TRANSCRIPT for 21_WP05_01_PRESENTATION_COMPARABILITY_VANBUUREN.mp4

Comparability is a Missing Data Problem

Hi my name is Stef van Buuren, and I am the lead of the work package 5, Statistical methods for individual patient data. In this lecture I would like to tell you something about the way how you deal with problems related to comparability of data, and how the missing data perspective could help. So first look at why we want want to use individual patient data.

Why individual patient data

So why do we want to combine the individual data and analyse the individual patients (IPD)? Well, the first thing is that it gives us the possibility to disentangle effects at that level of the study from the effects related to the study. IPD also gives you the possibility of studying effect modification, which would not be possible in meta-analysis. We may adjust for confounding variables, try to improve standardised definitions and analyses across different studies. We could also try to obtain complete follow-up the data by combing studies, analyse outcomes that may not have been sampled in one particular study, but others may have done so, for we could try to see, for example for quality of life, whether we would be able to analyse multiple outcomes. Combining of course is also useful if we have rare events. If you combine data then we have more power to do all good things, but it doesn't it doesn't come for free.

Why not individual patient data

I have also a slide “why not individual patient data”, since it is not so easy to create individual patient data because there are a lot of practical problems that you need to think about when you are combining data from different sources. For example, studies may collect different variables that measure the same thing. I will give you the example briefly where this is the case. Then you may wonder how to combine information. The timing of the measurements differs between cohorts. Different studies have different criteria for inclusion of patients participants. There may also be missing data, but those missing data may be missing for different reasons in different studies, so we need to think about those issues also. When we have data on same person from different sources, then sometimes the key the links persons is missing or incomplete, and we need to think about the issues that it could create. And, of course, the original data have been collected for different purposes, so what to do? Can we use the data for other purposes - we hope so - because we already have data, but of course the original investigators have made choices. Also if you are combining data, especially data that belong to the same person, then there is a risk of identifying the person from the combined data. So these are privacy issues that you also need to think about. Then there's a thing that classifications may change overtime for example, the ICIDH classifications that may change over time, and so we need you may need to have a strategy to go from one classification to the other. Access to the original study maybe restricted, so you need to think about in order to get into good things of individual patient data. I want to just study the first issue in this lecture, so the problem that we have two or more studies and that use different variables to measure the same thing. So let me give you an example.

An example

So here is an example of two countries, I've called them Antonia and Belmark, collecting data. Both these questions are about how well you can walk, and obviously the interest is in measuring some way how well the population in those countries can walk. If you look a little bit closer, the information that's being collected in Antonia then the item is actually the HAQ8, and that is “Are you able to walk out on flat ground?”. There are four response categories: without any difficulty, some difficulty, with much difficulty, unable to do. There's also some missing variables. In total we have 306 respondents in that survey. In Belmark there's a different item (GARS9), that looks like HAQ8, but it's different. So “Can you fully independently work outdoors if necessary with a cane?”. Also here we have four response categories: yes no difficulty, yes with some difficulty, yes with much difficulty, no only with help from others. So now suppose that you want to compare, based on these two variables, walking disability between Antonia and Belmark. Then, of course, we need to have some way of comparing them. In Antonia it is usual to calculate just the mean of the distribution. This mean, in this particular case, is equal to .24. In Belmark, it is more usual to calculate the proportion of people that have no difficulty. So here 145 divided by two 292, it's about half, so 50% is their benchmark. So, what is the problem now that we want to address?

Problem

The problem that we want to address is to compare walking ability between Antonio and Belmark, but the data differ in two respects. The questions are different between Antonia and Belmark, and also the statistics are different between Antonio and Belmark. Antonio use the mean, and Belmark used the PND. Well, how to solve that? Let's first look at a straightforward solution. The easy way out would say “well let's assume that the categories are actually working the same way”. So we say that category zero in HAQ8 is the same as category zero on GARS9. We do that for all categories, so that's easy to do. The big advantage is that now, under this assumption, we can calculate the PND for Antonia, and the mean for Belmark. Let's do that.

Easy way out, equate categories

If we do that, then we see that the PND is .8, so 80%, which was 242 / 306 is .80. So 80% in the category, whereas we had 50% in Belmark. So there is a 30% difference between the two, in this example, which is quite large, so 50% or 80%. Now let’s look also at the other statistic, so if you calculate the mean. Then for Belmark it would be .66, and this is on a scale from zero to three. So if you compare that to Antonia, which was .24, that's also a big difference on that scale. So, we see that both a statistics tell the same story: Antonio is better doing on both variables and by a last margin. Now what is the problem here?

But what are our assumptions?

Well, we need to be aware that we have made an assumption, and what exactly is that assumption, and is that assumption valid? Now think of these two variables as making up a contingency table. So we have the HAQ8 variable, which is of one way of measuring walking ability, and we have to GARS9 item, which is another way of measuring walking disability.

The number of cells is 16. What we are assuming is that the only counts that are allowed are on the diagonal. So that's basically the assumption that we make if we equate those categories. But is this a valid assumption? We don't know without any data. But now suppose that there is a third country for which we do have those data.

So this third country, called Citrus (which starts with C). And this is actually real data that we're looking at, it's not a made-up example, it's real data. Citrus was in a position to ask both questions to the same people. Then we can actually create this contingency table of HAQ8 and GARS9. What do we learn from looking at the numbers in this table? First of all, most of the numbers are on the diagonal, which is good because that's close to our assumption. But if you look little bit closer, then we see that its distribution is not symmetric. In general, the HAQ8 appears to be a little bit more difficult than GARS9, so there are more observations in the upper right corner than in the lower triangular corner. So there seems to be a systematic difference between those two items. Moreover its smaller on the diagonal and it's more spread. What are the consequences? If we would say "well it's clearly not diagonal this contingency table", what are the consequences of making that assumption for the comparison that we're making between two countries?

In order to say a little bit more, what we would like essentially is also to calculate the two statistics – the mean and PND – when we assume that the relation between HAQ8 and GARS9 in Antonia and Belmont is the same as in Cyrus. So that's what we're going to do now. And I would say that that this is probably a better assumption, a more plausible assumption, that assuming the thing is diagonal. So let's look how we do that.

Let's combine all data

Well the first thing that we need to do is to combine all the data. So we make a large data set which has the data of Citrus, Antonia and Belmark, all stacked under each other. So these are different countries, different people. For Citrus, we have observations for both hawk aid and part 9 so these are both blue in this in this this data set for Antonia we only HAQ8 and GARS9, so there are both blue in this dataset. For Antonia, we don't have GARS9, so it's missing. And for Belmark, we have GARS9 but we do not have HAQ8, so that's missing over there. And there are six observations in Antonia that have missing HAQ8 and GARS9. So, in this way we have brought together all information that we have into a data matrix. And now the idea is that we can fill up these red parts with things that are plausible, that we can learn from the other parts of the data. So, in the top rows we do have actually the information available on the relationship between the HAQ8 and GARS9, and we're going to extrapolate that information into the other cells. So, how do we do that?

Multiple imputation in MICE

We use multiple imputation for that. In short, multiple imputation is a way to generate synthetic, plausible values that act as replacements for the information that were missing. Let's look briefly how it works. We start from an incomplete data sets, in the way that I just showed you. What we now going to do is to fill up the red cells with imputed (synthetic) values that act as replacements. I can talk for hours how to do that, but I won't do that right now. Because we are not certain about what to impute, because that information is missing,

we need to do it not only once, but we need to do it multiple times. In this case, three times. So we complete the data sets three times, then we calculate our statistics of interest three times, and then we have three statistics and average those. And we can also calculate the confidence intervals around those statistics. That's theory that has been developed by Donald Rubin. I've been working to implement it in the mice package, which is now a popular way of imputing missing data.

Flexible Imputation of Missing Data

If you're insert more interested, you can look into the book that I've written on this subject. Now let's look at the results.

Results

If we look at the results for Antonia and Belmark, what comes out of, I've made the distinction between the equating method, which is the first method that assumes that every category is the same, and the imputation method, where we try to extrapolate the relationship between the two items to other parts of the data matrix. And of course for Antonio we had observed data with the mean of .24 and under imputation it is also .24 because it's observed. So there's no difference between them. But for Belmark we didn't have the mean statistic. For Belmont we can calculate it as 0.66, and the difference between those two is .4, so that's what we saw before. If you do the imputation, then of course the .24 of Antonia doesn't change, but Belmark changes because these are synthetic values now. It changes systematic actually. Instead of 0.66, it is now .45, so much smaller. So the difference that we previously saw of .4 is now about .2. We can also look at the proportion of persons that that walk well, that score in the zero category. For Belmark it was about 50%. If we equate the categories is about 80% for Antonia, so that's a 30% difference between those two. So both PND and the mean, based on equating, would say that Antonia is the more healthy population group for walking ability. If we do the imputation stuff, then of course the .50 is the same because that was observed. But the thing that is sensitive to our assumptions is this number. So instead of .80 (we thought that 80% would be in the zero category), if you do the imputation it's only 53%. So that's a big difference. And now the difference between those two countries only .03.

So what do we conclude from this? Well Antonio is still doing better but the differences are much smaller, and I would say more realistic, because our assumptions are based on actual data that come from Citrus. What we also see is that the imputation has more effect on the second statistic, the proportion of no difficulty. It is 10 times a small! So previously it was 30%, now it is 3%, so that's a huge effect. You can read more about this example in the link below.

Conclusion

So to conclude what you've seen. If you would do the simple equation thing relative to the imputation assumption, we see that simply creating exaggerates differences between countries. Exaggeration is not a good thing because overstated differences may induce interventions that are not appropriate, and sometimes these interventions could be very

expensive. So we better think about the assumptions that were making any consequences that they could have on the result. I've now treated the relatively simple problem where the first variable had four categories, and the second variable had also four categories. So it's kind of obvious to try to equate them. In many cases the number of categories differ and then it's not clear whether you should downcode it to the lower number of categories, or do something that you can go from the low number of categories to the higher number. So these are new problems and also the assumptions tend to be become more impactful.

In those cases, in order to make progress and do scientifically defensible analyses, I would suggest first trying to identify overlapping information between instruments. So of the type of information that we saw from Citrus study, or a country that had actually both observed, so we can say something sensible about the relationship between those variables. Then organise this is a missing data problem, put it into mice, apply multiple imputation, calculate your statistics and then you get a result based on the actual information that you do have rather than having an external unverifiable assumption everything is equal.

References

If you want to know more about this you can read in more detail Chapter 9 of multiple imputation book. There's also an online version. Of course you can buy the book so that I get rich by selling millions of copies, but the online version has exactly the same information and that's for free. If you want to dive in more deeply into the problem, I point to a previous lecture, Warsaw lecture, which also has more slides. If you want to recalculate this example, you can do it because the script is part of the book. Thank you for your attention and I hope it will this will give you some new insights.