Identification of the risk of overlap in meta-analyses

UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

Zhentian Zhang[1] **Tim Friede**[1] **Tim Mathes**[1]

[1]University Medical Center Göttingen, Germany

Motivation

- There is a risk of study sample overlap in a meta-analysis, when multiple studies in the meta-analysis are based on similar pre-existing data.
- Sample overlap potentially leads to increased type one error and distorted representativness of the result.
- In meta-analyses of observational studies including registry-based studies, the analyses are usually based only on aggregated data, and for such cases there is no systematical solution to handle sample overlap.

A theory for describing overlap

Overlap is an multivariate relationship. In contrast to bivariate relationship such as correlation, it cannot be fully represented using a single matrix. Figure 1 gives a simplified graphical explanation of it.

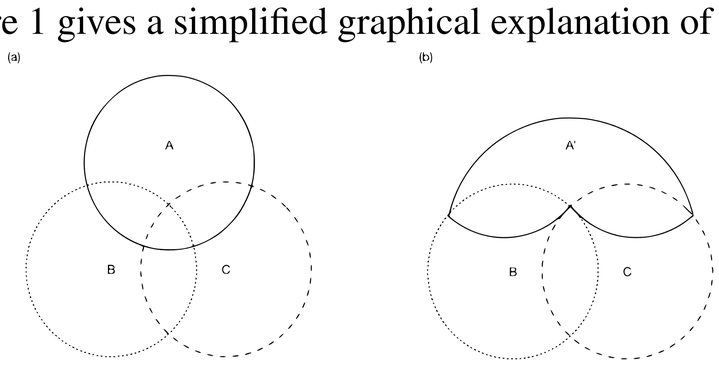

Figure 1: Overlap is not a bivariate relationship.

We developed a theory for describing and estimating the overlap structure in a meta-analysis. The theory could work as a basis for further steps of overlap handling, namely the evaluation of its impact and the correction of the result.

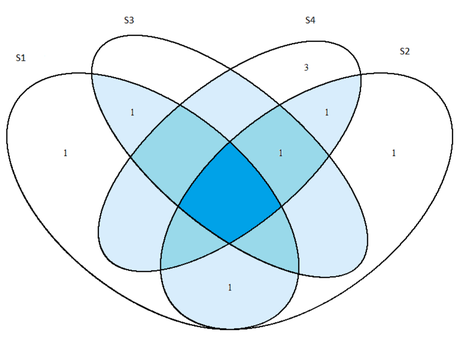

Figure 2: An example of overlap structure among four studies

We suggested a value to quantify the degree of overlap, the risk of overlap $\in [0, 1]$. It can be calculated for all subsets of the set of study samples of a meta-analysis, and only aggregated data is required for the calculation. 0 means there is no overlap among the sample of studies in the set; 1 means all samples in the group has the same ranges of key characteristics. Figure 3 shows our approach of visualising the risks of overlap.

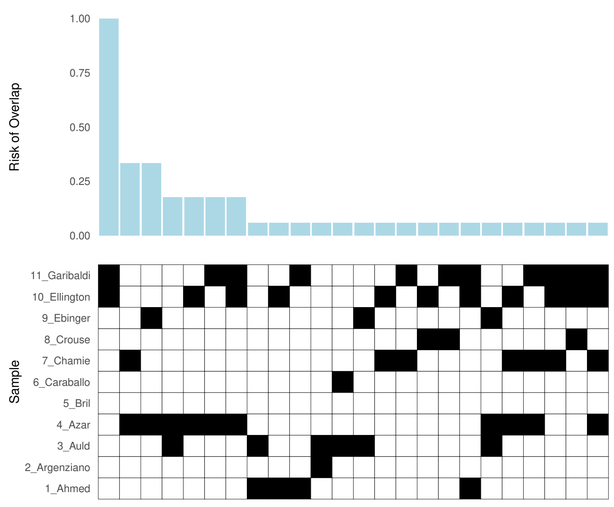

Figure 3: Visualization of the risks of overlap

Calculating the risk of overlap

Step 1: Collect the ranges of characteristics for each study sample.
Step 2: Choose the key characteristics.
Step 3: Vectorize the key characteristics.
Step 4: Calculate the risks of overlap.

code	1_Ahmed	2_Argenziano	3_Auld	4_Azar	5_Bril	6_Caraballo	7_Chamie
UK birmingham	0	0	0	0	1	0	0
UK Leicester	0	0	0	0	1	0	0
UK london	0	0	0	0	1	0	0
UK other	0	0	0	0	1	0	0
USA alabama	0	0	1	0	0	0	0
USA california LA	0	0	1	1	0	0	0
USA california other	0	0	1	1	0	0	0
USA california San francisco	0	0	1	1	0	0	1
USA Connecticut	0	0	1	0	0	1	0
USA Florida	0	0	1	0	0	0	0
USA georgia	0	0	1	0	0	0	0
USA illinois	0	0	1	0	0	0	0
USA Louisiana	0	0	1	0	0	0	0
USA massachusett	0	0	1	0	0	0	0
USA michigan	0	0	1	0	0	0	0
USA new york	0	1	1	0	0	0	0
USA ohio	0	0	1	0	0	0	0
USA other	0	0	1	0	0	0	0
USA pennsylvania	0	0	1	0	0	0	0
USA Texas	0	0	1	0	0	0	0
USA Utah	1	0	1	0	0	0	0
Community	1	0	0	1	0	0	1
Hospital	1	1	1	1	1	1	0
sample_sizes	17527	1000	217	14035	426	181	3849

Figure 4: An example of the result of step 3

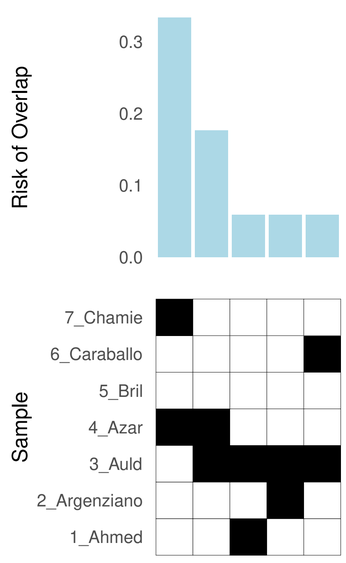

Figure 5: An example of the result of step 4

Utilisation of risk of overlap

One application of risk of overlap is to calculate the set of overlap-free samples with the largest sample size, following the steps below.

1. We find the set B_0 of all the sets of study samples whose risk of overlap is zero.
2. In B_0, we keep only the sets that have all their subsets also in B_0, and denote the new set as B_1.
3. In B_1, we keep only the sets who do not have a proper superset in B_1, and denote it as B_2.
4. Choose the set in B_2 that has the largest sample size.

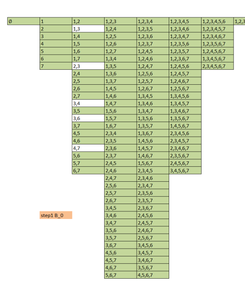

Figure 6: Step 1

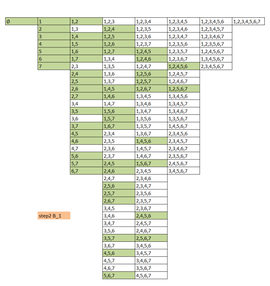

Figure 7: Step 2

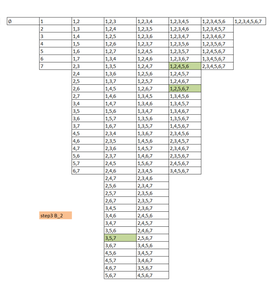

Figure 8: Step 3

Application on real world meta-analysis

We applied our method to several published real world meta-analysis. For example, in one meta-analysis that investigated the impact of ethnicity on clinical outcomes in COVID-19 [1], we calculated the overlap-free sample group with the largest sample size. The result is a set of 5 study samples (out of 51 tudies in total), which are marked in the figure 9. Although we only have 5 studies left, it has a total sample size of 14,317,870 compare with the original total sample size of 17,211,742 from 51 studies. In other words, we obtain a overlap-free meta-analysis population at the cost of only 16.8% sample size reduction.

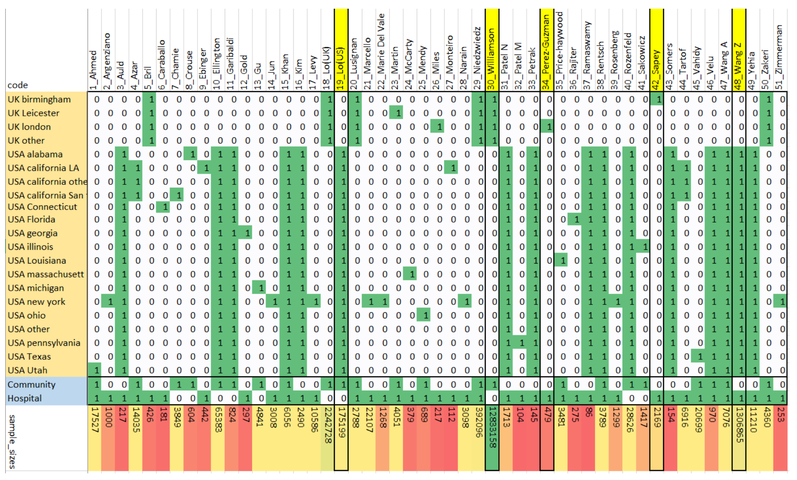

Figure 9: The result of vectorization, and the choice of overlap-free set of study samples with the largest sample size

Conclusions

We suggested a quantity ”risk of overlap” which indicates the degree of overlap . Our approach transforms the task of finding overlapping observations between study samples into identifying overlaps in samples’ characteristics, which is much more feasible in practice. We described the theory behind our method using a combination of common concepts in set theory and the coding of the basic information about the study samples. The method is viable even when only very limited information about the sample is given, for example when only the papers of the studies are available, which is often the case in practice.

We designed an algorithm based on ”risk of overlap” that find out the overlap-free set of samples that has the largest sample size, based on characteristics of the samples that are easily available. This set can be used for further analysis, which should provide valuable insight of the potential impact of sample overlap.

We applied our methods to existing meta-analyses in this paper, and were able to confirm the viability of our method to complex real-world scenario. The result also showed the necessity of such overlap analysis due to the high risk of overlap we discovered with the help of our method in the investigated cases.

References

[1] Sze S, Pan D, Nevill CR, Gray LJ, Martin CA, Nazareth J, et al. Ethnicity and clinical outcomes in COVID-19: a systematic review and Meta-analysis. EClinicalMedicine. 2020;29-30:100630.