Erin Farley: Good afternoon, everyone. We're gonna go ahead and get started. My name is Erin Farley. I'm one of JRSA's research associates. For those of you who may be less familiar with JRSA, it stands for the Justice Research and Statistics Association. We are a national non-profit organization dedicated to the use of research and analysis to inform criminal and juvenile justice decision making. We are comprised of a network of researchers and practitioners, which at the core include directors and staff from the State Statistical Analysis Centers.

Erin Farley: So it is my pleasure today to welcome you to our webinar, titled Tracking Defendants Across Databases. It will be presented by Matt Landon. Matt is a research analyst with the Washington SAC, where he studies issues affecting adult offenders at each stage of the criminal justice process. In addition, he earned his MS in Criminal Justice from the University of Cincinnati in 2015, and has also spent some time prior to the Washington SAC at the West Virginia Division of Justice and Community Services. So, welcome, Matt.

Erin Farley: Before we dive into the presentation, I wanted to first cover some logistical items, as well as thank the Bureau of Justice Statistics for helping to make this webinar possible. So we will be recording today's session. This webinar will be available on our website in the next couple of days.

Erin Farley: Today's webinar is being audio cast via the speakers on your computer, and also teleconference. So you can either use the phone number or your computer speakers to listen to the presentation today. If you are not automatically prompted with the audio conference information, you can go to the top of the menu bar, click on audio, then select audio conference. Once you do this, a window should appear where you can choose between calling in or joining the audio conference via your computer.

Erin Farley: Since everyone is muted upon entry, if you do have a question for the presenter, please click on the chat feature, which is on the right side of your screen. Then you can select to send a question to, which we prefer, to the host, the presenter, and the panelists. So that goes to everyone.

Erin Farley: This session is scheduled for one hour. At the end of the hour, we will have a quick poll for you all to complete. We would really appreciate you filling that out. It really helps us get feedback on what we're doing and how to continue improving. Also, we would like you to help us counting, help us count how many people are attending our webinars. We know that in some situations, there are multiple people watching at one computer. So if you are one of those people, what we would really appreciate is if you could, in the chat feature, write your name, whoever registered, and then the number of people including yourself, that are in the room. That just helps us, again, count how many people are attending these webinars.

Erin Farley: Great. Oh! And as a quick note, we do have another upcoming webinar. It is scheduled for June 14, and it is titled, From Report to Conviction: Understanding
Sexual Assault Case Attrition. It is going to be led by Melissa [Moribido 00:04:00], April [inaudible 00:04:02], and Linda Williams. Right now, there's no registration up for that, but I believe we will be getting that up in the next couple of days, or early next week. So keep an eye out for that. We have some other webinars that are also being planned out and finalized, for over the summer, so keep an eye out for those. There will be more coming.

Erin Farley: Okay, so I think that is all I have for you. So what I'm going to do is, I'm going to now pass the magic ball over to Matt and let him take it away. So, Matt, it's all yours.

Matt Landon: All right. Thanks, Erin. And wow! Hey, thanks all, for attending this afternoon. Really great to see so many people from all across the country. Always kinda blows me away. I've been an attendee in webinars before, and it is a fantastic feeling to see all of you here.

Matt Landon: So back when we were first thinking about doing a presentation, we in 2016 conducted a study. The name of that study was Jobs in Jail: Outcomes for Washington State Property Offenders. That study, as fascinating as it was, was rather routine as it goes for criminal justice studies. What I did not want to do was, end up giving a full presentation on just the study and its findings. Because while I find it fascinating, it's mostly relevant to the state of Washington, and probably something that a lot of you are very familiar with, either by reading the literature, conducting them yourself, and just your overall general familiarity with how criminal justice research works.

Matt Landon: But there is one particular piece that the Washington SAC had in the study that I find might be unique to us, perhaps a few other agencies out there, and is something that I at least want to highlight as a piece of criminal justice research that has been particularly useful to me in the last few years.

Matt Landon: To just being with this premise. Research in criminal justice very, very often ends up being restricted to a single database. What I mean by that is that you might be looking at, for example, a database from your Department of Corrections. Within that, perhaps just a database that pertains to prisons. When you're looking at that database, you're able to measure any variable inside that database against any other, but that's it. That's sort of your limit.

Matt Landon: If you have a question that goes beyond just that database, you end up in a little bit of a trap. Now, this can be completely fine if all of your questions exist inside of a single database, and I've included a couple examples here.

Matt Landon: How do the demographic traits of prisoners change over a given time period. You can look at ranges, averages, anything involved in that data set. You can ask how many offenders recidivate, but only as evidenced by a second entry in that same database. So we might be able to take a prison set and see how many
people return to prison, but we wouldn't be able to answer questions about arrests or perhaps repeat entries into jail. We could only really look at prison.

Matt Landon: So we ultimately often want to look at multiple databases. But then we run into a second problem. Sometimes these databases are linked for us and that's pretty handy, but that often only happens in the context of criminal justice, and then it's not always as great as we might hope. So we might have that prison data set. And then we have a question that relates to the courts.

Matt Landon: So we have a linking variable that links somebody who's sentenced to their ultimate prison sentence. That's great. Now we can look at things that happened during their time in court and during their time in prison. But we might still not have a linking variable to check out what happens in their arrest. Perhaps even more frustrating, we might not have a linking variable to find out what happens to them after they're released. We still can't get rearrest very well, just by linking courts and prisons.

Matt Landon: So one option that we can do in this case, is to examine each point of the question separately. So in example up here, what degree do judges choose a specific type of alternative sentence? And then, how did the overall recidivism rate change following that period? There we might have two different databases, two different measures, but unfortunately, not a very strong connection between our independent variable and our outcome measure.

Matt Landon: We just know that something changed at one point. We can measure the degree to which it changed, but we don't necessarily have a strong argument for how it was connected to that outcome. We can also ... I think one question that we get, do drug arrests drop after prisons begin offering treatment programs? Fantastic question. We can often track these. Sometimes that's a perfectly acceptable answer, to get a drop in drug arrests following that treatment program. But it is very hard to argue for causation much more than you can get correlation.

Matt Landon: But there may be another option in some of these cases. This is really what I'm driving at with this presentation is that, when you get enough personally identifiable information, or PII, available, there are ways in which you can take that, just the PII without any other research ID necessarily crafted for you to link it, and use the PII between two different databases to link offenders.

Matt Landon: With that information, once you're able to get people linked, you can begin focusing on more complex questions and a much broader array of variables. So for example, how often do those sentenced to jail find employment as compared to those in prison? Or maybe, what factors of a prison stay are associated with higher wages earned after release?

Matt Landon: Those question are not only richer, but we can actually start driving into individual characteristics, and perhaps their association with these things. You
can also begin looking at the interaction of various outcome variables on each other. For example, how someone’s employment, which might be housed in its own database, seems to alter their recidivism rates, and specifically the recidivism rates for rearrest. Or perhaps, for all the way through the system, a return to prison.

Matt Landon: That, I think, is an ideal. We want to have, as criminal justice researchers, a particularly rich set of data. But getting there can be the tricky part. So a little bit first about the study that is kind of underpinning this whole thing. I mentioned at the start, it was called Jobs In Jail. It was done in 2016 and it was based on some prior work with The Council of State Governments.

Matt Landon: In Washington, we previously took a look at our drug offenses and the amount of people who were going to prison under a drug sentence, and eventually restructured the drug sentencing grid in an attempt to better target and better address a lot of the issues that were coming up. There was a similar drive that came about with property offenses in this state. A lot of that stemmed back to the fact that our property crime rates were abnormally high. So a lot of questions started to swirl around what exactly was going to be done. The restructured property grid was a suggestion.

Matt Landon: At that point, this study started to get underway. Then that law did not go through. So we were in the midst of a study with no property offender grid, which was originally what we were going to lay the baseline for.

Matt Landon: So we, as researchers often need to, had to adapt. We switched over to saying, we’re not necessarily establishing the baselines for a property offender grid to be measured on later years. But we will still establish the baselines for property offenders and their outcomes, so that any future effort, no matter what it is, will be able to look back at 2016 and say, okay, here’s a study that gives us a look at offenders from about 2000 to 2015. Here’s what it looked like then as compared to, potentially in the future, what it might look like later. But to even begin to answer these questions, we needed to match a number of criminal justice system databases with some external information.

Matt Landon: So we took three primary data sources for this study. There was the Washington State Institute for Public Policies Criminal History Records. These get back into conviction. So this is just ... These people who are convicted don’t necessarily go to prison. They may end up in jail, or they might have community supervision, or an even lesser sentence. We also have our Department of Corrections offender files, these generally being people who went to prison.

Matt Landon: That is sort of our core database for this study. These were the main people we were focused on. Entry into this study as a property offender was graded on if you had gone to prison for a property offense at some point. That’s the point in the study where we started to track them. The reason for that was mostly because a lot of the mods were focused on this were aimed at reducing prison
rates for property offenses. So that was sort of the core group we were looking at. The people who are already getting prison offenses, what are they like?

Matt Landon: On the last data set that we wanted to include was the unemployment insurance database. That also contains wages, employment, and employer information. So that's not criminal justice related at all. Any linking data that we would normally have in there, that we could say, take a court ID and maybe the prisons track that same court ID, and we link them together. That's not in the unemployment insurance database at all.

Matt Landon: In fact, that database includes a lot of people who aren't even criminal justice involved. So having that as a resource is a huge thing if we're able to connect it, and in fact, we were. I wanted to first, talk about a couple selected findings, just really quick, to show the sort of things that came out of this. We were able to look at that first conviction for a property offense, to a re-conviction, that being in the Washington State Institute for Public Policy database. We ended up with a rate of about 68.7% of people who had at least one property conviction coming back and getting a second conviction. Those who were just in the Department of Corrections coming back to get any other conviction was about 59.3%. Going from the Department of Corrections and reentering prison again was a bit lower, but still at a 34% recidivism rate within this group.

Matt Landon: We also did a little bit of tracking for their average time to recidivism, giving a little bit of a scaler value to that measure. Then we also were able to look at their post release employment. A little under 60% of these people ended up employed post release, and they had about an average wage of $13.66 per hour. That is a mass simplification of a lot of the data. But giving just a quick overview, this is another one of the pieces that we were able to examine because we connected all three of these data sets.

Matt Landon: So here we have some bar graphs that show ... So, there's some different groups that we measured for having none of their time spent in DOC custody actually behind bars. They were always in some form of community program or supervision, being paroled, all the way up to their entire time being spent behind bars. We took a look at some differences in outcomes based on their recidivism or their employment.

Matt Landon: The different colored bars show you the three different data sets. We've got the recidivism by any conviction. That's the conviction data set. The recidivism by return to prison. That we could have just done with a single data set, no matching. And finally, their employment post release. That really relied on the wage data. So there's three different data sets worth of data, all matched by individual identifiers. This is made possible just by that. So this is just one of the sample finding that we were able to put out because of that.

Matt Landon: So, with that said, now I kind of want to delve into the meat and potatoes, The Match Method itself. So this is something that I want to frame in a way that is
Matt Landon: When we were working on our Match Method, we have a fantastic group over with us in our agency, called the Education, Research, and Data Center, or the ERDC. They're the real experts at doing this. They were a very integral part of this process. Because for our study, when we were proposing this to the IRB and getting approval, we're working with prisoners. I'm guessing many of you as well work with prisoners or other populations considered vulnerable by the IRB.

Matt Landon: Getting their personally identifiable information can be a big ask. Because that puts them at a higher level of risk in your study than just having de-identified information. One way that it is possible to mitigate that when working with an IRB, or at least in our case, was to say, "Okay, we're going to have an external source who just deals with their personally identifiable information, does the linkage, gets a code matched to each person's PII for each data set, and then sends that code back to us. We're the ones keeping the entire data as researchers. We won't necessarily know peoples' personal identification, but we now will have a code that allows us to link the data together."

Matt Landon: So that's sort of a two-party way of having the ... Nobody's viewing the entire picture. Somebody's able to just do the matching work and somebody's able to do the research work. It not being the same person can mitigate some of the risks when talking to the IRB.

Matt Landon: So for this study, the ERDC was working with the full name of the offender, which was contained in every data set we were using, their date of birth, and their Social Security number. With their full name, we've got their first, potentially their middle, and their last, the date of birth, and the Social Security number. So that's about five fields. In this study, we found pretty quickly that the Social Security number was not particularly strong. So we were able to match some records on it, but a lot of them ended up being matched without.

Matt Landon: With this initial match, what the ERDC folks did was, they took the data that we had in the DOC database, the [inaudible 00:20:13] database, and they already had the wage data uploaded to their ... They have a data warehouse that they keep this stuff in. So they already had a list of peoples' identifiers in there, and had that stable to begin with.

Matt Landon: In your case, if you're working with just databases and don't have any sort of pre-existing structure, the best thing that you can do is, pick the database that you have, that represents your core people in your study. What does it require in your study, to really be the target sample?
Matt Landon: So in our case, if we didn't have the data warehouse, we were just looking for the people who are in that Department of Corrections sample. We just wanted those property offenders. Eventually, we did narrow it down to them anyway.

Matt Landon: In your case, if that's what you're looking at, that's your fullest set. So you take the data set with the people that you really need for your study, and reduce it to their identifying traits. So perhaps we'll just go with their name, their date of birth, and their social security number. That's now your master set. You've got the list of everyone you need by their personally identifiable information. Then, take, if you're just working with one other data set, it's pretty simple. You just take that one and you compare it. If you're working with more than one, you're gonna need to repeat this process for each one, to try to get the match lined up for each case.

Matt Landon: For simplicity's sake, today I'm going to assume that we're just working with two databases. So you might take the PII from one database, you could say your master set. Then you take database number two. You want to run whatever you're using, SQL is good. You can run a lot of these codes in [inaudible 00:22:09], as well. Anything that you can do to run one set of these variables against the other, matching for a complete and perfect match. You will attempt to highlight and basically create a code for everything that does have a match.

Matt Landon: So, in this example, I've just created a randomly generated list of names, birth dates, and Social Security numbers. So, data set one is our complete set of everybody that we want. In this case, I've highlighted in green the ones that have successfully matched. The ones that are not highlighted did not get a match. So for each of these that is highlighted, what we were wanting to tell the program to do is, to create a linking ID for each of these that will tell us, Okay, Research ID 1 and Research ID 1001 in data set two are the same person. Then we can assign them any value that we want, as long as it's unique, for a linking ID. And that will remain their linking ID throughout the entire process.

Matt Landon: Now, once we've gotten this initial match, hopefully that is most of what you're dealing with, but sometimes it isn't. Once the initial matches have been made, you want to take everything that was a match and save it to its own data set. Everybody that matched, you can save and take their cross-link ID, or the ID that you created to link it up, and move it out. Once you've got that saved, delete them. They're out.

Matt Landon: That initial match is now done. So you've gotten a match and you don't necessarily want to take anyone who's been matched and continue matching them to other IDs. That can create a mess. So we want to reduce this list to just those that are left, that did not match. The next thing that we suggest is to wrinkle the data.

Matt Landon: Now, I went weeks thinking that this was an official term, but was informed just this last week that, no, in fact, it's made up. So I'm hoping that we push this far
enough that it becomes official because I like it. Wrinkling the data, as we term it, is making very small changes, very, very small, to account for little imperfections that can happen when crafting a database on the part of the people who maintain it.

Matt Landon: So that can end up being a wrong letter typed into a name. Or perhaps, the names end up out of order. Maybe the date of birth and the birth month end up switched. You might have one piece of your PII that is just exceptionally weak. In our case, it was a Social Security number. We had a lot of those that were corrupt, a lot of them that were suspiciously short, and a number of them that did not even exist.

Matt Landon: If you have a complete and perfect match in everything except one piece of PII, that might require a little bit of eyeballing. A number of these might require some judgment calls if you're able to go in and spot check your database, if it is of a size that, that's feasible. But for the most part, just doing some of these small switches will allow you to get a number of matches. It's really surprising how often and how many this can actually pull in for you.

Matt Landon: For this example of wrinkling the data, I've put some boxes around the data that, through just small little wrinkles, we're gonna be able to fix. So in that top line, there's an example of the names being switched. This will happen, particularly with names such as my own last name, Landon, that sound like they could be a first name. This will appear all the time, that you find somebody's name ends up switched in one database or the other, but the rest of their information is perfect.

Matt Landon: Just doing a wrinkle that switches these around will pop up a match. Just for those ones. Everyone else, it will ruin, but these ones, it will do it well. Those ones, once you find their match and you determine that it is, indeed, a good match, everything else looks like it lines up, you can give it its own ID, and like you did with the initial match, you remove it.

Matt Landon: With the Social Security number, in this case, it's blank. With weaker PII like this, sometimes you're able to just exclude it from the match criteria and see how many more you're able to get. If they all look like good matches, they look like close matches, then that's often an acceptable thing to do.

Matt Landon: The birthday. In this example, I've switched the date and the month. That last one has a wrong letter in there. For those cases, if you're suspecting that you might have some bad letters, or perhaps somebody has a last name such as Rodriguez-Gonzales, maybe there should have been a hyphen. For some reason, their name is Rodriguez-Gonzales in one database, and in the second one it's only Rodriguez, you can do a string code to search for just the first four or five letters, or perhaps only the last four or five letters of someone's name, and find out if that gets you your matches. It's very, very rare that you're going to find a false match that is that close.
Matt Landon: Now, I did put up a couple cases here that do not have red boxes on them. One of them is clearly a completely different person. That's gonna happen. You're probably gonna have people in some of your data sets who just don't line up at all. That's to be expected. They might actually be in one data set and never have entered the other for one reason or another.

Matt Landon: But then you're also going to have some of these weird ones. In one case, we've got Stephanie Riley, and in the other one, we have Stephanie Ridely. Could that be the same person? Maybe. The birthday is close, but it's not quite right. We just don't have the Social Security number on one of them to verify if that looks the same. My recommendation in these cases is to be conservative. I would much rather exclude a few cases than end up including false ones. Because that can create a much bigger headache down the road.

Matt Landon: So, once you have done that, you've gotten some of your close matches in. You get some more of those IDs linked. In general, I find it best to remain conservative. Once you've done enough wrinkling, you have no more close matches remaining, you want to go into that cross-link file, and check to see if any Research IDs align. You want to make sure the Research IDs are aligning with a single set of personally identifiable information. What you really don't want to find is, many Research IDs aligning to single sets of personally identifiable information.

Matt Landon: That is an indication of over merging, and you may want to backtrack and repeat some of those prior steps. Maybe look for some more conservative approaches. If you're finding that a single person in one database appears to be many, many people in the other, that also could be an indication that there just simply is not enough personally identifiable information to differentiate them.

Matt Landon: In the examples that I've been using here, it's unlikely. Using Social Security number, and their date of birth, and their name is a lot of information for things to be too similar. But perhaps without the Social Security number, it wouldn't be too unlikely. Perhaps if we have two John Smiths who share a birthday, they would be impossible to distinguish even though they genuinely are two separate people.

Erin Farley: Hey, Matt.

Matt Landon: Yeah?

Erin Farley: I just want to jump in. This might be beyond this conversation, but I was thinking about duplicates, and then you started talking about it. I was like, "Exactly." So I wanted to follow up real quickly and see if you had issues where, especially in Department of Corrections, or jail data, you have the same person that's in there multiple times, but it's not because it's an absolute duplicate, but it's because they were in there ... Maybe the unit of analysis is kind of muddled.
because you have people, but then people have multiple entrances into the system-

Matt Landon: Right.

Erin Farley: ... for different times they spent there. This might be going down the rabbit hole, but did you have any of those situations? And how might you have handled those?

Matt Landon: Right. Great question. This may, unfortunately, verge a bit on the abstract because I did not prepare, unfortunately, an example slide for this.

Matt Landon: But yeah, that's very common, that when you receive a data set like a prison set or a jail data set, you will have the same person's name entered maybe 50 times, maybe more, for every single action that happened inside of that system. Now, ours, fortunately for us, contain an ID that is assigned to that person when they enter the prison or the jail. So that every single action they take inside of that database is at least attributed to a single ID in that database.

Matt Landon: So we know that we only want that Research ID or that ID within, say, Department of Corrections database to match with one set of personally identifiable information in the wage records. So the ID that DOC has given them, we only ever want that to match once.

Matt Landon: So when we're preparing our files for cross link, what we'll do in advance is make sure that we reduce it to the point that each research ID and its associated identifiable information only appears one time. So that does require a little bit of work in advance, to prepare, and basically line it up so we're gonna attempt for a one to one match as best we can.

Matt Landon: In fact, later on, I actually have a brief example that shows that when you have this cross link, and when it's done that way, one to one, you can take information such as the person's sentence. Let's say that they were sentenced for theft. Then we can associate that record with each of the 50 records that appear in the Department of Corrections database, without multiplying out those records. We just do it by associating it with the Research ID.

Erin Farley: Mm-hmm (affirmative), thank you.

Matt Landon: That's somewhat of an answer. Sorry if it is a bit abstract.

Erin Farley: No! I think that, that definitely gets to it. In referencing the idea of going down the rabbit hole or opening up a can of worms, it's because I see exactly what you're saying. I think the process of getting there could be multiple avenues, depending on what your research question is, depending on what you're looking at, and sort of key decision making along the way to narrow it down to that one to one link.
Erin Farley: So I think, yeah. No, but you definitely answered my question, so thank you.

Matt Landon: Great.

Matt Landon: So in creating this cross link, once we've got those quality issues worked out and the matches are mainly one to one, we might even have a few errors left in this system, depending on how large your data sets are, that we're just unable to fully resolve. That can be okay. You just don't want it to be taking it over.

Matt Landon: We can create a cross link document. That's everything that you've been pulling out and saving throughout this process. We can use that document to connect the data sets. This really shows which research IDs match between the data sets. This is by the research ID level.

Matt Landon: So to your point, we may have a number of records in both data sets. Both data sets may have many records of the same person. But we might know that in each case, that set of personally identifiable information may be represented by that research ID linked to a specific other.

Matt Landon: So here's an example. In our, what we were going through, 10 and 15 didn't have a match, and they're in there with a missing number. We use -1. It won't confuse most programs, so [inaudible 00:34:35] or SQL will still recognize this as an integer, but none of your research IDs will actually be a -1, I hope. So this won't end up matching with anybody. But these others, you can match together.

Matt Landon: So you can see, for example, in data set one, we've got multiple records for these people, but only single records in data set two. We can associate those single records in data set two very easily with the ones in data set one, just by using the cross link to match the research IDs.

Matt Landon: Again, thank you all for bearing with me. I know that this stuff can sometimes be a little dry and dense, so I did try to include some colorful offenses for these people. I think the sort of person who is not smiling at a puppy might also be the sort of person who jaywalks, commits murder, and excessively honks their horn.

Matt Landon: So there's a sample here of what you might expect from a cross link.

Erin Farley: Oh, I just wanted to ... Sorry, just wanted to jump in real quickly. There were a couple questions.

Matt Landon: Sure.

Erin Farley: And one is, Did you control for shared Social Security numbers in the W2 data? And then the second question is, Are you able to control for non-W2 incomes? For example, day laborer, informal economy, et cetera.
Matt Landon: Great questions.

Matt Landon: So, for the question on controlling for the shared Social Security number. That is something that, specifically in the wage data, I don't know that we even quite got that far. Because we were having so much difficulty with the Social Security number in the prison system, mostly with people reporting false Social Security numbers, that when we had ... We started to get some matches on them, but I don't think we got very far past 20% before we scrapped that as a matching variable.

Matt Landon: So I would need to go back and check to see how the ERDC has been handling that within their master data management system because I do know that they have a full architecture built up. I would guess that, that's probably something that they have addressed.

Matt Landon: For the question relating to the non-W2 wages. That is something that we cannot account for, and was something that I needed to include as a limitation in the study, is for people who are maybe self employed, or they're doing some other work that doesn't fall under your average employer. Those wages we couldn't count. So the employment count and the wage count that we did include in that earlier slide, that is probably low. That is something we acknowledged in the paper.

Erin Farley: Great. Thank you.

Matt Landon: Absolutely. I'm happy to take questions throughout and more at the end. I am reaching towards the end of this piece on matching. I just wanted to talk about a couple things for troubleshooting.

Matt Landon: So, match methods, the stuff that I just discussed, they're not a good way to escape from a weak research ID. So when, speaking about that research ID before that might link the same person to multiple records in a single data set. If that's bad and we have a research ID associated with multiple people, or one person associate with a ton of different research IDs, that mess will follow you throughout your entire match method. Because the same set of personally identifiable information is going to appear like it belongs to four or five different people.

Matt Landon: Unfortunately, if there's a little bit of lacking internal validity in these data sets, this isn't a good method to go forward. The data sets really need to have some of that internal validity. Unfortunately, if they're lacking it, there's already some issues that will happen with any sort of analytics going on in that database to begin with. Because ultimately, you may begin double counting people or under counting people, depending on which way your problem runs.

Matt Landon: An indication of those sorts of issues coming up in the research ID can appear in the form of those many to one matches in the cross link file. That really does
mean that one person appears to match with five, ten, 20 other people in the other database. That can often be an indication of one person having a lot of different research IDs, so their personally identifiable information will match perfectly every time with a single set from the other. So that is something to look out for. Those many to one matches can be a real threat and can suggest some issues there.

Matt Landon: I did also, in talking with ERDC, want to include an example they suggested I talk about, and apparently this is more common than I realized. There are some odd quirks that you will pop up and that you may discover as you’re working with this. One that they talked about was, finding Mohammad Mohammad with the birthday of January 1. Apparently this is exceptionally common in areas with high immigration, partly because when people are first coming to the country, they might not be from a culture that actually regularly tracks their day of birth. They might know their year or they might know how old they are, but they might not know their actual birthday. So a common thing to say is January 1.

Matt Landon: So these cases can be exceptionally problematic because Mohammad Mohammad cannot be name flipped. Neither can January 1. It’s 1/1, and the same name, and the same name. So if you have a weak or an absent Social Security number and that sort of identifier, you might really be stuck in terms of getting a clear match, and also from differentiating them from each other.

Matt Landon: In these cases, although it’s somewhat painful to throw away data, your best bet may be to exclude some of these records, and as a researcher, to keep notes on any particular information that goes along with those records, to see if there is any slight skew that is being introduced to the database by removing some of these records.

Matt Landon: This, of course, is just one example. There are many, many others, and they particularly come up around common name sets. John Smith is always a classic example, but there are many of just very common names that seem to pop up frequently. They can muddle some of this work, trying to match with PII. I would be, someday, very fascinated to do an in depth study to find out just how that skews the data, if you ever do need to exclude high numbers of similar names.

Matt Landon: So that sort of covers the process, just the overhead process of matching. Like I said, I didn’t get deep into the weeds on how that might work with your code. That’s something that is specific to any particular program. But that gives you an overview on what sorts of things you’re looking out for, and the main steps of your process might be along the way.

Matt Landon: If any of you do have questions about how you might arrange code, or are interested in perhaps receiving an example, or having an example posted for SQL, SaaS, or something of the sort, we’d be happy to do that. But for now, we didn’t want to get deep into code, particularly if there’s other users of [inaudible]
00:42:03] or other programs, who might not be as familiar with whatever particular language we chose to go with today.

Matt Landon: But as a quick overview, just as a close on some of the benefits of the match data. A strong match going through this method really does allow for a comparison of a much wider array of variables, than might otherwise be possible, and sometimes to a more precise degree. We're able to look at the same person at multiple stages of this process, rather than look at separate stages of the process in isolation.

Matt Landon: So we might wonder, what is the relationship between the amount of time spent in community supervision and their later employment and earnings. Or how do employment and earnings themselves even relate to the later recidivism. Those questions are possible to answer once you've got data matched like this. You might still, as we always do, struggle to completely isolate causation, but you're gonna be a lot closer.

Matt Landon: This also is a particular boon to criminal justice data, because it allows us to track variables at each stage of the process. In criminal justice data, something that has always bothered me is how much variables really like to change between different stages of the process. Sometimes that's a completely legitimate piece of the data. It represents something that's accurate. Somebody had a plea deal and it changed. Other times, it's just recorded differently. This allows you to account for both.

Matt Landon: So we can wonder how many individuals were arrested on suspicion of a felony assault, but they ended up in jail on a lesser charge. Or maybe which traits among those charged with sex offenses are associated with a higher likelihood of participation and prison educational programs. Very specific question, but something that might be possible with a matched data set.

Matt Landon: So our study that we briefly just touched on, that's just one example. In my opinion, it's a very surface level example. There's a lot more that can be done with this. Just some suggestions that I'll toss out there, and I'm sure there's plenty more to be thought of. I think the best use for this often is, looking for external data, stuff that isn't necessarily in the criminal justice system, but that might still be of interest. So you might find external criminal justice records, such as doing a study where you're trying to see if any people from one particular state are offending in the state next door. And see what sorts of interstate, or perhaps checking federal records and seeing if there's any sort of connection between state crime and federal crime.

Matt Landon: You also might consider medical records. EMS data is sometimes a great source, a great alternate source for looking into some violent crimes. You might also find some [00:44:52] in medical records. Education data, as well, can be a great addition to anything working with the prison system, or education programs in prison. You might also be able to track offenders and see how their education,
whether they went to college, whether they graduated from high school, whether they had disciplinary records, how any of that pertains to their later involvement in the system.

Matt Landon: Or if there's any correlation, any association, or if they're able to get education at a later time. So any of those are just some ideas off the top of my head for where we might be able to take matched data sets. But really, there's an entire world out there. With that method, hopefully, all of you are able in some way, to expand just a little bit into another piece of data that enriches your research.

Matt Landon: So that's most of what I have to say. I am free and open, love to hear any questions or comments.

Erin Farley: Thank you, Matt. We will see if anybody has some questions. While we're waiting, we can also run the pole.

Erin Farley: Jason, if that's okay with you. I think you have the poll thing.

Erin Farley: But while we're waiting, I was wondering, is there anything, any work that Washington's doing right now that's continuing the work that you spoke about with this report, doing any analysis with these merged data sets?

Matt Landon: Yeah, absolutely. I kind of danced around this because I was trying to do less chattering about our studies. Once I get in there, I never stop.

Matt Landon: But one thing that we're doing now is a look at, actually, the education piece. We are re-running some of the data that we were doing on a study last year, that is trying to answer, using propensity score matching, whether people and to what degree they are benefiting from certain educational programs in prison. And looking at the difference between receiving a GED and perhaps receiving an Associate's Degree through prison programming. And trying to take that and look at their recidivism outcomes, and their later employment and their wages.

Matt Landon: So that takes their education records, which actually exist outside of the system. We have to be able to get the records from those colleges, connect them to their Department of Corrections records and their criminal history, because that's often a pertinent thing to include in a propensity score match. Then also links that to their wage and employment. As mentioned before, unfortunately we really can't look at self employment very well. So there is still that gap.

Matt Landon: Something in that study, and something that I'm interested in looking at, at perhaps a later day is, while we've got that education data linked, finding out if we're able to not only look at their later employment, but find out if they continued school later on. It would be fascinating.

Erin Farley: Great! Thank you. There is one more question. It is, Is the data wrinkling done manually, or with a matching or entity resolution algorithm?
Matt Landon: So that is with sort of the program, but it is a manual code. I believe our ERDC folks have a lot of code that they sort of have pre-written, having done the wrinkling so many times. So they sort of know which steps they're going to follow through. Then they put in the new names of the databases and the variables, and rerun it there.

Matt Landon: I don't know of any software that, in and of itself, necessarily is pre-made to do the wrinkling, but I'm certain that if this is something that you're gonna do regularly, and you kind of have the same process time after time, you can probably create a macro. Create it so that you can name the variables that you need and set it through the wrinkling process, so that you can get your standard thing done, and all those close matches out of there.

Matt Landon: I hope that answered the question.

Erin Farley: Thank you. Yes, we have one more question, too. What percent of data did you get as exact matches, and what final percent match after the wrinkling?

Matt Landon: That is a fantastic question. Unfortunately, I don't know the answer. Because in this particular case, and that's something I honestly should have asked. In this case, as I sort of referred to earlier, we had the ERDC run that entire match for us. Or for me, I should say. Because as the researcher, I wasn't supposed to be able to see any of the personally identifiable information of anyone in the study, so I never did.

Matt Landon: Because they're fantastic, I said, "Hey, guys. Could you please do this for me?", and they returned the linked data set. Then I said, "Okay, can you explain how you did that? So I can incorporate it into all of the explanation methodology of the study." I never did ask about the immediate match and the close match.

Matt Landon: I do know that, for the full match, it was somewhere in the 90's. We were, I think, about 90% matched. This was just the people that we had in prison. So we had about 90% of the people in the prison database match with people in the other databases.

Erin Farley: Okay.

Matt Landon: Yup.

Erin Farley: Great.

Matt Landon: So, sorry I don't have a more specific answer, there. That would have been a good one to know, wouldn't it.

Erin Farley: Well, I think that was our last question. So I want to take a moment to thank everyone for attending and completing the poll, and remind everyone that we
do have another upcoming webinar on the 14th. Not this 14th, I'm sorry, June 14. We should be opening registration for that very soon.

Erin Farley: Matt, I want to thank you so much for today's presentation. If people have a question for you, where can they find you?

Matt Landon: They can find me easily at my e-mail, which is just my name, Matt.Landon@ofm.law.gov. You can also contact us through our Sac's website. We've got a little contact page on there.

Erin Farley: Great.

Matt Landon: Those would be the best two. Again, to everybody who listened, thank you so much. Also, I do want to mention that if you are interested in reaching out or have questions, we would be happy to provide samples of code or answer more specific questions about the code process in this, that I didn't necessarily want to get into the weeds here.

Erin Farley: Oh, that's great. That's awesome.

Erin Farley: Also, another way to find you guys is, from our website, as well. If you go to our Sac map, and click on Washington, you can find some of Washington's information. That will lead you to your contact information as well. So there's an option number three.

Erin Farley: So, all right. Wonderful. Thank you all again for attending. Matt, thank you. Hope to see you all at the next webinar. Have a good afternoon.

Matt Landon: You as well.

Erin Farley: Take care.

Matt Landon: [inaudible 00:52:35] Six people on there.