Creating spouse pairs in UK Biobank data: a methodological note

David Hugh-Jones, Oana Borcan & Abdel Abdellaoui

2020-12-21

There are many reasons to be interested in human spouse pairs. Social scientists research family formation. Geneticists want to learn about assortative mating. The UK Biobank, which collects genetic and other data on about 500,000 respondents, is a promising resource. But UK Biobank does not record which respondents are spouses. To get round this problem, researchers typically match pairs of respondents by geography and other characteristics (Tenesa et al. 2015; Rawlik, Canela-Xandri, and Tenesa 2019; Xia et al., n.d.; Howe et al. 2019).

Howe et al. (2019) used individuals of European descent who "(a) report living with their spouse (6141-0.0), (b) report the same length of time living in the house (699-0.0), (c) report the same number of occupants in the household (709-0.0), (d) report the same number of vehicles (728-0.0), (e) report the same accommodation type and rental status (670-0.0, 680-0.0), (f) have identical home coordinates (rounded to the nearest km) (20074-0.0, 20075-0.0), (g) are registered with the same UK Biobank recruitment centre (54-0.0) and (h) both have available genotype data. If more than two individuals shared identical information across all variables, these individuals were excluded from analysis." They also excluded same-sex pairs, pairs whose parents both died at the same age, and couples with IBD > 0.1 This left 47,549 candidate pairs. We replicate this analysis and arrive with 40956 pairs.

These processes may include "false positives" – pairs who are not actually spouses, but who just happen to live in the same area, have the same accommodation, etc. These false positive pairs can cause trouble for empirical analysis. For example, we may wish to check for assortative mating, by testing whether people have similar scores to their spouses. If we estimate the correlation of pairs' scores in the dataset, and if some the spouses are false positives, then our estimated correlation $\hat{\rho}$ will be between the true spousal correlation, and the correlation for the false positives who just live in the same area. This will bias our results. Since people may have similar polygenic scores to others living in the same area (Abdellaoui et al. 2019), results could even be biased away from zero, leading to a mistaken finding of assortative mating.

UK Biobank fields 20033 and 20034 record the 100m northing and easting of respondents' home address, as inferred from their full UK postcode. (A typical UK postcode contains about 15 residents.) These are typically not provided to

researchers, for privacy reasons. Howe et al. match respondents on fields 20074 and 20075, which record the northing and easting truncated to 1 kilometer. One way to check for false positives is to use the untruncated home location data. Pairs who are in the same square kilometer, but not in the same untruncated location, are unlikely to be spouses. In our replication of the Howe et al. data, 3800 out of 40956 pairs never share an untruncated home location.

Another way to check is to use genetic information. Some UK Biobank respondents have a genetic child who is also in the UK Biobank sample. For any set of candidate pairs, we can examine the subsample of pairs where at least one person has a genetic child in UK Biobank. We can then compute the proportion of this subsample for which the child is the genetic child of both members of the pair. These pairs are "validated" in the sense that they had a child together – a fact which is of especial interest to geneticists. In the Howe et al. data, 650 pairs had one member with a genetic child in UK Biobank. For 366 of these, the child belonged to both pair membners. Thus 56.31% of these pairs were valid. This figure is a lower bound: having a child of just one partner does not exclude having another child who is shared.

Can we improve on the quality of these datasets? We try to do so by using untruncated location data. We match people who:

- had the same untruncated location;
- both reported the same homeownership/renting status, length of time at the address, and number of children;
- attended the same UK Biobank assessment centre, on the same day;
- both reported living with their spouse ("husband, wife or partner");
- consisted of one male and one female.

We then eliminate all pairs where either spouse appeared more than once in the data. This leaves a total of 35682 pairs. We again validate them using genetic relationships. 511 pairs had a genetic child among the UK Biobank respondents. Of these 441 (86.3%) were children of both parents. As a comparison, 11% of families with dependent children included a stepchild in England and Wales in 2011 (National Statistics 2014).

The subsample of pairs with at least one genetic child is not randomly drawn. In particular, since UK Biobank respondents are mostly over 40, the subsample who have a child as a respondent tends to be older than average, with a mean age 66.281). Thus, our figure of "validated pairs" could be a biased estimate for the whole dataset, for example if younger pairs are less likely to have a shared child. Within the subsample of the Howe dataset, this does not seem to be true: neither male nor female parent's age predicts whether a child is shared (logistic regression; father, $p = 0.455$; mother, $p = 0.594$). The same is true within our own dataset (father, $p = 0.617$; mother, $p = 0.73$). Nevertheless, the proportion of validated pairs using this method should only be treated as a rough estimate.

Signing the bias due to fake pairs

Even with stringent criteria for spouse pairs, the presence of fake pairs in the data could still bias results. In many cases, the expected value of estimates will be between the statistic for true spouse pairs, and the statistic for fake pairs. So, if the value for fake pairs is closer to (farther from) zero than the value for real pairs, results will be biased towards (away from) zero. Of course, researchers do not know which are the fake pairs remaining in their data. But they can create an artificial dataset of "known fakes" and estimate statistics within that.

Using the original Howe dataset, we test for assortative mating on a polygenic score for drinks per week [XXX cite]. The rows of Table 1 show the correlation between spouses' scores. The first row is Howe et al's original dataset. The second row uses only the Howe et al. "fake pairs" which came from different postcodes. The estimate is higher, suggesting that fake pairs biased the original result upwards. The third row shows the data with the fake pairs removed. The result is still lower than for the fake pairs, so any remaining "unknown fakes" may still bias the result upward. The last row uses our own data. Indeed, here estimated correlations are smaller again, and are not significantly different from zero at p = 0.05.

Table 1: spousal correlations for polygenic score of drinks per week, different datasets

Dataset	Estimate	C.I.
Howe et al. original	0.012	(0.003, 0.022)
Howe et al., different postcode	0.035	(0.003, 0.067)
Howe et al., same postcode	0.01	(-0.000, 0.020)
Ours	0.008	(-0.003, 0.018)

Notes on other researchers

Howe again: We excluded 4866 potential couples who were the same sex (9.3% of the sample), as unconfirmed same sex pairs may be more likely to be false positives. Although sexual orientation data were collected in UK Biobank, access is restricted for privacy/ethical reasons. To reduce the possibility that identified spouse-pairs are in fact related or non-related familial, non-spouse pairs; we removed three pairs reporting the same age of death for both parents (1807-0.0, 3526-0.0). Then we constructed a genetic relationship matrix (GRM) amongst derived pairs and removed 53 pairs with estimated relatedness (IBD□ >□ 0.1). To construct the GRM; we used a pool of 78,341 markers, which were derived by LD pruning (50KB, steps of 5□ KB, r2□ <□ 0.1) 1,440,616 SNPs from the HapMap3 reference panel56

using the 1000 Genomes CEU genotype data57 as a reference panel. The final sample included 47,549 spouse-pairs"

- Most articles include Albert Tenesa.

Tenesa et al. (2015): "Using household sharing information we identified a set of 105,381 households with exactly two members in the cohort that we considered to be couples. For 94,651 out of those 105,381 households, both residents report the same household size and relationship to other household members to be 'Husband, wife or partner' or both 'Husband, wife or partner' and 'Son and/or daughter (include step-children)'. Hence, for ~90 % of the pairs we have additional confirmatory information that these were couples. Our univariate and bivariate analyses included only those couples whose coefficient of relatedness (r) was less than 0.0625, of which only seven pairs had $r > 0.025$. Of those 105,381 identified couples, we used 13,068 White-British couples and 3,726 mixed-race couples (where one member of the couple was classified as White-British and the other as non White-British) that had been genotyped in phase 1."

Where only one couple members had been genotyped, they "used only pairs where both individuals were self-reported White-British; the genotyped person was classified as White-British based on genotype; individuals reported dif- ferent ages for one or both parents; and individuals had an age difference of less than 10 years, were of opposite gen- der, and reported to live with their partner or partner and children.

Rawlik, Canela-Xandri, and Tenesa (2019): like Tenesa et al. 2015. "Specifically, using household sharing information we identified a set of 105,380 households with exactly two members in the cohort. Of these, 90,297 satisfied all of the following criteria: (a) individuals reported different ages for one or both parents; (b) individuals had an age difference of < 10 years; (c) individuals were of opposite gender; (d) both individuals reported to live only with their partner or partner and children. We restricted our analysis to a subset of 79,094 couples for which both partners self-reported to be of White-British ethnicity."

- What is this "household sharing information?"

Xia et al. (n.d.): used the "methods of Tenesa et al." but now more people have been genotyped. They analyse 80,889 couples. (Self-reported Europeans and of European ancestry). They eliminate 138 couples with relatedness above 0.025. "For couple-shared phenotypes, we removed from 1,365 to 16,060 couples that had different phenotypic values and assessment centres." Hmm, but they didn't remove these from the database of couples more generally?

Abdellaoui, Abdel, David Hugh-Jones, Loïc Yengo, Kathryn E Kemper, Michel G Nivard, Laura Veul, Yan Holtz, et al. 2019. "Genetic Correlates of Social Stratification in Great Britain." *Nature Human Behaviour* 3 (12): 1332–42.

Howe, Laurence J, Daniel J Lawson, Neil M Davies, Beate St Pourcain, Sarah J Lewis, George Davey Smith, and Gibran Hemani. 2019. "Genetic Evidence for Assortative Mating on Alcohol Consumption in the Uk Biobank." *Nature Communications* 10 (1): 1–10.

National Statistics, Office for. 2014. "Stepfamilies in 2011." 2014. https://webarchive.nationalarchives.gov.uk/20160105222243/http://www.ons.gov.uk/ons/rel/family-demography/stepfamilies/2011/stepfamilies-rpt.html.

Rawlik, Konrad, Oriol Canela-Xandri, and Albert Tenesa. 2019. "Indirect Assortative Mating for Human Disease and Longevity." *Heredity* 123 (2): 106–16.

Tenesa, Albert, Konrad Rawlik, Pau Navarro, and Oriol Canela-Xandri. 2015. "Genetic Determination of Height-Mediated Mate Choice." *Genome Biology* 16 (1): 1–8.

Xia, Charley, Oriol Canela-Xandri, Konrad Rawlik, and Albert Tenesa. n.d. "Evidence of Horizontal Indirect Genetic Effects in Humans." *Nature Human Behaviour*.