

HRSA/MCHB 2007 FEDERAL/STATE PARTNERSHIP MEETING

Building Blocks for Promising Practice Models

October 14 - 17, 2007

Record Linkage 201: Vision for Data Integration to Action and Implementation

RUSSELL S. KIRBY: Okay. Now, there's a signup sheet we're passing around and one reason for that, it's not to take attendance. You might think we're doing that, but actually the reason for the signup sheet is, among other things, you'll have a better chance of getting the slide presentation if you put your email address on because I will email it to you that way.

I think you're going to have things on the website also, Scott, after the conference or—

SCOTT SNYDER: Yeah.

RUSSELL S. KIRBY: Yeah, there--it'll be available. Okay. So I have three objectives really for this presentation. One is to place Record Linkage in a broad framework, not just about Record Linkage itself, but within the context of planning analysis and public health action. Secondly, to focus on some of the key issues involved in doing that particularly with an emphasis on administrative public health databases. But really, the most important objective is this, to keep you

guys awake right after lunch and so, we'll see what we can do about that.

And we'll start out with this cartoon from the Birmingham News, which turns out that it's not just the highway department that has this problem in Alabama. Our community college system seems to have this problem, too, with legislators and their wives and cousins all having fulltime jobs where they never work and so on. But you never know, it could happen in your state, too, so you need to be aware of this problem, and it's a form of Record Linkage.

Okay. So what is Record Linkage? Now, this is a quote from Fellegi, and Fellegi and Sunter wrote a paper in 1969 called "The Theory of Record Linkage", which really is still the underlying statistical, theoretical basis for Record Linkage. And basically, what Fellegi said is that if we assume there is a single record as well as a file of records and all records relate to some entities, whatever they might be, then Record Linkage is the operation that, using the identifying information contained in the single record, seeks another record in the file referring to the same entity. So that's basically, in a nutshell, what Record Linkage is all about.

It has a long history. If we go by that particular definition, I'd like to say when I talk about birth defects around and said the bible was the first approach to birth defects around, but probably, also was a form of Record Linkage in there, too. But in public health, the modern methods that we have really only date back to the 196--I think Howard Newcombe might've written in his first paper in 1959, but

it's basically in the 1960's and with the advent of large mass storage and relatively efficient computers, desktop computers even, it's really been the last 15 years or so that Record Linkages become a major activity and something that's becoming an increasing focus in public health.

Now, Record Linkage should not be undertaken as an add-on to itself. I hope I don't sound too biblical here. I'm actually Unitarian so we draw from a variety of different text. But in any event, the whole point of this is that you shouldn't do Record Linkage just to do Record Linkage. There really needs to be some underlying purpose behind it. And ideally, it should be thought about in a broad informatics context, not only with the thought of what project you want to do, but how somebody else might be able to use the results after you're done.

And if you don't do that, then you're diminishing the potential impact of your work and making it so somebody else has to redo it potentially. But data quality is a paramount concern and needs to be a focus at all different steps in Record Linkage. And so, ideally, the Record Linkage should be done within a broad theoretical framework and hopefully with some kind of a research study design. So here's one theoretical framework, which--those of you who heard me talk before had probably seen this diagram. I like to show it in many of my lectures just because--mostly because nobody has seen it otherwise. It's published in 1990, but this is the—it started in Evans' framework for population health and what it basically posits is that within this, which is supposed to show up orange. I

don't know if it does here, but people get sick, they get healthcare and some of them actually escape from this box and they get better, they prosper and so on. And that's basically the traditional medical model, and it's also the model within which most of the data that we have in public health comes from. Okay.

The problem is that if we really want to understand the determinants of population, although in this case, what are the factors that influence the various maternal and child health behaviors and outcomes, we have to get outside the box. We have to think about genetics, the broadly defined physical environment, the social environment and then, of course, aspects of individual behavior and biology, and how those all interact.

And so we--as we think about the kind of data that we currently have, we need to realize that we don't actually have the data that we want, and we need to find ways that we can bring that kind of information into our framework wherever possible. So, as I just said, most of our databases are administrative data and most of them focus on aspects of disease processes or systems of care. And another one of our challenges is that--and I'm an ardently population-based person. And the--one of our challenges is that many of the data systems that we have are program-based, and some of our programs are population-based, but many of them aren't. And so when we think about, for example, working with a program like Medicaid or working with a program like Birth--we have to realize on some states, those might be close to universal, but in most states they're not.

And so, we always have to be thinking as we think about linking those records with other sources about some of the sources' variability in terms of who's actually a member of the different sets that we're trying to link and how they come together.

So and again, this is a diagram. I won't spend a lot of time on it, but a long time ago, I thought about what are all these different population-based data sources and how we might refit them all together, and this is what I came up with, and you will note that this particular diagram, which incidentally, also violates one of the rules of the top 10 list--the 10 best ways to do a bad scientific PowerPoint presentations because it doesn't show up the same on every version of PowerPoint that you use. So I apologize about that, but I didn't have time to clean it up.

But basically, what it basically posits is that you have vital statistics broadly defined as birth and fetal death records, some of those babies die, we link those records now for infants, but we don't typically link them for older children, and we really should be thinking about how to do that on a population basis. We have hospital discharge data. It's not really what we want. We'd like to have much more detailed clinical information about the context of each of the pregnancies and about the infant outcomes, but what we'd like to have is this. But at least we have hospital discharge data.

In some states, they actually do perinatal risk assessments. Florida has a program called Healthy Start where something like 85 to 90 percent of the women who give birth have an assessment, which can be linked to vital statistics, close to population-based. Then we have a variety of other programs. We have Birth Defects, we have High-risk Infant Follow-up. You see, this slides a little old. We now call this part C rather than part H.

We have a variety of screening and immunization programs, which really do strive to be population-based. And again, all of these need to be linked with the population that we're talking about. Then there's a variety of other programs in terms of looking more specifically at causes of fetal and infant and child death, which again need to be integrated with vital statistics. Some states have databases for clinical genetics, some states actually monitor some of the perinatal screening data as well. So you can think about how you can integrate all of those together, and of course, that leads to sitting around the table and coming to the decision to go ahead with the information matching.

So I expect to see some of you go back home and get dressed up on Halloween and look like this sitting around one of those meetings that Dr. Peterson was talking about this morning. Another kind of linkage that's becoming much more popular is linkages that create longitudinal data sets within some of the population-based data sources. So when you link—if you link all of the pregnancy outcomes for a particular mother, you can create a sibship file, which enables

you to look at the effect of a previous pregnancy outcome or health condition on subsequent birth outcomes. And there's a number of states that have put these files together. I personally think they should be something that we see in pretty much every state. I also happen to think that we should be creating a national birth index and actually have the ability to do these truly population-based rather than having to subset down by occurrence, residence, events within each state, which does have some generalizability issues. But then, why don't we link these records since we have all these data in computers. Wisconsin where I used to work, they had the whole birth certificate file with individual record data back to 1960, you could certainly match the records on the mother. They've done this in Illinois, Washington State and some other places, and you could certainly do that. And then that led me to thinking, "Oh, gee, wouldn't it be interesting if you had two women who were sisters and one of them, during her own gestation, had a mother who had no significant sequelae whereas in the other woman, she was affected by gestational diabetes, say. And what kinds of birth outcomes to those two sisters have, and to what extent does the in utero experience affect their outcomes?" That led me to thinking about that.

And, of course, you can do the same kind of linkage, but probably not this linkage, but this kind of linkage using other kinds of data sets. For example, hospital discharge records where you could link across successive pregnancies and it does imply some level of identifiers. Yeah. That's certainly an issue. But theoretically, it's possible to do that as well as it's theoretically possible to look at

(inaudible) of history experience of women who are on Medicaid for extended periods of time or eligible for WIC over extended periods of time. Then, I'm not going to talk a lot about this, but in terms of reproductive mortality, obviously, doing linkage is going to be a much more effective way of identifying reproductive deaths than that we get just by looking at birth certificates. And I'm still waiting-- this slide was dated 1996, and I'm still waiting to see somebody actually do this part. A lot of states now link death certificates for women of reproductive age with vital--the birth records, but actually also linking the hospital discharge survey, which would--there are many reproductive deaths that may not have a birth associated with them and you might miss still some of those.

Geo-coding is another form of record linkage and a very important form. If we go back to the diagram of the determinants of population health, frequently, geo-coding is what allows us to integrate data about the social and physical environment with our individual records, and so we need to be able to geo-code our records. And I personally feel that we should be geo-coding all of our public health databases to the extent that that's possible to do.

And then, finally, last diagram, this is the one that's much more theoretical. But I think that we should also be thinking not just about linking data around perinatal events, but also around pediatric events. Not a lot of states have developmental disability surveillance programs. There's a big overlap between developmental disabilities and birth defects, of course. We should be linking hospital discharge

survey records not just for pregnancies and infant events, but also linking them for older children, and likewise, cancer information as well. How many states have individual record level data on children with special health care needs? A few of you do, but not all. But this is a data set that also should be available and ideally, if you have these other kinds of data, you would be able to link them through having a database based on the certificate of live birth. And I call it the KMF that goes back to 1987 and I see Liz (inaudible) there. She probably remembers John Chapin, who was, at that time, the Assistant Administrator for the Wisconsin Division of Health. And we got together and start talking about how to link data. And, of course, I kept saying, "Well, you need to have a master file where you store the link IDs once you link the records." And so that's what this is. But this file needs to be more than a birth certificate file, because you need to have some mechanism wherever possible when you get information about people who are in-migrants to the region or out-migrants from the region to have some way of being able to quantified where they are at the particular point in time, so that you can actually know what the actual denominator is.

Then there's education, and this is one of the really challenging areas for us in public health. I personally believe that if we don't include education data in the things that we do in maternal and child health, we're missing a very large piece of the picture, and that was clearly brought out in the panel this morning talking about autism and related developmental disabilities. But again, how many states have individual level record databases for special education? There are some, I

know Florida does, but not very many states do. But these data, ideally, could be linked with vital statistics. You can look at a variety of perinatal risks. You can also look at educational outcomes, you know, this idea of no child gets behind or I'd like to call it no child gets in front. But the fact is that this should apply equally to children with special needs as it does to children who are mainstreamed in educational program. And we should be concerned about those kinds of outcomes. You can't really say, well, one's a health outcome and one's an educational outcome when you're dealing with such a complex situation. And we need to be thinking about how to measure those sorts of things in our work, as well.

And then, just to please everybody with the programs I might have left out, I made this diagram that also include your Medicaid and where can developmental disabilities, and so on, in terms of broadly about the range of programs that you might consider linking to at least the birth certificates.

So then, a famous 20th century philosopher, Elvis Pres--and last--the weekend before the last talks, I went to Memphis and we went to the Sun Studio, so I had a good dose of Elvis. But he sang, "You Don't Know What You've Got Until You Lose It." And that might be true, but when we're working with public health data, the fact is that you don't know what you have until you use it. And frequently, one of the biggest challenges that you have in working in doing record linkage is that there is an intense fear of the unknown on the part of program managers about

somebody coming in from the outside, even if they work in the same division, coming in from the outside and finding out all the things that are wrong with our data. But the fact is until you do, you'll never know. And I guarantee you that there hasn't been a public health program yet that didn't have some kinds of issues that could be corrected and improved by studying their data more carefully. So, that's very important. So, then, we have this question with record linkage, who, what, where, when, where, how. And you might ask, well, which of the questions is the most important. And I never could come up with an answer to that, so I decided to give you information about all of them, if that's okay. And we'll start with why. And in terms of why, the first question we need to ask ourselves is what is the purpose of the study? Why do we want to do a record linkage? Does it really make sense to do a record linkage? And although I'm a major advocate of record linkage, I am not an advocate that it is always the way to get the information that we need. Frequently, we can get the information we need just by calculating a numerator and a denominator from two different data sets. And when it's a congressman or a legislator asking for a statistic, they don't want you to spend three months developing a very complex algorithm and testing it and--they want to know now. And if you can give them an answer now, that might be satisfactory. And if you could then explain to them that a greater appropriation for your MCH program would be able to enhance the ability to provide even more detailed data, then maybe you do that. But you also have to ask yourself, "Can the linkage that you want to do be conducted in a manner that is going to create a database that might be useful in other ways?" Now, I'd like

to--the example of the State of Missouri, which has a maternally-linked pregnancy outcome file presently covering 1978 through 1997 and soon to be extended, I think, up to 2003 or 2004. And with that particular database, which they created, I think they had some very specific research questions they wanted to answer. But they have made the database available to academic researchers, and that is, over the past three to four years, has generated why to write, probably, is 30 or 40 different research papers in clinical and public health journals that have come from the fact that not only were they're willing to design their linkage in a broad way, but also make the data available. In fact, they didn't even charge us anything to get our copy that we have at UAB. So, there's--really be thinking about those kinds of things. And then, of course, is the record linkage technically feasible, and there are going to be some situations where a record linkage doesn't make sense, and you have to assess that and think that through. And then, of course, is it necessary? Then the question of how, where there's a number of different methods that are available between manual and automated approaches. And if you got all the individual level records, then maybe they aren't even in a database yet, manual approach might be an appropriate way to look at it. If you have a relatively small number of records in one of the two data sets that you're going to be matching, a manual approach might make sense.

I happened to be one of the PDIs for the autism project that was talked about this morning. And when I went down in July to the Center for Health Statistics with my list of about 65 cases that I needed to match for 2004, I had my list, and I

searched the birth certificate data base and did a quasi manual approach to that. There--it would have been overkill to do deterministic methods when I had 65 records. I was able to do the whole thing in about two hours doing it that way. So--but then, probabilistic methods are the other type of theoretical--from a theoretical point of view, the advantage of probabilistic methods is that they force you to do everything right in your whole process, and then it's easy when you use a deterministic approach to decide not to do the evaluation part and figure out, validate what you really have. But with probabilistic methods, you have to because it's inherited in the method. So, that's something to think about.

Now, obviously, there's this issue with identifiers. I personally think identifiers are kind of overblown as something that are essential in order to do record linkage. And that a case in point that I would give you, the first big project that I did, which I did in 1989, 1990, I had a file of two years of all the birth certificates for the State of Wisconsin to be identified. I had dates. I knew about the hospital of birth. I also knew which ones--I had the death certificates links, so I know which one has died. Then I had a file of--and I see discharges. Okay. And that particular file also had no identifiers, but it had some dates, and so on. But if you think about it, having a field like hospital in common between two different databases, and also having data birth in common between two databases, that gets you--if you think about binning your data across--the cross classification of those two different variables, for a hospital that has 500 births, there's an average one to two births per day. And just having those fields alone is going to get you to where you're

going to be fairly close, even without names and other more specific pieces of information. And even in a hospital that has a larger volume, I was able--a hospital that had 3,000 births per year, it was only around 10 on average per day, and I was able to bin that on, and I did actually use the deterministic approach for that. I didn't have any money to buy probabilistic software. The point is you don't necessarily have to have identifiers. And I'm going to make this point now, and I'll make it at least once or twice before we're finished, and that is you should never rely on a single identifier as the basis for making a link.

Okay. So, say you have--and of course, Alabama does have Social Security Number on its birth certificate, but we have a very forward-looking director of health statistics in our state who refuses to let anybody to use the Social Security Number, but a lot of states do have Social Security Numbers. But if you link a record just because the Social Security Number matches, and don't do any other evaluation to make sure that that's truly a match, you are setting yourself up for trouble. When I worked in an intercity hospital in Milwaukee, at least once every three months, we had a patient come in for delivery. Sometimes after she delivered, they've got some of the lab results back and they could tell--people's blood type doesn't change. These women had given somebody else's Social Security number or somebody else's Medicaid card as a basis for getting care. Well, that kind of stuff happens. It's hard to clean up the databases, so there are going to be errors. There's also key transcription errors and so on. So, be careful with that.

Then, again, in terms of software, should you buy a specialized program? Should you use a software package of your own? Should you develop your own? Well, I don't have a cut-and-dry answer to this. It's really going to depend on a variety of things. If you happen to be working in an agency that bought a copy of AutoMatch back in the 1980s, it still runs. They haven't really made any significant enhancements to it rather than making it run in Windows or Unix. Why not keep using it? You've already paid for it. You don't necessarily need to be doomed. But I would definitely be surprised to hear a state health department actually buying a license for AutoMatch right now. I think it costs about \$125,000. That's probably not a good investment of resources. So, you have to think through those things.

And then, just to make the point again and again, you have to evaluate the linkage results no matter what method you use. So, who should do the record linkage? Again, I don't know the answer to this, but there's a number of issues, the most important thing, of course, is to subject the staff to the personality profiles and make sure they actually have the right temperament for this kind of work. But they have to be willing to spend long periods of time alone, mumbling to themselves some--if they're like me, they have to switch to getting bifocals while they're in the process of doing record linkage, probably trifocals, soon. But--and should you have a dedicated linkage specialist in your agency, is it something that's more generalizably done within the staff? There are certainly

some pros and cons of that. I have heard of some states that have created a specific position for one staff person who does this kind of linkage. The problem you have there is don't let them leave. If you let them leave, then you can be in trouble.

And then, of course, the other issue of who is--or what data are you actually going to use for your linkage and how do you decide what constitutes a record that should be included and what shouldn't. Then we have the issue of what, again, not just in terms of the databases, but what are the functional relationships between the records in each of the databases. And when you get finished, are you actually going to have the data you need to answer the research question? And you need to think about that before you start, because sometimes it turns out that it actually doesn't work out the way that you thought. And then, again, what are you going to do when you're finished with all this? Are you going to burn your linked data onto a DVD and lock it in a file cabinet, and that's going to be the end of the process? Do you have some kind of a plan for more broad-based systematic data integration across different MCH programs and services? And then, of course, the issue of where should it be done in the health statistics agency, in the epidemiology agency, university, contract. I've actually heard states doing all of these, and I don't know what the answer is, but that's something that needs to be thought about, too. And then, of course, again, in terms of where, you don't forget about geocoding, as well, in term--that's another where issue of record linkage.

Then, how often should you do the record linkages? I have a number of views about this. I personally think if you're--if the data you're linking are infant death certificates with all of the emphasis on infant mortality prevention and FIMR and so on, I think those records should be linked to their birth certificates within two working days when they arrived in the Vital Statistics Agency. Not three months after the close of the calendar year, but right away, so that you can get the information to the people who need it, so they can do the important work that they're being paid to do. On the other hand, linking hospital discharges and birth certificates, that might make sense to do on a more periodic basis. Probably not annually, but probably quarterly, and there's some--there's different lags in terms of when data are ready in different data systems and you have to figure out exactly what's going to be the best in terms of that. And then in terms of other kinds of programs, what I call impulsive case finding, it's a sub-category of passive case ascertainment in which the program gets all of its data by automated record linkage across data sources. The State of Missouri for their birth defects registry, for example, links data from birth certificates, hospital discharge, special needs, NICU, pretty much anywhere they can think of that might have ICD codes for birth defects. But none of the records have actually been reported to a registry, and they're all coming in from other kinds of data systems that happen to carry that information. And that, again, depends on what the needs of the registry is as to how frequently that should be done.

So, just a little digression here, we'll take a look at some of the perspectives that experts on record linkage have--had to offer on how we should do this in the form of a top 10 list of how to do bad public health record linkage in keeping with the philosophy that you can often illustrate the Best Practices by saying what not to do. So, an anonymous correspondent told me that you should just have somebody else do the linkage and then use the "don't ask, don't tell" method perfected by the military. In that way, what you don't know doesn't hurt you. Okay. And maybe, I hope we don't actually do that, but you never know.

Okay. So, firstly, always trust the social security number in the database as the correct social security number for that individual. And if there are duplicate social security numbers for obviously different individuals based on age, gender, race, ethnicity, whatever it might be, then randomly select one, and if you don't know how to do that, use the latest state lottery to obtain the random numbers. That work for the people who won. They should work for you, too. Okay. And then, of course, change the linkage identifier every time you recreate the data set. And this will keep your data users guessing plus they won't be able to refer to specific records and that will ensure confidentiality. Yeah. I hate to tell you guys this, but I didn't make this one up. I was working actually with a city health agency that got a monthly download from the state vital statistics and they--and the state would not give them their birth certificate number and each record had a new number each month and they could never figure out--and there were records in the next month's file that were also in the previous one that had corrections and so on.

And they could never make any sense out of the data. So, this is a problem.

Okay. Of course, it doesn't matter if you get twins match correctly across files since they're identical. Anyway, and if they share a genotype then they should also share a link. Okay. And, of course, Henry Ford had the same idea with his Model-T. You can have any color you want as long as it's black. So, if a variable is listed in a data dictionary, it's safe to assume you can use it for linking because after all, it's always been collected in exactly the same way for the whole time period and whole study area. Okay. And this is especially useful to believe if you're trying to use variables like race, ethnicity or--did you know that adolescent mothers might actually get additional school between pregnancies? I mean, maybe not, but it could happen. Or that ICD--and we may someday start to use ICD-10-CM and we're going to be the last country in the world to start using it, but it could happen. And all these things are subject to change.

So, always assume that they haven't and you'll do bad public health record linkage. And then, of course, you always want to strive to have an out rhythm that over matches, because high-matched percentages are impressive. And if you get over 100 percent, you're done. If you can get there, you're done. And, of course, if you are over matching, there's no need to ever look at the data or evaluate the process. Right. Okay. But then, of course, checking for duplicate records just slows down the process. So, this is a step that should be eliminated.

So, instead, just take a look at your log file, oh, and another thing, of course, I don't know if I have this on the list here, but one of the ways to a bad public health record linkage is to never look at the statistical log file. But if you just happen to do it, just check and see that you have the same number of records in the output dataset that you had in the largest input file and that will prove that everything is matched, right? Okay. And then you go ahead and do your analysis. And of course, don't bother to check for name changes. It really doesn't happen often enough to change statistics. And that's, of course, especially true for women, for children who are adopted or in foster care or, of course, the rare family that speaks Spanish or some other language or comes from a culture like China, where surnames are listed first. And my favorite one, I just learned about this and I haven't figured out quite how to deal with it, but one of my colleagues just had their first child but he was from India. And so, in India, the way they work the names, they reverse--the son, they basically give the son a new first name and his surname is his father's first name. Is that--that's how--is that universal in India or is it mostly--so, how do you deal with that in terms of record linkage? Again, some communities have larger Indian ethnic groups than others. But it's certainly something to think about as well.

And of course, there's only one valid and reliable linkage spreadsheet, the one that you develop yourself. And firstly, you should definitely pattern it, because it is unique and it's an intellectual property and so on. I actually heard of state that was thinking of doing that. And then you should never test or evaluate the

strategy and, of course, never actually let anybody else see the computer algorithm. That would be even worse than sharing recipes to do that. Yeah. And then of course, deterministic linkage must be correct, because, after all, it's based on exact matches. So, why settle for a complicated, time consuming probabilistic matching when you can be certain.

So, and then of course, finally, you should spend months of time discussing in meeting after meeting after meeting, join a few conference calls just for a variety to do record, maybe even a webinar, to do record linkages. And you want to make sure that you include the staff attorneys, the department HIPAA privacy consultant, all of the division directors for any of the datasets that you're going to work with. And then you said, when you're done with all these meetings, you should assume, of course, that the linkage can be done in a couple of weeks and that once completed next year, it can be just run as overnight job. In that way, you will definitely have bad public health record linkage.

So, let's go back into the subject matter in a little bit more detailed. Okay. All right. Yeah. I actually heard of a story about this. Apparently, Descartes was in a bar and it was around closing time. And the bartender says, "Hey, Descartes, do you have another?" And Descartes looks at him and says, "I think not." And he disappeared. Yeah. My kids didn't get that one, but hopefully you guys do. But anyway, you really need to define the nature of the problem, be very clear about what the purpose is for the underlying reason why you want to do the linkage,

and what the records and each of the datasets represent. Doing some Venn diagrams to think through what the starts our relationships between the records might be, thinking about potential mismatches that might be present in terms of the catchment areas for the different datasets, reasons why there might potentially be records in one dataset that can't possibly match in the other dataset or vice versa. We're thinking for all those kinds of things and laying that out in a fairly clear manner before you get started. And then of course, thinking about what are you going to do with the results? So, here's a handy primer about why to link. So, here we have--you can select your best answer that--it might be a different answer for different problems. One question would be we can't answer the research or policy question without linking this data. Another might be we have to, under the terms of our SSTI grant, or maybe under the terms of our CDC cooperative agreement or whatever. And another might be integrating record linkage into routine data management process enables us to asses the programs effectiveness and efficiency on a continual basis. If this is your answer, we need to talk, because--and I know that it is a potential issue in some states that so much time spent linking the data and then when it's all finished, they turn around and start doing it next year, and nobody ever does anything with the data.

And that's not what this is all about. If we're trying to improve the newborn screening program in our state and we never actually take our dataset down to the newborn screening director and say, "Give me three or four things that you'd like to ask now that we have these data and we'll analyze the data and give you

some suggestion.” You need to do that, because it’s not just about record linkage. So, then we have the other side of the question, why not link? So, lack of funding? Okay, maybe. You can always--I’m in Alabama and it’s a little difficult to send my students to your state, but we can figure out something. That wouldn’t cost that much. The staff don’t have training. Well, that’s why we come to these kinds of sessions or maybe more comprehensive sessions. We don’t have the hardware, software or data storage available, maybe. But, again, that’s not necessarily something that we can’t overcome. Bureaucratic inertia, that’s certainly an issue. Turf battles between programs. How many of you have gotten involved with the record linkage project where you had issues with one or the other of the programs? A lot of people are shaking their heads, putting up their hands.

This is a major challenge in terms of this. The question doesn’t warrant linkage. That’s possible, definitely. Some of the above, all of the above, could be any number of different issues. And of course, here’s the business, the bureaucratic flowchart for why it’s hard to do record linkage. Yeah. Let’s see, I think, my office is right about here.

Okay. So, some of the--again, the first steps that you need to think about before you conduct a record linkage. And my first step is to see--we’re doing fine. Okay. Carefully consider the context in terms of informatics, the program and research issues. It’s--on one level, it’s fine for the analyst just to look at data dictionaries,

but it really helps to go and talk to the program staff and find out what the key issues that they're concerned about are or maybe find out what the key issues are that they should be concerned about that they have never been able to look at before, whatever those kinds of things are, because they will help you to think through more clearly how to do the linkage. And be thinking about who's going to use the data and what kind of users might--to which it might be put. And so, as a hand in terms of this, if you only ask the people on your team about the potential users, the users identified are going to be within the same frame of reference you already have.

And a lot of times, interesting ideas come about from people who'd come from completely outside of the box and (inaudible) a lot of our--public health breakthroughs are coming from those kinds of approaches and you won't get that if you don't find a way to involve individuals who can look at it from a clean and fresh perspective. So, I just put a few slides together just looking at how you might explore the issue of linking birth certificates and Medicaid pregnancy claims data. And I want the prefaces by saying, "I have never actually done this kind of a linkage myself but I was thinking about from an informatics point of view of what kinds of steps you might need to go through." There might be some in the audience who have work with this a little bit more closely. So, what do the records represent? Well, you might have a Medicaid claims database. It might include women who are pregnant or women who recently had a baby. It might include women who are not. And potentially, women who may be elderly and

probably not at particularly high risk of that, although, nowadays, a lot of strange things are possible. There could be men. There could be infants and children included in that database as well. But you also have to think about in order for somebody to be in the claims database, firstly, they have to be enrolled in Medicaid. Okay. So, how do they get to be enrolled and what is the relationship between the population was eligible and those who actually are enrolled. And then, in addition to that, in order to be in this claims database, there need to actually be a--but, again, it depends on what level of Medicaid data you can get your hands on. But most of the files that people in public health get are paid claims databases. So the ones that failed to get paid, they're not in there. So, you're not going to have that information either. So, those are some of the things you have to think about. Then, of course, with birth certificates you've got potentially live births and--well, fetal deaths are separate kind of certificate, but those are now that you should certainly be thinking about.

Then, again, what records might be included? And this is a big challenge. In some states, the vast majority of women on Medicaid are in managed care programs. And some states actually have programs where they generate and count their records that are, kind of, like individual claims for service. But in other states, they just have a global bill, which summarizes the overall set of services that were provided. And how do you analyze those records in relation to the fee for service claim after claim after claim? What are some of the issues with that? Do you want to just decide, "Well, I'm not going to look at the ones that are

managed care?" Well, there could be some systematic bias in terms of that. So you have to really think carefully about this.

And then there's also the possibility that there could be records in the Medicaid database that don't represent prenatal services. And it's, kind of, an interesting thing, but in almost every state, there's a large number of women whose labor and delivery services are paid for by Medicaid retroactively, who didn't actually get any prenatal services at all. So, there's not going to be any claims for them. And if that's what your interest in studying, that group is going to completely subset out of your analysis. So you have to think about the implications of that.

Then there's the possibility that one particular--might have multiple records. And she might actually have multiple records for a lot of peculiar reasons, some representing inefficiencies in Medicaid eligibility. She could, for example--I hope it doesn't happen--with presumptive eligibility it shouldn't. But she might have been on Medicaid early in the pregnancy, and then for some reason, it shouldn't have happened, been terminated and then became re-eligible again, has a new number, and you have no way of linking those, necessarily those two pieces up.

Anyway, I'm just throwing out that there's a lot of complexities that you have to think about in terms of what these data are before you think about how you can link, because they'll have implications for the broad linkage strategy that you need to put together.

And then, of course, why are you doing these linkages and other question to think about. And some of the answers to that question might have to do with the focus of the study. Are you interested in looking just at maternal outcomes or on live birth outcomes associated with the prenatal care services that were provided? Are you interested in focusing on the kinds of health services that the infant receives based on the birth outcomes? Are you interested in looking at mother-infant dyads and potentially linking this up? A lot of states have enhanced prenatal care services. Some states have enhanced infant or family benefits. Are you interested in pulling that into this as well? And, again, that has implications.

Then there's issue of residence and occurrence and the possibility exists that there could be services that you don't have good access to. There could actually be the situation where both the birth itself and much of the health care occurred in another state even though the person is a resident of your state. And what are the implications of that for your getting the data that you need in order to include them in the database? And again, there could be some systematic patterns in terms of who--for whom that happens. And when I worked in Arkansas that was a major issue. The eastern third of the state of Arkansas, virtually, all of the high-risk perinatal services went to Memphis. And in case you don't remember there's a thing called the Mississippi River, and Memphis is in Tennessee rather than in Arkansas. And that was a major issue in terms of--Tennessee had a different birth certificate file, the records didn't come electronically. I don't know what was going on with the Medicaid, but there's all sorts of issues.

Then, of course, there's the issue of Medicaid eligibility versus utilization. And, again, I pose this particular problem in terms of thinking about claims data, which implies utilization, but we still have to be thinking back to--anytime we do a study, we want to think about the reference population. And so, do the subjects that we have data on in terms of actual claims, what do they look like in terms of the universe of women who are eligible for Medicaid? And, again, are there systematic biases and so on.

And then, do we have our priority expectations as to what types of records will and will not match? And how have we built that into our algorithm so that we can try to address them, so, again, some possible purposes of the linkage. I mean, I don't know what these would be, but there could be--in your state you could be looking at this, and again, I think I've actually gone through most of these already. I don't think I talked about items four or five, but these are definitely things to think about also. Maybe you're just interested in linking with Medicaid so that you could have another proxy measure--socioeconomic status. You're either on Medicaid or you're not on Medicaid. And maybe you don't believe that the checkbox on your birth certificate for that is accurate, so you want to actually match the records with the Medicaid, it could be.

Also, pregnancy episodes of care. One of the things that--I've been waiting a long time for somebody to get beyond the Kessner Index, the Kotelchuck Index,

the Greg Alexander's RG Index. They're all based on numbers of visits and when prenatal care started. And actually looked at what kinds of services that women received and whether they are getting quality prenatal care continuously through the pregnancy. And at least, theoretically, with the Medicaid claims data, which are coded CPT codes, you can get down to some more specific levels in terms what kinds of services, and you certainly have dates of service so you can look a little bit more specifically at the timing of care as well. And to do that, it's going to require that before you even link the records to your other dataset you need to do it within data set linkage. Because there are potentially going to be one to 50 separate records for--or potentially even more than that, depending on how they're billing for each of the different services, if they're billing separately for labs, billing separately for ultrasounds or whatever. There could be a lot of records for a particular woman. So you've got to think about how to put that altogether into an episode of care before you can link it to the vital statistics. And, of course, there could be some other things that you could be doing as well.

So, a little bit more about residence--do I need to talk more about this? The issue here is that, for vital statistics, there are exchange agreements for records and it works fairly well. I think they might have even finally figured out how to make this an electronic exchange. Although, I'm not completely sure how that's working.

JOHN SENNER: Not yet.

RUSSELL S. KIRBY: They haven't yet? Yeah.

JOHN SENNER: It keeps coming, but not yet.

RUSSELL S. KIRBY: Okay. Well, John and I--that's John Senner from Arkansas. We used to work together, and in 1989, I went to what was not called NAPHSIS, before they've changed the name in NAPHSIS, and at the meeting, I thought, "We really need to put together a mechanism for coming up with an electronic exchange." So it's only 18 years. It'll happen eventually. It's certainly feasible, anyway. I think they actually are getting close to figuring it out now. But Medicaid is another potential issue and I don't know how this all works with Medicaid, but it's also a potential issue with some of our other data sets, our WIC data sets, our newborn screening, thinking about how to exchange data so that since our program has--really interested in residents of our state, how do we meet and make this all match up. And it's definitely a challenge to work through. And then, of course, when the records fail to match and you potentially have this outside of the study area, kind of, issue then could it be not necessarily that you're unable to link, but due to some other issues in terms of reporting requirements or eligibility that might be different between the two jurisdictions. Okay. Then of course, eligibility versus utilization, and I don't know if I need to belabor this, but the issue is that for many of our programs, you establish that you're eligible and then you enroll, and some portion of the individuals who enroll in a particular program never actually receive any services.

And when I worked in Milwaukee I did a study of the prenatal care coordination program. And at our hospital, which was the biggest site in the entire state, 40 percent of the individuals who were enrolled in the prenatal care coordination benefited Medicaid never had an additional service after the evaluation that the term--that they were eligible. But of course, in Madison, they thought that they had thousands of women who were getting this benefit. So, again, you have to think through those kinds of issues and how they might affect your data. And I would caution that you shouldn't do your analysis just based on eligibility that you really need to find out whether the individual you're studying have actually received some kind of service. Then when these things, when you get down to it thinking about matching some of the pregnancies that involve a woman who's got, say, got a Medicaid pregnancy claim doesn't result in a live birth, it could be a fetal death and hopefully you can--now we do know, by the way, that there's the--consistently in all of the studies that I've seen, five to fifteen percent of fetal deaths are under--are not reported in vital records, and then of course, spontaneous and induced abortions on most states have induced abortion reporting systems that are so be identified that you would never be able to link those records with anything else and for a variety of good reasons. And then, of course, some of the women in the Medicaid data set might not have been residents of the study area at the time of the vital event, and people do move so there'd be another potential issue. And then of course, over reliance on unique

identifiers can lead both to mismatched and unmatched records. So you have to be careful about that as well. Okay.

So just--what I'm trying to do here is in an hour and a half give you the highlights of about a two-day training seminar. So I've got a little bit--I'm--not any computer programming or simulations or anything, but I—what I wanted to do is just briefly go through some of the issues in how you do deterministic and how you do probabilistic linkage and then I'll conclude.

So, first of all, in terms of deterministic linkage, first thing you need to do is think about what variables are common to both of the data sets. And my example is going to be based on SAS so if you happen to use some other program, just substitute whatever the relevant would be. But first thing is to do prod contents and look at the variables. And I did want to make this point here. And that is that if you're maintaining all your data in Microsoft Access and you think that you can link the record just by doing a direct join on variables in data set A and data set B, don't do that. There's so much detailed data cleaning that you need to do that you're--