International Journal of Population Data Science

Journal Website: www.ijpds.org

The Mysterious Case of the Delayed Twin: using research data to resolve linkage questions.

Avery, D[1]

[1]University of Oxford

Introduction

In a large biobank of over half a million people, we have several pairs of participants who appear to share their genome. As more individuals are sequenced, more pairs are likely to be found. If these are twins then this is great news, but it isn't quite that simple.

Objectives and Approach

Where 2 people share a genome we need to be able to confirm that these pairs are twins. However, there are a number of issues which could cause 2 people to appear to share a genome; for example being recruited twice, donating blood on another's behalf, etc. We already identify and exclude participant data based on these conditions. We developed our methodology by looking at the first identified pair in great detail, looking for evidence which specifically ruled out possible alternate explanations, and then applying and refining the method on later pairs.

Results

We were able to demonstrate the pair were almost certainly twins using their biochemistry and family questionnaire data as principal sources. We also identified a number of variables which were useful in indicating the likelihood of a twin, and now form part of a methodology which we are still developing. Even more usefully, we identified a number of variables that seemed like useful measures but proved extremely misleading. To date we have 26 pairs of possible twins, with 9 confirmed as twins and the remainder looking likely to be twins but falling short of a threshold for confidence. We also have 75 pairs which confirm duplicate participants we have already excluded.

Conclusion/Implications

We formed two lessons: even very simply linkages come with pitfalls, and you should gather more administrative data than you think. We're proposing the collection of additional familial relationship data in our third resurvey. We are also looking into machine learning and statistical techniques to better identify twins and duplicates.

http://dx.doi.org/10.23889/ijpds.v3i4.643