Hello

Today's talk is about Data Linkage and I'm going to give you a very brief overview.

My name is Natalie Shlomo and I'm based at the University of Manchester.

So what do we mean by that Data Linkage? Well Data Linkage brings together information from two different records that we believed to belong to the same person based on a set of matching variables. Now if the two records agree on all the matching variables it is unlikely that they would have agreed by chance and we can be quite assured that the link is correct and will be high. In other words the pair will belong to the same person.

If all of the matching variables disagree, the pair will not be linked and it is unlikely that it belongs to the same person. And the problem is of course when we have intermediate situations where some matching variable agree and some matching variables disagree and we need to predict whether the pair is a true match or not a true match and often this will require some sort of clear clean intervention to determine the matching status.

The problem of course in data linkage is the presence of errors i.e. would we collect the data and where we don't have unique high-quality identifiers in order to carry out the linkage. So the challenges of data linkage or the errors, the variations, the missing data on the information that we have to link to records together. The differences in the way the data is captured and maintained in different databases. For example we might have different versions of the date of birth compared to age. The dynamics of the databases the changes over time for example name changes due to marriage, divorce, address changes etc. these are all challenges in a successful data linkage. The typical problem in strings, matching variables that are strings we can have misspellings, transpositions fused or split words for example the first of the last name may be fused together and we need to split them. Missing or extra letters the way these strings are a typographical errors in the strings extra information missing punctuation typical problems also arise in numerical variables where the numbers may be transposed or there may be insertions or deletions.

So data linkage typically involves three stages: the first stage is the pre linkage. This is where we need to edit data and clean the datasets, parsing fused strings, standardizing the matching variable and this requires quite a

bit of work to make sure that the two databases are able to be compared through their matching variables.

The second stage is the linkage itself to data linkage. We need to bring together all possible pairs for comparisons and determining the correct matches. In other words do they belong to the same person? All possible pairs are produced within something we call blocks which I mentioned in a few minutes these are determined by blocking variables. And finally the third stage is the post linkage this is where we need to check for residuals or pairs that were not linked. We need to determine the error rates and to make sure that we have error rate so that we can carry out analysis, taking it into account any linkage errors that we have in the data set. So what are we looking for when we need to determine the matching variables? Well they have to be unique, they have to be available and known, accurate and stable over time.

So obviously we are conducting data linkage to carry out statistical research and to inform policy and the two main methods of data linkage and their combination is deterministic exact matching and probabilistic matching. So I will be going over both of these two types of data linkages. In deterministic matching this is based on exact 1 to 1 character match of the matching variables. In probabilistic matching these are based on partial identifiers which may be available such as names and addresses and a score is computed for each potential pair based on the individual probabilities of agreement for each matching variable. Deterministic linkage or exact matching this is where the records into datasets must agree exactly on the matching variables in order to conclude that they correspond to the same individual. And usually this is done when we have a high-quality identifiers such as an ID number. Now what happens here is that all matching variables have the same weight associated to them so for example matching on gender would carry the same weight as matching on the last name. So this will be quite different then in the probabilistic linkage. Now even in deterministic linkage we can incorporate some errors for example in fuzzy matching this is the exact matching carried out with a a wild card or a set of substring for example a wild-card of a *a *a can be any number of words such as banana and pyjama . This method for example is used in search engines on the internet or we can transform the data such as using a phonetic code such as soundex for the names or we can truncate the names for example the first three or five letters of a name which must match exactly. So there are ways to incorporate errors in deterministic linkage the important thing to remember is that each character

has to agree exactly and then there's no weight associated to the matching variables.

In contrast what is probabilistic data linkage? This does not require that all identifying fields match exactly in order to be able to conclude that the records belong to the same individual. Basically we carry out of frequency analysis of the data values necessary and in order to calculate for each matching variable a wait or score and this indicates for any pair of Records how likely it is that they refer to the same entity. Now uncommon value agreements would give stronger evidence for the linkage. Large weights will be assigned to fields that match and small weights are assigned to field that don't match. And then we would sum the scores overall the matching variables, compare the sub to a threshold and from there determine whether the pair should be declared a match a non-match or if we are undetermined or undecided we can send to clerical review.

Now in probabilistic data linkage the method relies on calculating scores and these are based on probabilities. This determines agreement between the matching variables between the pair of records as well as the disagreements. So either from a previous experience of record linkage on a similar application or perhaps we take a preliminary linkage exercise, produce some sort of gold standard linkage, we need to calculate how likely it is that the variables that do agree between the pair would have done so by chance or if the pair were not correctly matched. And this is compared to how like the agreement would be in correctly matched record pairs. Now we can also use latent modelling, latent class modelling, EM. algorithms to estimate the matching probabilities without the need for a previous experience of linkage or any test data but this is a topic for another day.

Now probabilistic record linkage is more computationally demanding and more difficult to program but it reduces the number of overlooked matches by being able to model the inconsistencies in the data and taking them into account.

So what are the criterion for good matching variables? We need to have the agreement between variables which are more typical of correctly matched pairs rather than those that might have occurred by chance in unrelated records. So for example variables that might agree by chance an unmatched record pairs are those which don't divide the population into many subclasses for example gender that would be a fifty percent chance

of having a correct match on gender for example. The key technical issues in the development of data linkage procedures are good quality identifier that are available to discriminate between the person to whom the record refers and all other person's. Being able to decide whether discrepancies and identifiers are due to mistakes in reporting for a single individual and being able to process the large volume of data within a reasonable amount of computing processing time.

So there are three key parameters for a probabilistic data linkage and I will be going over each one separately. The first is the quality of the data, the second the chance that values of a matching variable will randomly agree and finally the third the ultimate number of two matches that exists in the database. So not all fields for matching variables give you the same amount of information and uncommon value agreements should show stronger evidence for linkage. To incorporate this discriminating power of the matching variables the weights are computed as a ratio of two frequencies which I will then translate into probabilities.

The first is the number of agreements of a field in record pairs that represents the same individual. And the second frequency is the number of agreements in a field and record pairs that do not represent the same individual. In order to determine these agreements and disagreements we need to define something called an agreement pattern which I'm denoting here by a lambda. For example three matching variables with binary comparison test whether the pair agrees for example on last name or disagrees whether the pair agrees on first name or disagrees whether the pair agrees on street name or disagrees. So a simple agreement pattern in the case of three matching variables for example would be 101. The pair agrees on last name, disagrees on first-name and agrees on street name and in fact for three matching variables there would be eight such comparison vectors.

Now agreement patterns might be complex and they might be based on string comparators or not necessarily binary agree/disagree. For example we might have a 0.66 percent proportion in agreements on last name due to a string comparator. Now the first parameter that i mentioned is based on the data quality. Data quality is the degree to which the information contained for a matching variable is accurate and stable over time. So data entry errors missing data false dates obviously this diminishes the accuracy and produces low quality data. The higher quality data the more likely we are to be able to make correct match. So data quality the first parameter is

reflected in one of the probabilities needed for this process. And in the Fellegi-Sunter framework of 1969 they define this as the M probability. The M probability is the conditional probability that a record pair has an agreement pattern gamma given that it is a match in other words the same person and we write the M probability as the conditional probability as you can see in the notation there. Now this is approximately 1-the error rate the computer science literature might refer to it as the reliability. So how much the matching variable has errors associated to it due to data entry errors, missing data etc.

The second parameter depends on the number of random agreements and this is denoted in the Fellegi-Sunter framework as the U probability. More formally the U probability is the conditional probability that a record pair has an agreement pattern gamma given that it is not a match. So you can see that the in the notation the condition is on no match not a match. The third parameter is the overall number of matchers or potential matchers in our datasets and this is denoted as the probability of M, the probability of match. Now the parameter or the probability of interest of course is the match probability this is the probability of a match given and observed agreement pattern. So according to the basic theorem we can calculate that using the formula as shown on the slide. The probability of a match giving agreement is equal to the M probability which is the probability of agreement given a match times the probability of match / the probability of agreement. This matching probability is based on something called a likelihood ratio. This is called the likelihood ratio is based on the agreement likelihood ratio. In other words the ratio of the M probability / the U probability. Now Fellegi-Sunter assumes conditional independence. This means that the errors associated to one particular matching variable is independent to errors associated to another matching variable which is quite a strong assumption and in that case the comparison vector that we defined on the previous slide can be decomposed into its separate components. So the likelihood ratio is a ratio where on the numerator we have the M probabilities for each one of the matching variables separately and in the denominator the multiplication of the U probabilities for each matching variable separately. So the likelihood ratio is our overall score our test statistic and we can order these comparison vectors by the agreement ratio R of gamma and choose thresholds and upper and lower cutoff values to determine the correct matches and the not correct matches. Now it's a little hard to multiply probabilities together and in fact the framework of the Fellegi-Sunter takes the log of the likelihood ratio and therefore instead of multiplying these ratios we add them up we can sum them by taking the log

it can be any log transformation but we take the log of the M / U for the first matching variable plus the log of M over U for the second matching variable etc. etc. in order to produce an overall score. And this is what in essence is what the probabilistic data linkage is doing.

So here's an example let's assume that I have prior test data and I give you the M probabilities the probability of agreement on a particular characteristic x given it's a true match. Now obviously these are quite high 0.9 if its first and last name age, 0.8 if it's house number street name so little lower quality for house number street name but nevertheless our M probabilities are generally high we do expect our data sets to have high quality and minimal errors associated to the variables.

The U probability the probability of agreement on a particular characteristic x given it's not a match is 0.1 if its first name last name age and 0.2 if its house number and street name. So you have a set of M probabilities, you have set of U probabilities, there they are again on top of this slide the M probabilities and U probabilities. I am now going to put together all potential pairs in my two databases and here's a particular record i have a Samantha smith and Sam smith names, I have address both of them have 435 Main Street, birth year 1954 for the first, 1955 for the second record. One is the mail and one is a female. So what is my agreement vector now , my comparison vector we can see here that there is a disagree on first name, there's an agree on last name, there's an agree on house number, there's an agree on a street name. There is a disagree however in the birth year and a disagree in sex. So by comparison vector in this case would be 0 1 1100. So now i have to put together the log in this case I’m using the natural log of the likelihood ratio. Now if there's a disagreement instead of the M / U, we take 1- M over 1 - U as you see that they're in the first term there is a disagree on first name so we take the natural log of 1 - 0. 9/1 - 0.1. We have an agree on last name so that is the log of 0.9 / 0.1. We have an agree on house number which is the log of 0.8 / 0.2 etc. etc. For each potential pair we can now calculate an overall score which will be used to determine the cut-off for determining the match status. In this case we got a -0.81. Now you can probably thinking to yourself we can probably improve on this algorithm. For example we can use a dictionary and string comparator metrics which might give partial agreement weight to Sam and Samantha perhaps Sam is a nickname for Samantha. We also see a deviation of one year in the birth year and perhaps that might be sensible to think of that as a partial agreement instead of a disagreement and so there are ways to incorporate dictionary, string comparators etc to give partial

agreements to matching variables.

So data quality as i mentioned is quantified by the M probability with respect to the accuracy and stability of the matching variable. And for any given field, any given matching variable the same value for M probability applies to all records it doesn't matter what the value is, it's all about the error of the quality of the data in that particular matching variable. But you can see that the U probability this is where the distinguishing power the discriminating power is. This can be obtained for example that simply by thinking that the probability of the two records randomly agree i gave you the example of gender which has a random agreement of one out of 2, a 1 out of the number of values. Month of birth for example would be 1 out of 12, age perhaps could be 1 out of 100 and we can think of the U probability as just the overall probability of a random agreement.

Now in contrast to the M probability a matching variable may have multiple values of U probabilities each corresponding to specific value in the matching variable. So for example the U probability for last name perhaps you might want to give more weight if there's an agreement on a name such as the Brewski compared to Smith. So the U probability typically is estimated by the proportion of records with a specific value based on the frequency seen and say a large primary data source.

And I can't end the probabilistic data linkage without discussing a notion called blocking. Now there are number of possible comparisons increases with the product of the file sizes so for large files let's say I have two files each of size 10000 that produces a hundred million comparisons that we need to look at and produce comparison vectors and overall scores. So what we do in data linkages that we restrict the comparisons two blocks of data where one or more variables need to match exactly. So now you can see are introducing exact matching into the probabilistic matching framework and the idea is to institute exact matching on a blocking variable which are likely to refer to the same person and therefore we reduce the amount of time that we need searching through the file through the searching through the pairs. So we utilize in a deterministic approach to help us with the probabilistic method of record linkage and we can block sequentially so in a typical framework of data linkage which would be a 1 to 1 match let's say we have a post enumeration survey that needs to be matched back to the census file typically that's done sequentially and iteratively using different variables for the blocks. This reduces the amount of pairs that we need to look at because only the potential pairs that match

on a blocking variable will be produced.

Deterministic matching for example we could block on postcode and surname carry out the data linkage carry out the clerical review on the set of designated potential matchers and then we put our match dataset aside and we proceeded to another iteration through the residuals of the two files that were not matched perhaps with another blocking criteria such as year of birth and this is the way we carry out the linkage using iterative process interchanging blocking variables and matched variables.

So once we have our overall scores we need to determine thresholds and these thresholds determine the match status of whether we will determine them as a true match not a two-match or we're not decided and we need to send clerical review. So as in classic decision theory in statistics these decisions are thresholds. They are determined by minimizing two errors the type one error and the type two error. The type 1 error in the framework of the Fellegi-Sunter record linkage is the error of linking unmatched records. So we put together a pair we said these are a match we put them in the match dataset , these are not ready, the matched dataset is ready for analysis but we have errors the type 1 error we have pairs that should not have been matched. The type 2 error is the error of not linking match records so we have still in our residual datasets potential pairs that we did not find and as I mentioned since record linkage can be an iterative process we might find them again in a subsequent pass through the data.

Now as in classic statistical theory these thresholds are determined by how much you are willing to be wrong based on these two error types and these are predetermined by you as the data linkers so then you determine how much you willing to be wrong and the type 1 error type 2 error and this will determine the cut-off of the thresholds. Off course the high values of the overall score suggests the correct match. The low values of the overall scores would suggest an incorrect match and in between we might be undecided and we'll need to carry out some clerical review.

So what constitutes high and low? Well! suppose we have a frequency distribution determine the critical values of high OR low values based on these levels of significance the type 1 type 2 error, how much you are willing to be wrong so perhaps I have some training data again where we know the true match status and some gold standard data set and we can derive these distributions for the two matchers and the two non-matchers the lower distribution of course are the non-matchers because they have

the smaller weights and they are also very large group because there are lots of potentially not matcher and this is where we need to look at the the upper tail to determine the error type and to determine the threshold and of course the matchers would have the upper distribution of the higher weights.

So in this slide is a figure representing these distributions of the non-matchers and matchers assuming that I was able to make them into a nice normal bell-shaped curves. So you see there are mixture on the lower end the small overall scores you see those are the non matches and on the right-hand side where the scores are high you see the matchers and you can see of course that the curve is quite large for the non matches there are quite a few more non matches than there are matches. And the idea of courses is to choose thresholds based on your type 1 , type 2 errors. So we're looking for two thresholds a score which I denote by w+ above which we will automatically classify the pair as a correct match and a w- below which we automatically classify the pair as an incorrect match and in between the W- W+, we would need to carry out a clerical review. So i'm going to now zoom into this figure in the middle there looking at that exactly where that mixture is occurring between the non matches and the true matches in the next slide.

So the decision rule is based on three types of pairs those that we believe to be a correct match those that are we believe not to be a correct match would be on the left-hand side and in the middle where that mixture is occurring you can see there the tails of the matches and the non matches coming in from different directions as i zoomed into the mixture part and we can look at the preset our type 1 and type 2 errors, on the non matches the upper tail would be the type 1 error as you can see that's pointed out with the error and on the coming from the matches on the lower end of the tail of the matches is the type 2 error and these are preset right! You are the data linker, you preset how much you're willing to be wrong and this is what derives our cut-off thresholds as you can see by the vertical lines in the mixture. So anything above the w + is considered a match everything below w- is considered the non-match and you can see in the figure what the errors are in each of these decisions based on the type 1 type 2 error and all of those in the middle between w- and w+ will be sent off to clerical review and manually reviewed to determine their match status.

So after record linkage once we have carried out our decision rule we should check for errors, we check out carry out perhaps logical checks in

the data for evaluation using other key variables not necessarily those that are matching variables. So for example if you're matching death records with a hospital discharges you would want to make sure that you don't have any hospital discharges after a death lots of ways to use logical statistical data editing and logical checks to make sure that your datasets are having little errors and no errors in them and obviously the poor quality data we can get errors. What is recommended of course is those pairs that are declared a match even if you're doing an exact in deterministic matching you have your matched data set this is what you're going to use for your analysis which you know could be for informing policies. it is are very important to understand if there are any errors in that linked dataset, carry out a small random sample, check for the accuracy in the matching status, particularly for those that might be close to the threshold cut-off values. And on those pairs that were not matched also carry out a small random sample check for the accuracy in the match status and of course these errors can be used to compensate for linkage errors when you analyse your linked data using classical measurement error framework measurement error models but again this would be a subject for another talk.

So just to look at what our decision or dispersion matrix would look like in any decision theory and classical statistics, we would have our columns defined as the true status the null hypothesis versus the alternative hypotheses. In this case the column of the null hypothesis is the non matches they're not matched and the alternative hypothesis are the matches. On the rows, we have our decision we had a test statistic, we had an overall score based on the likelihood ratio and based on our criteria, our thresholds we determined and decided whether they were a not linked pair we did not put them together as matches in other words we failed to reject the null hypotheses versus we linked the pairs right we decided they are a match and rejected the null hypotheses. And now on the diagonal of our decision matrix you can see that we made the correct choice so if we did not think the pairs and indeed they were not matches great! We made the right choice not linked on matches this is also known especially in epidemiology studies as the true negative. On the other diagonal we have the linked matches made the right choice also known as the true positive where the type one error and type 2 error comes in are the off diagonals. So if the true status was a non-match and we linked them together as i previously mentioned this is the type 1 error is also known as the false positive we linked not matches. And the other half-diagonal we have not linked matches this is the type 2 errors I previously defined also known as the false negatives.

Now what you can see here outside of the matrix is a various quality your evaluation of parameters that should be calculated by data linkers and delivered to researchers. For example the proportion of the false positives, the proportion of the false negatives these are very important criteria to understand the quality of our linkage process. The 1- the false-positive rate is also known as the specificity and a sensitivity or in computer language and computer science literature is also known as Recall. And also in the computer science literature you find another measure which is called Precision. Now noticed that precision is actually the number of true positives out of the total linked pairs. So in the computer science literature you'll find Recall and Precision in order to evaluate your linkage process.

So what is the overall process? We first need to select our matching variables and our blocking variables, we need to edit, parse, produce some code, the string comparators standardized believe me the pre-processing stages is the most cumbersome and the most work to carry out in a data linkage. Once we are satisfied that both datasets have high quality and are consistent and standardized together we then block in sort both files. Now first we carry out a deterministic method if you don't have an ID a unique ID you might want to concatenate the matching variables, carry out a deterministic method perhaps you'll find some correct matches. Those that are matched put them into the match dataset, those that are not matched we carry out a probabilistic method. Again using probabilities and agreement ratios based on test data, determine the thresholds and match the two datasets. Now off course we could have good matches, some have to be sent to the undecided clerical review and then those are sent to the true matches and once you have the true matched file, in order to carry out your analysis you need to make sure that there are no errors check for errors check for the logical error in the data set and then use those to correct for your statistical analysis and the measurement error model and you can go ahead and use the dataset for informing policy research whatever the reason is for linking the data.

Thank you very much for your attention.