>> So, anyway, what we're going to do for this afternoon is we're going to start looking at the files and we're going to start with the--what used to be called the Denominator Files, currently called the Master Beneficiary Summary File, but basically, it's the file that tells who's in our study, who's in the Medicare program.

So back in the Stone Age when we first started using these data, there were two files that CMS created especially for researchers. And when I say they created them for researchers, what I mean is they actually compiled them and reformulated them in a way that would make them particularly useful and easy to use for researchers, and that one of them was the Denominator. It was a single file and it was 80 columns wide. And back in the early '90s, columns mattered, and adding extra columns to a data file was a lot of work. And we'll talk about that later 'cause there are some historical things that the Medicare program did that--so the current computing technology seems absolutely crazy but at that time, it was really hard. And--but this was 80 columns wide, and so what was nice about it when people would be new to the Medicare program would say, "Well, you can see the data" 'cause there's this thing of--when you're transitioning from having collected the data yourself with all of the pain of having abstracted the charts and paper forms to like getting something, people got very nervous about it. So where is the data? And you're just like, "Okay." And you could pull the denominator up in an editor, a text editor, and it was beautiful, and you could say, "See, that column right there, that's what it means." And you can't for any--you couldn't for any other file and frankly, you still really can't.

So everybody loved the Denominator File 'cause it was really sort of--you could figure it out. Well, of course, we can't anymore. So what's happened is what used to be the Denominator File, which was the name like a proper name, is now a class of files. We've got the Beneficiary Summary File, the CMS Denominator, the Part D Denominator, the SEER Medicare program has denominators. And what--and so what we've done is we've changed the names of them all and they each differ a little bit in terms of the exact content they have but what in fact they are is their--so what we need to think of more generically now is this idea that denominators are about the basic enrollment information and demographic information you need to do to start building your cohort, your panel study, your cross-sectional study, or whatever study design you have.

And I think, and I would encourage you to really think this way because I have every expectation that denominators and what the files contain will continue to evolve. They will continue to get renamed. And so if we get really caught up in the name of the file, we're going to sort of miss the more general point.

And we're now going to focus mostly on the Master Beneficiary Summary File which has four segments to it. So it's--again, from this simple, pull it up on a text editor file, you have this growing complexity. So there are four pieces to the Master Beneficiary Summary File Denominator. There is the basic summary which has enrollment information which is what we care about for the most--minimum that we care about. There is a segment on

chronic conditions which include flags that have been created for us about whether or not the person has evidence in their claims that they have one or more chronic conditions. There's a cost in utilization segment that's a summary of how much was spent on the person and what types of utilization they had during that year. It's very high-level, very summarized, but you could use it to find out how much was spent, I mean, if all you cared about was total spent on hospitalizations and you didn't care what they were for, this would be a great way to get it.

And finally, there's the NDI death information segment. And as we said, CMS has gone back and forth on that in terms or when and for which years they provide NDI, National Death Index, death matches. So if having cause of death is essential to your study, it's really important that you call ResDAC and talk to them and make sure that your years of your study match the availability of this cause of death information. It has changed over time in part with finances and I expect that there will be continued movement back and forth in part as we work on data sharing arrangements between the National Death Index and CMS and between states and the Feds and so on. This is one that's going to move back and forth for a while.

But I'm going to really focus primarily on the Beneficiary Summary segment because that is by far the most generic and I would probably argue the most essential. So as I said earlier, you know, when we think about denominators, we could get really caught up in a specific file and say, "What is in column 17?" But I think that's not going to help because they keep changing.

So what I want to do is talk about the sort of this broad conceptual understanding. And when you think about denominators, what you really are looking at is patient identification, looking at demographic information, and you're looking at eligibility to receive certain types of services. That's really what it's about, because I--as we--remember, when we said everyone in the denominator needs to be eligible to be in the numerator, it is--these are the fields that help you figure that out. So if--we recommend that your demographic information for your study be pulled from the Denominator File. That would be date of birth or age, date of death, sex, and race. It's the easiest, it's the most consistent, and it means that your file won't be dependent on somebody having a certain type of care and won't be dependent on anything outside of your control like when things were processed or anything else. So this is the place where you should be pulling this basic demographic information.

So when we talk about the Denominator File, the first question you should be asking is, "Well, who's in it? When you say it's a Denominator File, what do you mean by that? You, guys, were all thinking that, right? Okay. So what it--denominators are annual files. So every year, there is a Denominator File created then a summary file, whatever we call it, and it contains one record for each person who's eligible to be covered by the Medicare program for even one day, okay? So if somebody joins Medicare in February 1st, and dies in February 2nd, so it's two days technically, they will be in that file, and they will be there just as much as somebody who is in the Medicare program for the entire year. So the file is not

limited to users, so there will be people in that file who never used a single service and there will be people who were getting frequent flyer points from their doctors 'cause they were in so often. You will have everything in there. But this is really--these are the--this is the list of people that CMS says, "They are eligible to be in the program. If their bills come through, we will deal with them." Okay. Managed care, too.

Eligibility, as I believe Marshall told you this morning, is determined by the Social Security Administration. So CMS does not actually determine eligibility. Social Security Administration does and then transmits that eligibility information to CMS. The basic Denominator File will contain all benefit groups. It will contain the disabled, the elderly, those with ESRD, unless in your selection, your data request, you asked for a certain subgroup. So if you say in your request, "I only want the elderly," thus other groups will be there. But if you said, "I just want five percent random sample of all Medicare beneficiaries," you will get all benefits or benefit groups with it.

There's no indicator in the file specifically for new beneficiaries. There are fields that you can use to create it, but you--it's actually--but in order to really select who is new this year, you--there are ways of doing it, but it's not just a quick selection.

So why are all these denominators challenging? There's a whole list of reasons, right? But one of them is what I call date stamping, which I'm sure there's a technical computer science term but since I'm not a computer scientist, I don't know it. And basically, it's my--it's the idea that things can change over time and the question is, what is the rule for deciding which value is going to be kept in the file? Okay?

So if you could have--if you think about something like address, over the course of a year, most people don't move, but some people do. And how do you decide which address you keep? And so the general rule--so the general options are the first address, so we could say, "We're going to put for their address the address they had on January 1st of the calendar year." We could do the last address, we're going to put in the address that they had on December 31st. At the end of the end of the year, that's what we're going to say as where they live. You could actually even say, "We're going to have whatever information is the most current when we actually kept the file which could be well after the end of the calendar year." And so you can see those rules, they each have their own logic and they're each justifiable but you can imagine that if you had two different files, you use two different rules, and when you do cross tabs, you would get different answers. And the answer--and so this is what--this is where this gets challenging is you will find things change.

^M00:10:06 The good news about the elderly is that they don't move a lot. It's actually not a particularly mobile population so we don't get a lot of movement but we will get some, and the question then is, what is the right value for what we want to do and how can we make sure that we understand what value we have and the implications of it for our studies?

So the Denominator File is created when CMS pulls a whole bunch of stuff together. They use their own internal data, these data from the Social Security Administration, they use data from the Railroad Retirement Board, they get data from the states, they get information from claims that they use to trigger certain events, and they get information from managed care organizations. And that stuff is all combined by CMS in order to manage the program and then it's--and then that stuff moves down and trickles down and makes its way through the system for the Denominator File.

So underlying all Denominator Files is this mass of database which is currently, now, I emphasize currently called the CMS Enrollment DataBase or the EDB. Now, I say currently because I have renamed the file a couple of times since ResDAC has started. They've never changed the content, but they changed what they call it as they migrated probably from one computer system to another. So EDB is the current thing. And what CMS does is they take all of those sources and they combine it in the Enrollment DataBase. And in that database is a historical cumulative record of the Medicare program. So it contains, though I've never tested it, but it technically contains eligibility and enrollment information for every beneficiary ever entitled to Medicare.

Nobody ever gets access to the EDB. So you can't say, "I want 100 percent historical denominator." That would be really fascinating, wouldn't it? But we're never going to get it. But what happens is from that Master File, once a year, they extract the annual Denominator File, they pull off the information for that particular year. The CCW or the Chronic Condition Warehouse which is, again, just a different cut of the same data, is updated for a full year. So you can see there's differences in terms of when dates are done, but they're all done the same way, which is there's this big Enrollment DataBase that contains everybody who's in the program or who was in the program, they're combined down, and once a year, using similar but not exact algorithms, your denominator record is pulled out. So you can see that if you really wanted to focus on why one file was a little bit different than another file, you could keep yourself busy for years.

And so one of the challenges when you're using these data is to sort of, I believe, to get the right priority which is to say you need to do enough data checking to make sure you understand what you have and that you're comfortable that what you think you have is what you have. But you can drive yourself crazy. So if somebody stopped me during break and said, "Well, I've got--I've been using some data and I've got problems in less than one percent of my cases. What do I do?" And my answer is, "Delete them." There's always going to be a few history records. We've--you've done surveys and you get somebody who answers two questions like, "Wait a minute, you can't say you've never had a cigarette," and when we--but then answers the question and says smokes two packs a day. Like, look at these--people--you will get inconsistent information all the time. And at some point, you're going to have to decide whether the value you're going to get from trying to reconcile the inconsistency is worth the amount of time it's going to take to do it.

And so--and the problem is, is because we've got 43 million people, the

number of inconsistent records always looks really big, right, 'cause one percent of 43 million is a really big number. And so the challenge is, is to be--is to have this reality check of why do we think this is going to change, why do we think this is? Is this just 'cause sometimes, there's stuff that doesn't line up or do we think that this is something more systematic that we really need to understand? That is the challenge that you need to focus on and focus on putting this on to perspective. So under--when CMS identifies people, if you were good, you'd say, "Well, how do they know they have everybody and how do they know they don't have the same person in there twice?" 'Cause we're all thinking that, right?

And we have what's underlying all of this is called a HIC or Health Insurance Claim number. And it's an 11-digit number, okay? And like all things, CMS, the names for the parts have changed over time. And right now, what we call it is we call it a nine-digit CAN and two-digit BIC. And a CAN is Claim Account Number which just basically identifies how somebody justified getting their benefit. So it's usually a Social Security Number, but not always under which the benefits are claimed. So if I'm in Medicare and I'm getting benefits from my work history, the nine-digit CAN would be my Social Security Number, okay? If I'm getting work--if I'm getting my Medicare benefits under my husband's work history, that nine-digit CAN would be his Social Security Number, okay?

The two--let me finish this slide then we'll go. The two-digit BIC--so then the problem is, is if two people share the same number, how do we know we're not combining people into a single record, which could cause all sorts of problems and keep us really busy? And so what--and Social Security Administration also wants to track this because they don't want to send one check to two people or two checks to one person, right? So they've got the same problem that we do, which is they want to know how many people they're sending checks to every month.

So they have this two-digit BIC which is a Beneficiary Identification Code, and that explains how the person relates to the work history, okay? So it could be that I--for me, of course, my own work history would be my Social Security Number and the BIC would be an A and then a blank which would say, "She relates to--it is her own personal work history." If I got it under my husband's, it would be his Social Security Number. For his benefits, he would have an A, his work history. And for my benefits, I would have a B, saying, "It's his wife." Okay? Like, not even that hard. It gets hard 'cause it gets chaotic when you think about all the number of ways somebody can justify their work history, justify getting benefits under a work history. So there's a whole long list of things that frankly doesn't really matter to us, but it's just important you understand those.

So now, the question is, what do we get? Well, we don't yet, we used to get Social Sec--we used to actually get HICs, we used to get their real identifier with Social Security Number, and it used to be a huge problem. It was huge problem because my IRB didn't like me getting Social Security Numbers and so I was constantly fighting with the IRB saying, "No, no, no, no, this is okay," and the IRB is saying, "No, it's not okay." And I would say, "Well, I, you know, Google search your Social Security Number, you

can't find it unless your--" but it was a real problem. So what they--what CMS has done is they've said, "Well, you know what, you guys don't need their actual numbers. What you need is a consistent identifier so that you are sure you've got one person and one person only, and you can put the denominator, the hospitalization, security records. That's what you need. It doesn't matter what it is, it can be a random number as long as it is a consistent random number across all files." And they said--and that's what they did, and it was beautiful, and it really is. And it's called the BeneID which is a unique study-specific ID that can be used--that is used to differentiate people, okay? So that's what has happened.

But I wanted to--I want to go over the HICs and the identifiers because there are places where we still need to know about it and we still need to understand it. So one big thing is, you can tell IRB, you will not get Social Security Numbers. So when you go through that list HIPO waiver, and that the answer is you do not get Social Security Numbers, you do not get phone numbers, you do not get exact addresses, you do not get any of those things that causes--cause IRBs to get all nervous anymore.

The HIC, as we said, is based on a Social Security Number, and this matters because the Social Security Number is not a random number. I don't know how many of you knew that. And the Social Security Number as we all know is a sort of three segments. So the first segment is a state in which the Social Security Number was assigned and it goes from low to high, starts off in Maine with zeroes, goes down the East Coast, back up to Minnesota, down to Texas, back up the West Coast, and ends up Alaska and Hawaii at the very end. So if you--you can--if we did a little survey here about what numbers people's--what are the first three digits in their Social Security Number, we would find is sort out geographically by where people were born for the most part. The next two digits, so those are the first three digits.

The next two digits are a grouping number. Years ago, there were articles about how to use a group number to identify identical twins and stuff like that. I don't understand it.
^M00:20:01 We've never used it. We've never worried about it.

And then the final four digits, and these are the ones we care about, are random number. They're randomly assigned and without replacement. And we use that for sampling. So a systematic sample of a random number is a random sample, okay? So what that means is if we want to take--we don't want 43 million people on our computers 'cause even the really good ones would get swamped, right? We say, "Well, I just want like five percent." So now what are our choices for getting a five percent sample? You could put 43 million people into a random number generator and randomly sample them but that's going to take huge ugly resources to do that, right? So what they did instead is they said, "Let's take all possible four digits of this and let's randomly sample five percent of those. And then we will take anybody whose Social Security Number ends with those digits and that will be our five percent random sample." Pretty cool, I mean, and that is how the five percent random sample is created. And it's the same four digits. Actually, I think they are done, they used two digits. It's the

same digits every year so you don't have to worry that your 2008 five percent file is based on different people than your 2009 file. But this is how we do it.

The sampling algorithm was developed by the Census Bureau in the mid '90s maybe. So if you're using a five percent file, you'll notice if you start reading the literature, there's a lot of confusion about how researchers explain what that five percent random sample is. It's more likely wrong than right and--but you'll discover that in the literature about a lot of things, but that's what it is so this is how it is. It's based on Social Security Numbers which are randomly assigned.

Okay. The BIC, as I said, is assigned by SSA to keep track of how many people are claiming benefits under the same work history. So no two people complaining--claiming benefits under a particular work history can have the same BIC. So we say, "Well, this person had seven spouses where they would have a B1, a B2, a B3, a B4." Seriously, you can that many. You can have 14 disabled children and they would each have their one C1 through C14.

So if you look, the SSA has got over 60 categories BICs. So if you look at that record layout, you're going to have all of these possible justifications for benefit and it's quite fascinating in terms of understanding the complexity of the Social Security program. It doesn't matter for us because we can still--our goal is mostly to make sure that we know that we're looking at one person at a time and only one person at a time, and the rest is just curiosity while we're waiting for our programs to run.

So just to sort of summarize, the HIC is unique. Like I said, Social Security Number plus BIC, it's unique. No two people ever share the same HIC, either current or historical, okay, 'cause Social Security Numbers haven't--aren't reissued, no two people share it. Multiple people can claim their Medicare benefits under the same work history but with the addition of the BIC, we're able to sort that out. And most people use their own work history for benefits right now. So they're based upon the Social Security Number used for benefits. The vast majority uses their own work history but there are some who won't--who don't, who use her spouse's. In general, it's based on maximizing benefits.

So there's this algorithm that says you get this much money for this. And so it really comes down to if there's a wage difference, are you better off having two people under one income or not? And often times, when there are changes, it is women who are widowed who then use their spouse's work history if their spouse had higher wages. So that's--there are places where that process happens. I think it's personally very interesting in terms of understanding the program but, again, we don't need to worry about it, it's just so you know.

So if you're reading the literature or you could ask, I mean, I better tell you, I get asked sometimes the craziest things from journal reviewers. I tried really hard not to ask crazy things myself now because how annoying

it is. But they'll say, "How do you know that in, like, you know," like, well, "I don't care" but, you know, you're going to sort of politely say that. So some of this information is because these are the things that people who don't know much about the program may ask you about your data and so you need to at least have some basis for understanding the origins of these so that when you get asked the questions, you can answer them.

So as I said, HICs can change, it doesn't happen real often. We're lucky because back in the early '90s, we used to have to manually realign HICs when people changed it, and it didn't affect a lot of people but it was a massive amount of work. It was--because you're to take this ID and replace it and it was just--it was ugly programming even for people who are good at it. That's all handled now automatically in the BeneID. So the BeneID smooths over changes in what ID some--what Social Security Number somebody uses. They use--they smooth over which BIC they have. You don't event know if two people are using the same work history anymore, okay? So it's really a good thing and it's made our lives a lot easier.

The one thing you need to know is that the BeneID is uniquely assigned for a study. So different studies--so we could--so two of--we could both--Lindsay [phonetic] and I could both have the exact same cohort. Okay. We both said we want people who have--anyone who is hospitalized in 2007 for hip fracture and we went all of their claims for 2006, 2007, 2008. Exact same specifications, exact same study, we would have the same people 'cause we wrote the same study but her IDs and my IDs would be different. So, if we merge our two things together, we would combine people who didn't match. So this is--again, these are study-specific IDs. So what this means, the big thing, it means is that just because I have a record starting with 0001 and Lindsay does that we can't assume that her number one and my person number one is the same.

And that happens a lot, not like the example I gave you, but I have a data set and somebody else has a data set down the hall. I've got data from '04 to '07 and somebody else has '08 to 2011, because this is great, we can put them together and we have a longer sequence, and the answer is you can't just merge it. You will be combining people who aren't the same. So if you decide you want to extend it and you can justify it, you've got to write to CMS, you've got to tell them you plan on combining this data, you've got to get permission to do it, and then they will give you a--and you can request a crosswalk that allows you to crosswalk your Benes from your study with your colleagues, but you can't just do it on your own, you will combine people who are not the same. So, as we said, a crosswalk is available if you need it. And, again, this is really a huge improvement because there's less risk.

I mean, I think one of things you've got to remember is that, you know, CMS takes disclosure of subjects very, very seriously that's why we have the rule, for example, that no cell sizes of 11 or under, actual or calculated, can be released. Well, the reality is none of us want to have a breach. And so--but at the same time, what we also can do is we can say, "Well, let's make sure that we've done everything we can to minimize the chance of a breach and to minimize the chance of any real problem should

there be a breach," right? And that's the best we can do. And so this, getting rid of the Social Security Numbers, making it impossible to link to anybody else's data down the hall, anywhere else, is to everybody's advantage. Okay.

Residency, 'cause we often got--geography is one of the, one of the--geography is a big issue whether you care about geography or not, either way you put it, right? A lot of stuff will vary by geography. So even if you're not interested in it, you need to be aware of it. The state, the county, and the zip code of residence are in the file. So you can do things at any level. The question is, "Well, what are we getting when get this residency thing?" It's officially the mailing address for official correspondence. You know, like, well, great, you know, I suppose this was my South Florida issue. So, like, if they have their mail sent to their daughter in Philadelphia, I'm going to think they're living in Philadelphia when they're really in the South Florida. How are we going to get that right? Or the snowbirds, I don't know how many of you are from the north but I can tell you that in Minnesota, a lot of seasonal migrants are all migrating along with the birds. Everybody is going. They will come back when the ice melts. And so the question is, "What happens with them? How do I do who is part of my state?" And what's interesting there--so residency is based on different things. For the Denominator File, it's based on where somebody officially lived when the record is finalized which is usually end of February of the year after the file.

^M00:30:05 The Bene Summary File is at the end of the calendar year. So there's like a three-month difference, not a big deal. When I--when we've done studies sort saying, how much of a problem do we think this is, the answer seems to be not much.

So a few years ago, I did a study trying to figure out hospice service areas. And that's a really good one to figure out, quality of address, information, 'cause hospice is--hospice services are generally home-based and so people aren't going to be traveling a long way. So one of our concerns was that we might see like crazy things happening, and we did see a very, very small number of people who officially lived in Arizona who were getting hospice care in Duluth, Minnesota. Now we know that hospice wasn't driving. Okay, so there's right--so we got a few of those cases but it was--I mean, it was maybe a quarter of a percent, I mean, it was not--we were really impressed with it, you know, 'cause the challenge is like, if you try to do concordance by state, you can't really automate it, you got to stop and look at the state and say, you know, how close are things, and so between services and people. So hospice was a really good one because it's home-based and it was maybe a quarter percent of people where we had to say, you know what, this is so far that we can't think of any rational way that this person could be living where they are, being treated by those hospice, where it is.

So I came out of that study which was about something else, really impressed with sort of the quality of the data. I think what really goes on is that people use more mail forwarding than having things directly said. So I think it's more likely that somebody will have--that have there mail dropped

where they are and then forward it on to their daughter, but they keep their mailing address where they are. And I don't--some of that is probably taxes, right? 'Cause if you live in Florida and there's no state income tax, the daughter lives in Philadelphia where they pay tax, you want your mail going to Florida, right? So I think it's some of that stuff, but I was actually again pleasantly surprised, and I think we can trust that address information, again, little craziness but not bad.

Medicare beneficiary is the vast majority of beneficiaries are elderly, old age security supplemental income, okay. See if I can do this. So vast majority are elderly, 86 percent, the next biggest group is disability, and the smallest group is end-stage renal disease. Just--I think Marshall told you this but some--when people turn 65, they are no longer disabled. Disability means people who are not age-eligible for the Medicare program. So once you become age-eligible, you lose your disability status 'cause you're not disabled anymore. Okay, so this is how you are justifying your benefits. And the default justification is age and when you can have that, that's what counts.

So there are fields on entitlement and these can be useful if you really are interested in people who started out in the disability program. So what we have is the original entitlement and the current entitlement. So for example, somebody who was originally on disability and then currently on old age, it would tell you that they started off in the disability program, they aged into the base program and that's where they are now. So you can tell how somebody started in the Medicare program if you want.

The Medicare status code combines current entitlement with end-stage renal disease information. So technically, so although I said that you lose your disability status, you do not lose your end-stage renal disease status. So you can be aged without end-stage renal disease, aged with end-stage renal disease. Disabled without end-stage renal disease, disabled with end-stage renal disease, or only have end-stage renal disease. So that's--this can be very useful if you're looking at the ESRD program and if you're trying to understand how that all fits together.

Medicare status code or the disability is important and I--this is one of these things that--is probably like highest on my pet peeves. So if you submit a journal article and you make this mistake and I review it, I'll tell you that I'm going to come back at you and it drives me nuts. So we'll see these articles. So we use the Medicare population and the average age was 65. Now, how do you do that with the benefit at start? You know, but--and what happens is we get--we've got these people who are disabled or with end-stage renal disease and they will pull the average age down enough so that we--they're starting to dip below the typical Medicare age, and what that means, it's not about age, but it's--these are really different populations.

So first of all, when we look at basic demographic, mean age, mean age of the elderly is 75 years, 74.6. For disabled, that's 49, for end-stage renal disease it's 46 years. So we can see that these two programs are much younger by definition almost, than the elderly program. But look at this,

in terms of percent male. The elderly program is dominated by women. Women have longer expectancies than men and more likely to out-survive their husband and it makes sense demographically, that there may be more women than men in the elderly--in this elderly cohort.

Disability is tricky because disability in the Medicare benefit is about two things. It's about having a disability and making it through a process to have that disability formally recognized. Okay, so who is more likely to go through that? But it turns out it's men and there are many, many studies suggesting it's not that men are more likely to be disabled, it's that men are more likely to be the primary wage earners which means that the loss of their income is more important--is important enough to their families that they need to make it through this process. Women are more likely, if they have a work-limiting disability, they just stop working.

But look at the annual mortality, 6 percent for the elderly, 2.6 percent for the disabled, and 8 percent for end-stage renal disease. So why are the disabled so low? The reason that disabled have such low mortality has to do with what you have to do to qualify for the benefit. So once you qualify, there's a 29-month waiting period before your Medicare benefits kick in. So what we end up having for the most part is non-fatal disability, right, because you've had to survive long enough for your benefits to kick in. So this is actually a relatively low mortality group.

And then if you look at the top DRG, so what is it? General classification for why people get hospitalized. Heart failure for the elderly, psychosis for the disabled, and dialysis-related procedures for end-stage renal disease. So they're a very different population.

And so, the thing that drives me crazy, 'cause you're waiting to hear that, is when you put them all together, what you end up doing, I believe, is you end up losing the uniqueness of each of these populations. They are each important, they are each different, they are each deserving of study. And when you put them together, you get sort of like a bad stew where you can't--it's all there but it doesn't--nothing taste like itself. And that's my analogy. So when people put it all together into one pile, I worry.

So that we've got a population basis for the elderly, right? We've said this. We've got 98 percent of the elderly, 99 percent of the elderly by the time they are dead. We can talk about this in a population basis, we can calculate rates. For the disabled and the end-stage renal disease, it's a lot trickier 'cause we know that this is not the vast majority of people who have a work-limiting disability. So work-limiting disability of a certain magnitude, who have made it through a certain process will have enough quarters and so on. And it in no way, reflects the level of disability in the US population.

Okay. Age and date of birth. Age is calculated differently. So this is again the date stamping. In the denominator, the age is the youngest the person will be. In the Bene Summary File, it's the oldest they could be. So I would be honest with you, I actually think this rule is better. My

staff had always had a really--we all have a hard time with this 'cause we do the shorthand. We say, usual restrictions, right? Age 65 or older and you can almost see the program--if they have age greater than or equal to 65, right? Easy, easy as programming except for, in the denominator, you actually want to have age greater than or equal to 64, because everyone who is 64 in the denominator file turns 65 during the calendar year and becomes elderly. In the Bene Summary File, they said, well look, everybody gets it wrong so we're just going to make it so that you can get it right by saying, we're going to select people who, at the end of the year were 65 or older and not worry what they were at the beginning. So it's really important to look at what file you have, to read your documentation, and to make sure you know which of these two rules you were following, because it will affect the new enrollees. Okay?

^M00:39:59 So we don't want to include disabled people if you're using a Bene Summary File, you don't want to exclude all your new enrollees if you're using a denominator. There are these concerns. And I don't think--if you haven't checked this again in a couple of years. But we have sort of these persistent problems that Medicare misses deaths. And that's there are too many people who are too old. And that--when you look at the Medicare enrollment information, there's too many of them relative to what you would expect from the census or the Guinness Book of World Records.

And so this is was from 06, it's what needs to be repeated again. What we did is we compared the number enrolled in the Medicare program and in the census. We see that between ages 90 and 94, there's a slight excess in Medicare, that's actually offset by the slight deficit in Medicare in 95 to 99. So I suggest there's maybe just some rounding error. But look at the difference for people who are a hundred or over. The census says there are 68,000 people, Medicare says there are 177,000. That's a massive difference. And we can see the census doesn't even break these numbers down. And yet, according to Medicare, there are 5,000 people who are 130 years old or older. And so, the question is, what do we do with these? I mean admittedly, these are really small numbers, but these are troubling.

So what you find and some of this makes sense is that, you know, when we write a method section for a journal article, we never really tell the whole truth about everything we had to do to get our study done right. So if we ever told the whole truth, no reviewer would ever accept it. Like what do you mean you get rid of the 130 year olds? So even though people do it, nobody talks about it which means there's no standard practice of reemerging out of it.

So we've tried a couple of things. And I'll show you sort of where I am. First of all, we said that anybody over 90 or over a hundred who has no healthcare use in year, we get rid of, 'cause those people are unlikely--they're unlikely to really be alive 'cause it's really hard to imagine a 100 year old who doesn't at least get a flu shot or something. We also looked at getting rid of anyone over age 90 who doesn't have part B coverage. This means that they've got coverage for hospitalizations, but not for any physician services. But this is an important restriction because if you remember, you have to pay for your part B coverage. You

don't have to pay for part A. Part A is an entitlement. So what we're basically saying is that nobody was willing--if no money has changed hands, then it's less likely to actually be real, given how much services have shifted to part B.

And then we got sort of silly and said well, anyone older than the oldest person in the US we should probably get rid of. But the reality is that the rule that I like is sort of this first two which says "for the oldest old, I want to use and I want some sign that they're fully insured." But since my studies generally require part A and part B coverage, this rule happens anyway by default. So we've got to remember that it's a really small number, it's like a third of a percent. But depending on which population you're studying, it can still add up. And if you're using the data and you're not expecting this and you see that one of the things that I think happens is at least for me and for my group, is that we say "Oh my gosh, what else has gone wrong here? What else am I not understanding, that sort of causes this whole--these series of alarms to go off?" And so by knowing that this is an issue and having a plan upfront to handle it, it's just less troubling, we just sort of, we can move on, we can be aware that this is not perfect. But we can be more comfortable with the imperfections.

Sex is coded one is male, two is female. There are no missing values from this field. So instantly, we should be nervous, right? 'Cause there's always missing values. And then when you read your record layouts, it's had these great little charty thing here which cause my blood pressure to go up dramatically very quickly. Persons with missing information filled in according to this rule. If ages less than 65 and sex is missing, then they're male, if they're greater than 65 and sex is missing, they're female. And that makes sense, right, we just showed you the demographics. If you had nothing--if you know nothing else but their age, that's a reasonable guess. But that's like not making me real happy. Like we would prefer to have it missing and we could choose to apply the rule than just have them do it for us. So we--so I had to go and like figure it out for myself like how uptight should I be.

So first of all, we look--we had to think that what are some examples of things that should be exactly correlated with gender, so prostate cancer. 100 percent of the men with prostate cancer--100 percent of the people with prostate cancer were male, that's good. Ovarian or cervical cancer, 99.98 percent of the people with ovarian or cervical cancer are female. But of course, these two cancers also happened to younger women, which might explain this a little bit is that--if there are people who are disabled, that would be where they would show up. And then we look at breast cancer. And breast cancer looked awful. And the reason is, is that man can have breast cancer. And there's technically a different ICD-9 code for male breast cancer than there is for female breast cancer, I don't know how many of you knew that. Anyone? But what it appears happens and so we've checked this a couple of places is that, you know, female breast cancers are so common.

And I can tell you. One of the things I did in grad school, one summer

is I did ICD-9 coding. And you stopped looking up there, you know the codes, right? And so you can imagine if you're working in a college office is like "I know my breast cancer codes. This is all I code." And we think in some cases, they missed the fact that there's a different code for man and for woman in breast cancer. So that they code the man using the female breast cancer code. So then we looked at breast reconstruction which would be much more female-based. We found that 100 percent of the people who got reconstruction were women. So you will hear things like, you know, I was looking at this data and I found a man with a hysterectomy. If they can't even get that right, what can they get right? And I've heard people say that to me and you will hear that and the answer is "Small problem would be better if we could figure it out for ourselves so we could choose to remove the people with missing gender or do some sensitivity analysis." We can't but all of the examples that I personally have gone through, I'm just showing you a couple--have lined up saying that this has probably much to do about nothing. But, just be aware of it. And if you see a man with a hysterectomy I probably would remove him from my analysis, okay?

Race, race was originally back in the stone ages coated as white/black, other unknown. And in 1994 the race codes were expanded. And it would add Asian, Hispanic, Native American or other and unknown, okay? And what's interesting here--so what you might ask is how come Hispanic is part of the race and why isn't there enough ethnicity field like the Census Bureau had asked, right, and if any of you were thinking that. And the reason is, is back in 94, CMS had enough space in the denominator file for one--they have one extra--they didn't have extra columns. So they could expand the number of values that a column could take, but they didn't have spare column for that file. And they didn't want to have to recreate the files in 81 column file. So they combined the fields. This is, again, if you think about it, it's like early 1990s computer technology driving our ability now to study race, 'cause it's never been fixed.

So what they've done is they worked on efforts to update racial classification. So they've done things like they've reached out to minority communities, they've reached out to Native American bands and the Indian Health Service to try to get those. The biggest problem that we have is on the Hispanic race and ethnicity code because of course you can be white and Hispanic or black and Hispanic or Asian and Hispanic, right? And so this Hispanic race ethnicity code has an estimated sensitivity of about 35 percent. So that one's our problem.

We can show that we're getting better, there's greater racial diversity especially with Asian and Hispanic, some--this is actually doubling with Native Americans but this is Puerto Rico. And if you see, this is a percentage of Puerto Rico that's Hispanic. And it's now up to--in 07 it was up to 21 percent. Okay, so this one--I would say it should be a whole lot higher. But again, in Puerto Rico, we still have people who are classified as white or as black and not listed as Hispanic. There are--there are ways as a race variable, surname race variable. And what you can see with Puerto Rico is that surname variable gets up almost to a 100 percent of Puerto Rico as Hispanic. So there are ways that we're using, that is they're preprogrammed in the Bene Summary File to supplement

this information. So if you want to find Hispanics and look at that, you can. So what you'll see is that when you compare them, the RTI variable which is the race reclassification doesn't change whites, doesn't change blacks, but what it does is it moves an awful a lot of Hispanics mostly from white to Hispanic. And a few from black to Hispanic and then there were some that were classified correctly in the first place.

^M00:50:12 So if you are interested in Hispanic, in sorting Hispanic people, I encourage you to use the Hispanic surname variable combined the race, with CMS race variable rather than using the CMS race variable alone.

One just--other thing to point out is that often times we talked about how amazing statistical power we have? Race is still one of those places where it's really easy to end up in small number situation. There's no standard breakdown or grouping of races. For a while, I actually used to argue that we should do black, non-black versus white, non white, with the idea of sort of where the most classification was. Journal editors don't like that. But it gets really hard because, you know, combining Hispanic, Asian, Native American doesn't feel right. So--but just a warning, it's not uncommon, how's that? It's not uncommon to have to combine racial groups. And just think about it and try to think about it, I would not argue not just mathematically but really conceptually about what it is about race you think is going on, what is your mechanism you think is working and try to do groupings that are actually consistent with the phenomenon you're studying.

Mortality. The mortality drives everybody crazy because it's really confusing. There are two fields about mortality. There's a date of death and there's a death date validation field. And what happens is, if somebody is alive, their date of death is blank, right? Not dead, no date of death, it's blank. Once they died, their date of death is filled in. So anybody whose got a filled in date of death is dead. The problem is, the death date isn't always exact. And if it is exactly, it's noted with a V. And if it's not exact, it's blank. It makes sense? Okay. And so what happens is that the way date of death comes down explains why this is happening.

So the date of death information comes from Social Security Administration. But CMS every night, checks the claims for people from hospitalizations, from hospices to get people who are discharged dead. Anybody with the hospital or the hospice said they were discharge dead or the emergency room that's transmitted--as I say, to trigger an investigation. Okay. And the process happens--I actually got to see this letter from a family member saying "We have received a word that you might be dead." If you are not dead, please let us know. It's a crazy letter but it's like how else do you say it, right? And so, yeah, I mean, it's like you can see--everyone's, while you've seen, like a newspaper article if someone holding this letter up, saying they think I'm dead. But this is what happens, you know, how do you explain this.

And so this is the process that is triggered. So what happens is that being dead triggers a stop to Medicare benefits, a stop to your social security benefits and so on. So this is real tension between not wanting to live,

not wanting to pay for services for people who aren't alive and not wanting to cut off insurance benefits for somebody who may still be alive. Right, that's this tension. So what they do with Medicare is they say "Look, we know you died and we don't have the documentation, we really expect to have to be absolutely certain, we're going to set your date of death to the last date of the month. So like, this is the last date you could have been alive, so that we don't want to cut off healthcare benefits too early. So I mean, so when we're doing survival analysis, we could say, "Well, let's set it to the 15th." 'Cause on average, that's right. And from a study that makes sense. But if you're thinking about health insurance, like on average, you cut it off correctly, it doesn't quite fly, right? And so that's why they move the death date to the end of the month 'cause for CMS and for SSA, this isn't really about death date it's about the last date of services.

So if you can remember that and understand that, this makes total sense. It's the only decision CMS could make. If they don't have information they want which is usually a certified death certificate. So if somebody who lives in another country or has a benefits coordinated by the Railroad Board or whatever thing. If they don't have that certificate the way they like it, they are not willing to certify that--you're not willing to certify an exact day. In those cases, they will say "We will grant the month. You will take the year and we will move into the end of the month," and that's the last date that person could have been alive. And now it's when everything stops, makes sense? So that when you're looking at validated dates of death, what you discover is they're pretty steady across the entire month. And that's exactly what we'd expect and because it's such a huge file, our trend line is really flat, just click. When you look at non validated dates of death, there are none until in this case I think the 29th so it must have been a leap year 30th or 28th, 29th 'cause of how the deaths are recorded in there, sometimes getting a little bit over a year, 31st and so on. So what happens is you will get this tail here depending on how long the months are, how many days are in each month. So that's what's going on but this causes a huge amount of confusion because people will say "well, their death wasn't validated so I shouldn't count them as dead." And the answer is no, their death--they are dead. What is--what's left to some--what's left open to debate is the exact day that they died.

If you really care about this, so if this matters to you, so sometimes to some of our studies and we really struggled with like, could we make a recommendation about how to handle this analytically? So, the problem is, if we include the non-validated death dates as an actual death date, we're going to overestimate survival. On average, by 2 weeks, which depending, on what you're studying, probably isn't going to be a big deal. You're studying survival with some long chronic condition, two weeks aren't going to matter. If you're starting post discharge mortality, two weeks is massive. So the implication and the magnitude of problem for this overestimation is going to be very study-specific.

So here are the things we went through, if you sense your people, if you say "Well we're just going to get rid of everybody with non-validated deaths" it's actually what's called informational censoring. That is to say we're nonrandomly moving the people who--removing the people who have

the events. We may actually be causing trouble and changing it and misestimating them, right. We could put it to the 15th, which sort of has some appeal. The other thing that some people will do is they will create an algorithm and they will say "Well, I'm going to look in the claims to figure out and I'm going to set the date of death to the date of the last claim, if it was on, if it was in an in-patient setting, or the day after the last claim of an outpatient basis. So--'cause people would say and we've seen it ourselves, you'll find the hospitalization where they're discharge dead, why can't they just use that date? And the answer is you can. If somebody was only seen in a clinic, it's a little trickier to think they died in that clinic day, so some people would give them an extra day or two, just to kind of match what seems reasonable.

But this is the dance that we need to go through, so before you go through the gymnastics of finding the hospitalizations and finding the hospice deaths, ask yourself how much precision do I need and if you need the precision of figuring out the last day they were in contact with the healthcare system, then do it. And if you don't, pick the end of the month, pick the middle of the month, pick the beginning of the month, make a rule and move on because you're not going to gain enough from that precision to be worth the amount of time it's going to take. It's my personal advice.

You can see why CMS and SSA want to cut off benefits as quickly as they can, they sort of frown upon providing healthcare for dead people. You know, sometimes when I looked at those, some of that actually looks like it's probably a date error. So, looks like, you know, where something happens like you know, like--like they were listed in the MedPAR as dying exactly a year before they did or people would get just the classic sort of October 6th to June 10th kind of problems and so some of those problems are that they--it generally gets cut because one of the checks that will happen whenever a claim is submitted is, is this person eligible for benefits? And as soon as somebody dies their benefits are cut off.

So some of these that you may be seeing are these exact ones that we talked about where the hospital mistakenly says somebody was discharge dead. So they filled in the wrong discharge status code, which doesn't matter to the hospital. It triggered an event, it triggered an investigation, which was of course rejected. They got rid--and then eventually they died and if that happens close enough it's going to look like you know, some hospital is turning and admitting dead people for extra care. But that--and it maybe that the hospital has never corrected the initial discharge status or discharge status code. So again, if you really want to know discharge death use the denominator, it's the better source which is sort of one of the themes of this talk.

Benefits, as you know, it was Part A and Part B, the basic benefits. ^M01:00:00 94 percent of the people, so just about everybody have both Part A and Part B. 6 percent of more months Part A services than Part B, you're not required to have Part B services and the big issue is you can waive part B and restart without penalty if you have, if you can provide proof of another insurance it's equivalent. You can waive it and start with penalty if you do not have other insurance. The issue that we have and

the reason we care about this is about what--the sort of the phrase that has emerged is likely to have complete claims. And that's really what we care about is we want to limit ourselves to those people where we're likely to have the whole picture of what's going on.

And so we've done some analysis and one of the things that we find is that if you compare hospitalization rates, if you go with Part A and Part B service and people with A only. Now these are hospitalizations which are Part A benefit. So everybody has Part A coverage. What we see is much, much higher hospitalization rates among those with complete Medicare coverage than with those without complete Medicare benefits. It's likely two things. Some--like this deep here might be those dead people that we talk about or those really, really old people without any healthcare use, so they're just staying in the system which is why this rate looks low. The other thing is it probably reflects people who have other insurance. They have other private insurance, they use the VA, they use the Indian Health Service and so they don't use Medicare, they waive, they choose not to pay for their Part B benefits and therefore, they're Part A only. And that's the likely explanation, it is not that the people who are A only are really healthy, it's most likely that the people who are A only have other sources of healthcare coverage and therefore, we are not seeing complete claims in the Medicare data.

And we actually compare--confirm this with SEER-Medicare aid only. The advantage, this is old but the advantage of the SEER-Medicare data is that the SEER cancer registry is actually looked for use. And so we'd look at people who'd had colorectal cancers treated with colectomy which is a major procedure, major hospitalization. Now what we found is that our ability to find, in claims, the hospitalization was dramatically lower for the A only and those with A plus B which again would suggest that there's, this loss is due to people having other sources of insurance or other sources such the VA. So this again gets back to this idea of my strong recommendation is always limit your cohort to people with both A and B coverage even if you're looking at a Part A service.

Medicaid, every state exercises the option of paying Medicare Premiums. And this can be done in terms of the state pays the Medicare Premiums only. It do premium plus cost, this is your Part B insurance, this is Premiums and Cost Sharing which means both paying the insurance price and paying the co-pays and deductibles and then full Medicaid benefits which includes long-term care, it used to include pharmacy and in the denominator, this is field state buy-in.

And all the state buy-in fields says, is that one of these three things is going on. That doesn't tell you which. The biggest mistake with state buy-in is, I have to tell you, this is my pet peeve, is everybody uses this as a surrogate for poverty. People say we controlled for poverty. State buy-in is not poverty. State buy-in is a state helping with these things. So while it's likely correct that everybody who's in one of these three programs is poor, it's also likely correct that there an awful lot of really poor people who aren't being helped with any of these programs. And every once in a while I pulled it out and should redo it. We actually mapped,

rate of state buy-in and we could see the outlines of States to the outlines of South Dakota. We could see the outline of map at the county level and we could see the outline of New Hampshire and the outline of South Dakota to pick on too. Now when you can see that, that tells you, as you'd expect, that part of state buy-in is state policy. It's not just poverty. So be careful using this as a surrogate for poverty. And it's--that's again another one that I've argued with editors and reviewers about, so why didn't you control for poverty and I say, "well I would if I could and I refuse to set state buy-in variable." They don't know who'd appreciate that.

You can count on a monthly basis whether that's state buy-in A, B or both, most people are both and they--and you can look at it. The state tells CMS when they're state buy-in and then CMS builds a state directly for the premium instead of billing the beneficiary. So that's why the state, the CMS manages this. This is why they care is because it is about sending the bills for the premiums and the bills for the co-pays and deductibles. As we said, it tells you whether the beneficiary is covered by the programs, but it is not a clean proxy for income. So just be careful with that.

They're also our monthly indicators for enrollment status, Part A, Part B and this enrollment. For every month, you'll get an indicator that says not entitled, A only, B only, A plus B and then combining that with state buy-in. So this is sort of what they look like. Like if here's a whole row of Cs that means they had 12 months, A plus B state buy-in. Here's somebody who had four months of A and B and then nothing. Here's somebody who had 10 months of nothing and then 2 months of A and B coverage.

So the challenge is, you can't tell what happened here. We can guess right? We can guess that they died and that's why they lost their coverage. If they went from A and B to A, we'd say they must have stopped paying their premiums when they go from A and B to nothing, they lose their part A also. Mortality is a likely explanation but the point of it is, that the monthly indicators don't tell you why the changes. They just tell you what the status is and here's other information to figure out where the deaths are.

Oops, I'm going around with so--beginning in 06, the Bene Summary File gives you additional information about the type of state buy-in. QMB, QMB in full Medicaid so if you care about which program somebody is in, and whether or not they've got full benefits or premium assistance, it used to be that we couldn't tell. And now the answer is now we can. So if you're interested in these programs, there are groupings of these and you can find documents that sort of say combine this, this, and this to mean this program, it's again there's a lot more detail than we need but this is now a huge benefit to anybody who is looking at the effect of state support for Medicare beneficiaries.

Managed care enrollment is a challenge and it's important to really focus on. So what happens is with managed care, CMS pays the managed care company a fixed fee and that person, per member, per month fee and that managed care plan is responsible for all of their care and they must at least provide equivalent care to that which somebody would be eligible for in the Medicare fee for service program. If somebody were to--if a managed care, if a claim

were to--mistakenly goes through and be sent to CMS, CMS would reject it, send it back to the plan. So the challenge is figuring out who these people are. So you can get both monthly indicators and the summary count. How many months of coverage do they have in this calendar year? It's important to know that none of these fields right now give information on the specific managed care plan.

So this is something that we've gone back and forth about, which is, I want to know what plan it is and I want to be able to compare people in plan 1 to plan 2. You can't do that. We've tried--the policy may change so I'd certainly encourage people to ask if you've got a good justification for why it's important. But this is really more to use for selection and to limit your cohort. There have been a straight--demonstration projects, but for the most part people in managed care are required to have both part A and part B and just about everybody does.

So this is what the plans, what the fields look like. Not in managed care, risk managed care, this cost managed care is something where the CMS processes some of the claims and not all of the claims. And this is not particularly common in on demonstration programs.

So this is what the claims look like. Again, so you see here somebody who's never been in managed care. We don't know from that how many months they were in the Medicare program. We only know they weren't ever in managed care. So we just have to remember that these fields don't line up. You could be not in managed care because you got tired of or because you died.

Many years ago, back when I was in Florida, we were looking at factors affect--we're looking at hospice use in fee for service and in managed care. My analyst called me up one day, she's very excited and she said it's amazing, she had about a month after they enter hospice, just about everybody drops out of managed care. And this was at a point when sort of the policy discussion was whether managed care plans was skimming, I'm like "Wow it's really interesting." Well we finally figured out that about a month after people enter hospice they die.
^M01:10:04 And the reason they were leaving managed care was because they weren't alive anymore and just as CMS doesn't like to pay for health care for dead people, they don't like to pay managed care enrollment for them either. And so what happen is because the fields were two separate sets of fields, we had to put them together properly. And so we misinterpret it a switching out by choice versus switching out from death.

I tell you those stories 'cause they're sort of funny, they're sort of obvious right, you're listening, you're thinking, "How could they make such silly mistakes?" But it's really easy to make these mistakes. So if you get something that looks too good to be true, you're imagining the headline and the greatest story you've ever had, you probably did it wrong. It's just you know, so just take the time to look it up and make sure that you really are sure you didn't make a programming error.

Most people in managed care, either have none or have a full year, there's policy changes happened in the late 90s and early 2000 so people couldn't

switch in and out on a monthly basis. It's an annual enrollment and that certainly stabilized things.

There are these things called cost managed care organizations and they're a hybrid. And what it is and we--CMS processes the hospital, the skilled nursing facility and the outpatient claims and a few selected carrier claims. And so the challenge that we've had like, just one of my projects is, can we use the information we have? So we see hospitalizations, can we keep them and this is our denominator-numerator analysis right? Can we keep them in our, you know, we can find them in our numerator mostly.

And the way we came down to it and the study that I was on, actually I've used it twice, and in one case we decided to keep them 'cause it was really focused on hospitalizations which we know we will have, 'cause CMS process it. And in another case, we decided to exclude it. And it had to do with how much we thought we needed the carrier and how much we were going to lose by not having the carrier and how incompatible those two populations would be. So I've made a decision to exclude them, and I've made the decision to include them. And I was very study-specific. So what I'd recommend is that this is one of those where you want to stop, you want to think about it. Figure out what you're studying and our decision was based on completeness of data and ease of the method section. Because this is still such a small number that we decided in one case to have like a separate subcohort halfway through our paper, wasn't going to buy us anything and was going to make things a whole lot trickier. And in another case, it wasn't going to cause anything to be anymore difficult.

As always, you cannot tell whether somebody disenrolled by choice, or because they were dropped. You can't tell if they died, you can't tell why they're not in managed care. But there are other indicators that let you do that. We also know from the Bene Summary File we know what type of plan they have, managed care organization, regional PPO, description drug plan and so on. For part B benefits, we know whether they get premium subsidy and co-payments. We do not know specific plans. And prior to 2010, we don't know the exact formulary of their plans.

So sort of to finish up, this is idea that I want to--you don't think about, which is what has come out in the literature, is likely to have complete claims. You'll often see it. It's almost become a term of art. We restricted our study to that population likely to have complete claims. What that is, is equal months of part A and part B coverage and no managed care enrollment. Often we'll require the conditions for some period of time, one year prior to one year after, 3 months prior to 6 months after, we make a window and we say, that's the window I want to search. If you want us to consider part B coverage too, just remember that that's another subset. And you have to figure out how you put that in.

You can also, and I'm not going to do this, you can read about this. You can do person-months if you really want a technical epidemiology. You could rather, than--coming each person once, as one, everybody equal, you could come in months there in the program, so you could end up getting this person-years at risk which are going to be always be lower than the number

of people, right, so you can see that you can count the number of person-years at risk for a population or the number of people in the cohort that year. So if you want to do it, you have all of the information you need to do that. And you can see that by doing that, what you'll see is you'll adjust your rates particularly with the oldest and with the--a little bit less with the youngest. So if you want that level of precision, it's completely possible.

So just to summarize, CMS has this enrollment database which combines information from social security and managed care, Railroad Retirement Board, states and others along with their own data. And then from that, the denominator and the Bene Summary File are cut each year. So this database here is underlying and that contains the information. So it's just an extract for us.

So one of the things that happens is that people say, "Well, do I really need this?" You know, my data cost are getting so high and I just have these start-up funds, and I'm trying to figure out how to get the most. And do I really need the denominator file because I see that I can get gender and date of birth in the med part, so can I just stick with that? We get that question all the time. And I just want to remind you, first of all, it's always good and not just because I'm teaching it, it really is. If you're following a cohort overtime, you really need to have it because people can drop in and out particularly if you're looking at the disabled. If you're combining with an external source, you have to have it. Otherwise, you can't differentiate between didn't link and linked but didn't get hospitalized. Okay, so if you want to combine a source, you have to have it.

And if you're tracking somebody from a Part A service to a Part B service, you need to have it to make sure that they're in the part B. So if you've got a hospitalized cohort and you want to look at what happens to them in the 30 days following hospitalization, and their use of procedures, in-patient or out-patient, you have to know whether they're eligible to have any part B claims. So that you can differentiate between, they don't have part B coverage, we don't know what happened to them. And they have part B coverage, they just didn't use any.

And then just a reminder, if that both--end stage renal disease and the disability programs, it's important because in both cases, people can lose benefits. You really can't lose your OASI benefits part A. You can lose your part B benefits by failure to pay premiums. But you can lose your end stage renal disease and disability benefits.

And so we have cases, we have stories. This is an old one but I like it where somebody called me up and they said "We've got the 5 percent carrier files and the number of people in the file is decreasing." We know at this point, given the demographics, the US population should be increasing, physician office visits shouldn't be affected by other changes. And then this is always the question we get, "is there a problem with the file?" And so we pulled the denominator and we find that the denominator was rising exactly as we expected. So then we say "What is it?" Well it turned out

that managed care enrollment, during the same period was rising incredibly steeply. So it's a period of huge growth in managed care. So the growth in managed care was greater than the increases just due to natural aging of the population. And so there was a net loss to the carrier file because it was a net loss of people in the fee for service program, over this period.

So the point of that all is, is that, with the denominator file, these researchers could have figured it out. With the denominator file, they could have figured out the bias and the shifts in their population over this period, due to managed care enrollment, that they couldn't see, just looking at the carrier file alone. So there are some proposed changes to the denominator file.

I've been told I have to tell you, some of these are my proposals, so not necessarily ones with anybody who's got any say, but I keep putting them forward and sometimes they get picked up. There's been discussion about creating indicator about whether the geographic information reflects the beneficiary's physical location or mailing address. As I said, that would be nice. My experience from the hospice study at least would be that it's not a crisis. And that in fact, they line up far better than I would expect. And there's a primary payer code. So if you remember, and I think Marshall told you private insurance pays first, right? Private insurance, Medicare pay second, MediGap pays third, Medicaid pays fourth, patient pays whatever is left. Not everybody has every type. So these primary payer people are really tricky. They are the ones who sometimes don't have part B coverage. They are the ones where the price is sometimes left off because somebody is paying ahead of Medicare.

And we'll see when we look at claims tomorrow morning. You can see a primary payer amount. So when a particular claim, you can know whether a primary payer helped. But what we don't know in our denominator is whether they have a consistent primary payer. I know from family members that once a year, they get a letter from CMS, saying you know "please confirm all health insurance you have. Do you have a primary health insurance, what is it?" So every--so they, so CMS knows this. And so the question is whether we can figure out a way to put it on the file. So when we're studying utilization, perhaps we would consider, either stratifying by primary payer or even restricting to people who--where Medicare is the primary payer. So this is proposed. They don't know--at least my proposal, maybe other people's as well.