing available as quickly as the patient want it.

These are questions about why the patient is responding to that letter saying maybe in two months from now, I will get testing. Now, I might do that with my car. You might say well, if the engine light is on. I will say well, I will just keep driving it. The interesting question when you get a letter that says you might be exposed to HIV, HCV, or HBV. Now the vast majority of testing was done at VA. We saw almost no testing done at VA facilities; or, sorry fee-basis facilities or Medicare facilities.

What is interesting to me is that when they get these letters, people respond. They come back to VA for the testing. Despite that, there does seem to be these other offsetting unintended effects. The effects tend to be bigger for African-Americans. We see some of it switching. We some evidence for example, in the dental care where people are what I would suggest delaying their care; but eventually do come back for care. You cannot look at things like colonoscopy and endoscopes. Colonoscopy is an example. You might say well, let us look for colonoscopy like we did dental. Our people are not coming for colonoscopies. Well, the recommended non or low risk for colonoscopy is every ten years.

We just want to have that luxury of a ten year follow-up and say are they coming back for a colonoscopy? But I could also argue that based on these data most people would have forgotten ten years later or other issues would be much more relevant to ten years to determine where they are getting their colonoscopy. Then what happened ten years ago? But these data actually mirror what the literature shows. It is that when TV shows – and here I am going to show my age or with the TV show ER. When they showed things, people respond to that level of information. But people then, those effects dwindle quite quickly over time. The limitations and I will leave you with some limitations again. I said we do not have the exact limitations. We had that notification letters. You try to make sure we drop one of the sites if they \_\_\_\_\_ [00:56:15] feel comfortable with that site and identifying the cases in that site.

Then, even though we use this information on African-American and Blacks, the race data are problematic. My understanding is they getting much better and much more complete or non-missing in the CDW. But if you use sort of the historical data it is a particular challenge and then in the data. You have to understand the data generating process for all of these things and understand missing \_\_\_\_\_ [00:56:41]. The conclusion is really – there is two-fold. One is you can read the conclusions for this study. I think that the conclusions for you as good scientists are trying to figure out things about can you develop a good solid method for your data? Can you build in strong data checks to your data?

I cannot stress enough how important it is to understand the data generating process and sit down with the clinician at your VA site; and understand how they are doing things. If you are thinking like I want to check everybody who has got cancer, you would want to sit down with an oncologist here at \_\_\_\_\_ [00:57:15] site and say, so tell me about what happens when a patient comes in with the suspected cancer. How does that get coded? Or, for example, if you are interested in cardiovascular issues and a patient had a history of an MI. You are interested in MI in the future. You would want to see, okay, there is the diagnostic codes that are history and diagnostic codes. The clinician might be interested in getting through the patient CPRS EMR and be clicking that because of the history of that. They talked about are you eating well? Are you reducing your salt? How is your hypertension and so forth? You would see the diagnostic code for MI. But it is an historical thing and not an actual event.

You have to be very careful of that data generating process. If you have any information or questions for me specifically about this talk or data generating at all, you can reach out to me. If you have questions about HERC products, you have the Heath Economic Resources Center with the HERC at VA dot gov. Then anything about data in general, I will refer to my \_\_\_\_\_ [00:58:16] and close colleagues at VIReC. Because they tend to save me on almost all of the errors that I make. Because they straighten me out. Hopefully Linda, that is a good plug for you.

Unidentified Female: Alright, thank you Todd. It looks like we might have time for one or two questions. Molly, please let me know if we need to cut things off.

Unidentified Female: No. We are good on this end.

Unidentified Female: That is great. Todd, can you describe the structure of VA and Medicare data before you linked them?

Todd Wagner: How do you mean the structure? These are rectangular files. You have to make sure you have the capacity to link them. In this case, we are using information on scrambled Social Security Number to be our linking factor. But because our analytical files are summarized data. By that, I mean, we are looking at sort of ethics of time for cases and controls.

We are going to have a one record per case that says here is the patient's use of services in the VA in the past period of time. You would need to do the same amount of reshaping of the Medicare data but have in that same form before you merge it. Does that make sense?

Unidentified Female: Yes.

Todd Wagner: Now, I have seen other people do things where you stack the data. Then you collapse it in a uniform method. There are different ways to do it. I do not want to say my way is the only way. But that is generally how I would do it.

Unidentified Female: Alright. Okay, maybe one more question. Do you have any good resources for good data check methods?

Todd Wagner: VIReC – yes. I do not actually. I have bumps, and bruises, and old war wounds that have taught me not to trust even myself as a programmer about what you think has happened to the data. It is often things about merging, reshaping, and collapsing what you expect is happening with missing data. I do not.

Unidentified Female: Alright, and thank you so much, Todd. You can of course contact VIReC, if you have any questions or that we can get the answer. You can also contact Todd. His information is on the screen. Our next session will be presented by Denise Hynes. The first session is entitled, Decisions, Decisions, Decisions; Selecting Methods and Tools for Data Analysis. It is scheduled for Tuesday, December 22 at 2:00 p.m., Eastern. Molly, can I turn things over to you?

Unidentified Female: Excellent, and thank you so much, Hera. Thank you, Linda. Of course, thank you Todd for presenting for us. For our attendees, we really appreciate you joining us today. We look forward to seeing you at the next session that Hera just told you about. I am going to close the meeting up now. We do appreciate it, if you would wait just a second for the feedback form to come up on your screen. Fill that out; we do look very closely at your responses. It helps us to improve the sessions we have already provided. As well as it gives us ideas for new sessions to support. Thank you once again, everyone. This does conclude today's HSR&D Cyberseminar presentation.

[END OF TAPE]