Identification of the risk of overlap in meta-analyses

UNIVERSITÄTSMEDIZIN GÖTTINGEN **Zhentian Zhang¹ Tim Friede¹ Tim Mathes¹** ¹University Medical Center Göttingen, Germany

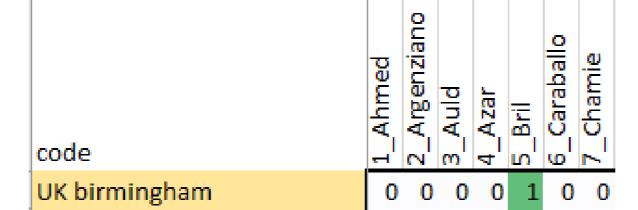
Motivation

- 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

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.



Application on real world metaanalysis

We applied our method to several published real world meta-analysis. For example, in one metaanalysis 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.

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.

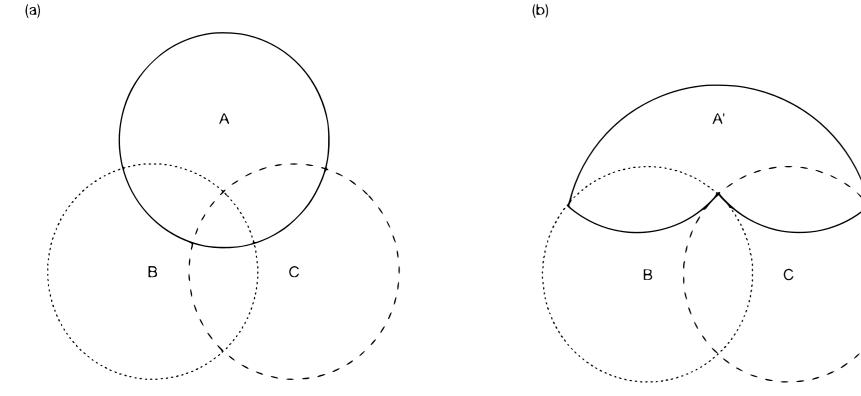


Figure 1: Overlap is not a bivariate relationship.

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 im-

0	0	0	0	1	0	0
				- -	v	0
0	0	0	0	1	0	0
0	0	0	0	1	0	0
0	0	1	0	0	0	0
0	0	1	1	0	0	0
0	0	1	1	0	0	0
0	0	1	1	0	0	1
0	0			0		
<u> </u>						0
0	0			0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	1	1	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
0	0	1	0	0	0	0
1	0	1	0	0	0	0
1	0	0	1	0	0	1
1	1	1	1	1	1	0
17527	1000	217	14035	426	00	3849
	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	0 0 1 1 0 0 1 0 0 1 1 1	0 0 1 0 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0	001000011000110001000010000100001000010000100001000010001000010000100000100010000100001000010000100001000101001010011110	00100001100001100001100001001001000001000001000001000001000001000011000001000001000001000001000001000001000101000101000111111

Figure 4: An example of the result of step 3

code			- 11		5_Bril	6_Caraballo	7_Chamie	8_Crouse	9 Ebinger	10 Ellington	11 Garibaldi	12 Gold		14 Iun	1 L LAON				18_Lo(UK)	19_Lo(US)	20_Lusignan	21_Marcello	22_Marie Del Vale	23_Martin	24_McCarty	25_Mendy	26_Miles	27_Monteiro	28_Narain	29_Niedzwiedz	30_Williamson	31_Patel N	32_Patel M	33_Petrak	34_Perez-Guzman	35_Price-haywood	36_Rajiter	37_Ramaswamy	38_Rentsch	39_Rosenberg	40_Rozenfeld	41_Sakowicz	42_Sapey		44_Tartof	45_Vahidy	Velu	47_Wang A	48_Wang Z	49_Yehia		51_Zimmerman
UK birmingham	0	0	0	0	1	0	0	0) () () (0	0	0		0	0	0	1	0	1	0	0 (0) () ()	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
UK Leicester	0	0	0	0	1	0	0	0) () () (0	0	0	0	0	0	0	1	0	1	0	0	1	L O	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
UK london	0	0	0	0	1	0	0	0) () () (0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
UK other	0	0	0	0	1	0	0	0) () () (0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
USA alabama	0	0	1	0	0	0	0	1	. () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA california LA	0	0	1	1	0	0	0	0) 1	1 :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	1	0	1	1	1	1	0	0
USA california othe	0	0	1	1	0	0	0	0) () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	1	0	1	1	1	1	0	0
USA california San	0	0	1	1	0	0	1	0) () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	1	0	1	1	1	1	0	0
USA Connecticut	0	0	1	0	0	1	0	0) () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0 0	0 (0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA Florida	0	0	1	0	0	0	0	0) () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0 (0 (0	0	0	0	0	1	0	1	0	0	1	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA georgia	0	0	1	0	0	0	0	0) () :	L :	1	1	0	0	1	1	0	0	1	0	0	0	0	0 (0 (0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA illinois	0	0	1	0	0	0	0	0) () :	Ľ	1	0	0	0	1	1	0	0	1	0	0	0	0	0 (0 (0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	1	0	1	0	0	1	1	1	1	0	0
USA Louisiana	0	0	1	0	0	0	0	0) () :	Ľ	1	0	0	0	1	1	0	0	1	0	0	0	0	0_0	0	0	0	0	0	0	1	0	1	0	1	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA massachusett	0	0	1	0	0	0	0	0) () :	Ľ	1	0	0	0	1	1	0	0	1	0	0	0	0) 1	. 0	0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA michigan	0	0	1	0	0	0	0	0) () :	L :	1	0	1	0	1	1	0	0	1	0	0	0	0	0 (0 (0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA new york	0	1	1	0	0	0	0	0) () :	L :	1	0	0	1	1	1	1	0	1	0	1	. 1	0	0 0	0_0	0	0	1	0	0	1	0	1	0	0	0	1	1	1	1	0	0	1	0	0	1	1	1	1	0	1
USA ohio	0	0	1	0	0	0	0	0) () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0 (1	0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA other	0	0	1	0	0	0	0	0) () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0 (0 (0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA pennsylvania	0	0	1	0	0	0	0	0) () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0 (0 (0	0	0	0	0	1	1	1	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
USA Texas	0	0	1	0	0	0	0	0) () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0 (0 (0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	1	1	1	1	1	0	0
USA Utah	1	0	1	0	0	0	0	0) () :	L :	1	0	0	0	1	1	0	0	1	0	0	0	0	0 (0 (0	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	1	1	1	0	0
Community	1	0	0	1	0	0	1	1	. () :	L	1	0	1	0	1	0	0	1	1	1	0	0 (1	L ()) 1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	1	1	0	0	0	1	1	1	0	0	1	0
Hospital	1	1	1	1	1	1	0			L (0		0	1	1	1	1	0	0	_	1	. 1	1	L 1	. 1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1
sample_sizes	17527	1000	217	14035	426	181	3849	604	2442	22000	024	167	107	J000 /9/1	8005	6056	2490	10586	2242728	175199	2788	22107	1268	4051	379	689	217	112	8605	392096	12833158	1713	104	145	975	3481	275	98	6845	1299	28626	1417	6912	154	6916	20699	970	7076	1306865	11210	4360	253

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



pact and the correction of the result.

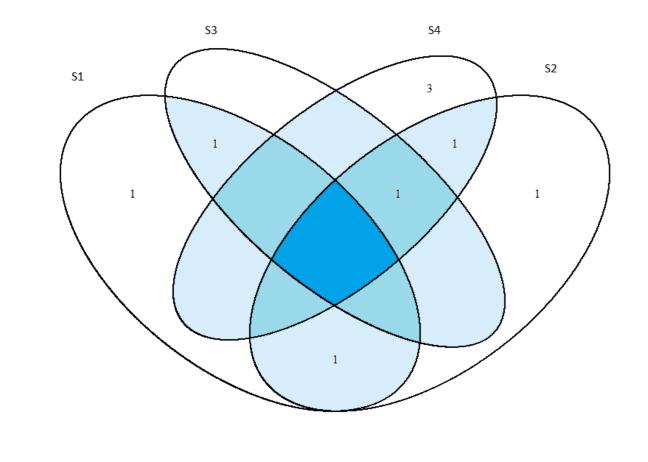


Figure 2: An example of overlap structure among four studies

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 metaanalysis, 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

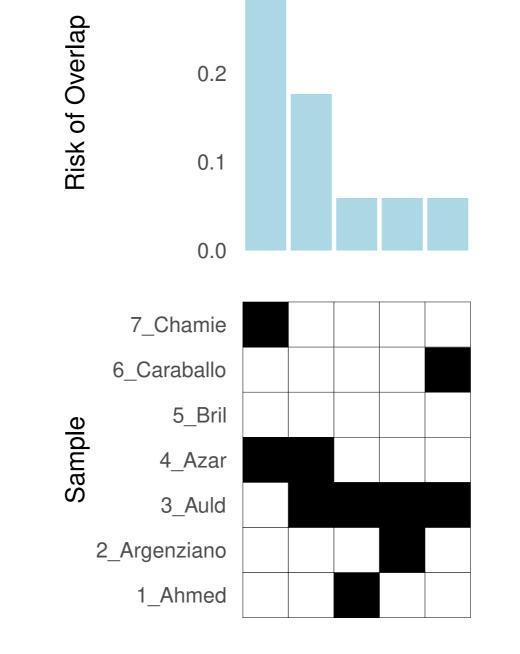


Figure 5: An example of the result of step 4

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.

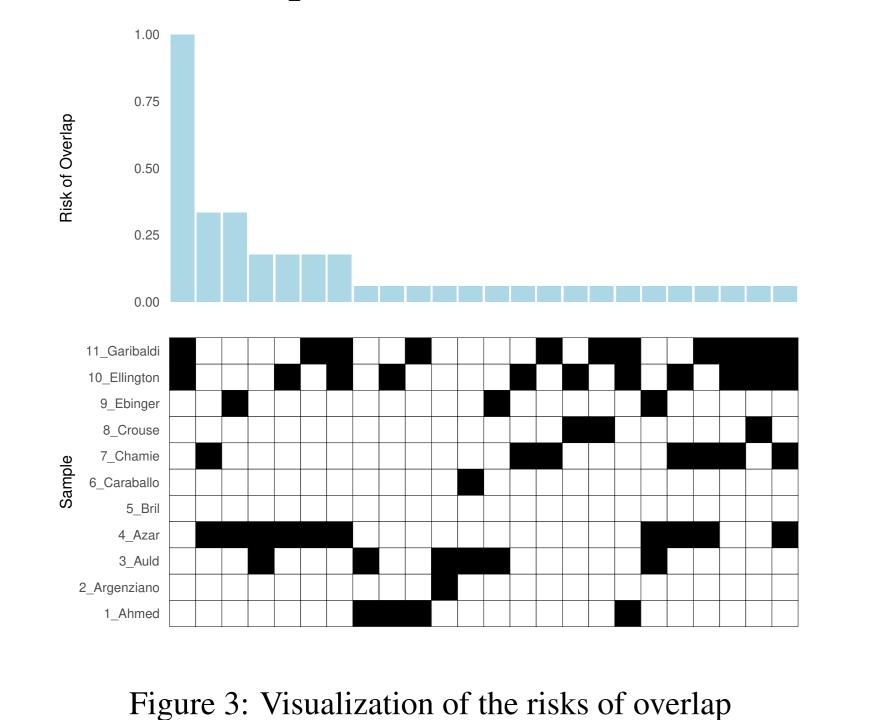
1. We find the set B_0 of all the sets of study samples whose risk of overlap is zero.

2. In B_0 , we keep only the sets that have all their

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.

the risks of overlap.



subsets also in B₀, and denote the new set as B₁.
3. In B₁, we keep only the sets who do not have a proper superset in B₁, and denote it as B₂.
4. Choose the set in B₂ that has the largest sample size.

1 1,2 2 1,3 3 1,4 4 1,5 5 1,6 6 1,7 7 2,3 2,4 2,5 2,6 2,7 3,4 3,4	1,2,4 1,2,5 1,2,6 1,2,7 1,3,4 1,3,5 1,3,6	1,2,3,4 1,2,3,5 1,2,3,6 1,2,3,7 1,2,4,5 1,2,4,6 1,2,4,7	1,2,3,4,5 1,2,3,4,6 1,2,3,4,7 1,2,3,5,6 1,2,3,5,7 1,2,3,6,7	1,2,3,4,5,6 1,2,3,4,5,7 1,2,3,4,6,7 1,2,3,5,6,7 1,2,4,5,6,7 1,3,4,5,6,7	1,2,3,4,5,6,7	Ø	1 2 3 4 5	1,2 1,3 1,4 1,5	1,2,3 1,2,4 1,2,5 1,2,6	1,2,3,4 1,2,3,5 1,2,3,6 1,2,3,7	1,2,3,4,5 1,2,3,4,6 1,2,3,4,7 1,2,3,5,6	1,2,3,4,5,6 1,2,3,4,5,7 1,2,3,4,6,7 1,2,3,5,6,7	1,2,3,4,5,6,7	Ø	2 3	1,2 1,3 1,4	1,2,3 1,2,4 1,2,5	1,2,3,4 1,2,3,5 1,2,3,6	1,2,3,4,5 1,2,3,4,6 1,2,3,4,7	1,2,3,4,5,6 1,2,3,4,5,7 1,2,3,4,6,7	1,
4 1,5 5 1,6 6 1,7 7 2,3 2,4 2,5 2,6 2,7	1,2,6 1,2,7 1,3,4 1,3,5 1,3,6	1,2,3,7 1,2,4,5 1,2,4,6	1,2,3,5,6 1,2,3,5,7	1,2,3,5,6,7 1,2,4,5,6,7			3 4	1,5	1,2,6				1		3				1,2,3,4,7	1,2,3,4,6,7	_
5 1,6 6 1,7 7 2,3 2,4 2,5 2,6 2,7	1,2,7 1,3,4 1,3,5 1,3,6	1,2,4,5 1,2,4,6	1,2,3,5,7	1,2,4,5,6,7			4			1,2,3,7	1.2.3.5.6	122567	-								- 1
6 1,7 7 2,3 2,4 2,5 2,6 2,7	1,3,4 1,3,5 1,3,6	1,2,4,6					c					1,2,3,3,0,1			4	1,5	1,2,6	1,2,3,7	1,2,3,5,6	1,2,3,5,6,7	1
7 2,3 2,4 2,5 2,6 2,7	1,3,5 1,3,6		1,2,3,6,7	124567			5	1,6	1,2,7	1,2,4,5	1,2,3,5,7	1,2,4,5,6,7	1		5	1,6	1,2,7	1,2,4,5	1,2,3,5,7	1,2,4,5,6,7	
2,4 2,5 2,6 2,7	1,3,6	1,2,4,7		1,3,4,3,0,7			6	1,7	1,3,4	1,2,4,6	1,2,3,6,7	1,3,4,5,6,7			6	1,7	1,3,4	1,2,4,6	1,2,3,6,7	1,3,4,5,6,7	
2,5 2,6 2,7			1,2,4,5,6	2,3,4,5,6,7			7	2,3	1,3,5	1,2,4,7	1,2,4,5,6	2,3,4,5,6,7			7	2,3	1,3,5	1,2,4,7	1,2,4,5,6	2,3,4,5,6,7	
2,6 2,7	127	1,2,5,6	1,2,4,5,7					2,4	1,3,6	1,2,5,6	1,2,4,5,7					2,4	1,3,6	1,2,5,6	1,2,4,5,7		
2,7		1,2,5,7	1,2,4,6,7	_				2,5	1,3,7	1,2,5,7	1,2,4,6,7					2,5	1,3,7	1,2,5,7	1,2,4,6,7		
		1,2,6,7	1,2,5,6,7	_				2,6	1,4,5	1,2,6,7	1,2,5,6,7					2,6	1,4,5	1,2,6,7	1,2,5,6,7		
		1,3,4,5	1,3,4,5,6	_				2,7	1,4,6	1,3,4,5	1,3,4,5,6	-				2,7	1,4,6	1,3,4,5	1,3,4,5,6	_	
		1,3,4,6	1,3,4,5,7	-				3,4	1,4,7	1,3,4,6	1,3,4,5,7	-				3,4	1,4,7	1,3,4,6	1,3,4,5,7	-	
3,5		1,3,4,7	1,3,4,6,7	_				3,5	1,5,6	1,3,4,7	1,3,4,6,7	-				3,5	1,5,6	1,3,4,7	1,3,4,6,7	-	
3,6		1,3,5,6	1,3,5,6,7	-				3,6	1,5,7	1,3,5,6	1,3,5,6,7	-				3,6	1,5,7	1,3,5,6	1,3,5,6,7	-	
4,5		1,3,5,7	1,4,5,6,7 2,3,4,5,6	-				3,7	1,6,7	1,3,5,7	1,4,5,6,7	-				3,7	1,6,7 2,3,4	1,3,5,7	1,4,5,6,7	-	
4,5		1,5,6,7	2,3,4,5,7	-				4,5	2,3,4 2,3,5	1,3,6,7	2,3,4,5,6 2,3,4,5,7	-				4,5	2,3,4	1,3,6,7	2,3,4,5,6 2,3,4,5,7	-	
4,0		1,4,5,7	2,3,4,6,7	-				4,0	2,3,5	1,4,5,7	2,3,4,5,7	-				4,0	2,3,5	1,4,5,7	2,3,4,5,7	-	
5,6		1,4,6,7	2,3,5,6,7	-				5,6	2,3,0	1,4,6,7	2,3,5,6,7	-				5,6	2,3,7	1,4,6,7	2,3,5,6,7	-	
5,7		1,5,6,7	2,4,5,6,7	-				5,7	2,4,5	1,5,6,7	2,4,5,6,7	-				5,7	2,4,5	1,5,6,7	2,4,5,6,7	-	
6.7	2,4,6	2,3,4,5	3,4,5,6,7					6.7	2,4,6	2,3,4,5	3,4,5,6,7	-				6.7	2,4,6	2,3,4,5	3,4,5,6,7	-	
	2,4,7	2,3,4,6		_				-7-	2,4,7	2,3,4,6						-7-	2,4,7	2,3,4,6	-1-1-1-1-		
	2,5,6	2,3,4,7							2,5,6	2,3,4,7	-						2,5,6	2,3,4,7			
	2,5,7	2,3,5,6							2,5,7	2,3,5,6							2,5,7	2,3,5,6			
	2,6,7	2,3,5,7							2,6,7	2,3,5,7							2,6,7	2,3,5,7			
	3,4,5	2,3,6,7							3,4,5	2,3,6,7							3,4,5	2,3,6,7			
step1 B_0	3,4,6	2,4,5,6					step2 B_1		3,4,6	2,4,5,6					step3 B_2		3,4,6	2,4,5,6			
	3,4,7	2,4,5,7							3,4,7	2,4,5,7	_						3,4,7	2,4,5,7			
	3,5,6	2,4,6,7							3,5,6	2,4,6,7							3,5,6	2,4,6,7	_		
	3,5,7	2,5,6,7							3,5,7	2,5,6,7							3,5,7	2,5,6,7	_		
	3,6,7	3,4,5,6							3,6,7	3,4,5,6	_						3,6,7	3,4,5,6	_		
	4,5,6	3,4,5,7							4,5,6	3,4,5,7	_						4,5,6	3,4,5,7	_		
	4,5,7	3,4,6,7							4,5,7 4,6,7	3,4,6,7	_						4,5,7	3,4,6,7 3,5,6,7	_		
	4,0,7	4,5,6,7							4,6,7	3,5,6,7	_						4,6,7	3,5,6,7	_		
	5,0,7	ייטיניי							5,0,7	4,3,0,7							5,0,7	4,0,0,7			

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 Metaanalysis. EClinicalMedicine. 2020;29-30:100630.