

Harmonised Health Outcomes Across Administrative Data: Lessons, Opportunities, and Challenges from UK Biobank

Authors names: Eleni Domzaridou, Ben Lacey, Megan Conroy, Naomi Allen, Yangmei Li

Objective

To develop a resource that maps health outcomes across coding schemas in linked administrative data in UK Biobank, addressing the challenge of identifying equivalent outcomes from multiple sources.

Methods

UK Biobank is a prospective cohort study of ~500,000 adults, recruited between 2006-2010, with follow up for health outcomes through linkage with administrative data. Clinical codes include Read Version 2 (Read2) and Clinical Terms Version 3 (CTV3) from primary care, and International Classification of Diseases (ICD) 9th and 10th editions (ICD-9 and ICD-10) from hospitals, cancer registries, and death records; self-reported conditions were also reported at recruitment. We reviewed existing mapping resources and mapped clinical codes in different schemas to 4-digit ICD-10 codes.

Results

We successfully mapped 81% of Read2 (N = 12,448), 93% of CTV3 (24,188), 92% of ICD-9 (3,060), and 100% of self-reported codes (509) to ICD-10 codes. Although existing resources frequently allowed one-to-one mapping of ICD-10 codes (94% of the mapped codes for Read2, 58% of CTV3, and 79% of ICD-9), the remaining codes required extensive clinical review, which is ongoing. The conversion increased the granularity of health outcomes by 3.8 times from 2,000 3-digit to 7,500 4-digit ICD-10 codes. The mapping quality will be evaluated using phocodes, and by assessing consistency across data sources.

Conclusion

Our approach preserves clinical detail, increases coding granularity, uncovers nuanced outcomes, and enables precise, internationally comparable research using enriched UK Biobank data. Although harmonisation supports cross-cohort research, validation of outcomes across data sources remains essential to avoid misclassification and minimise bias.