

International Journal of Population Data Science

Journal Website: www.ijpds.org



Enhancing Health Outcomes in Linked Administrative Data: Development and Validation of an Open-Access Mapping Resource using UK Biobank

Eleni Domzaridou¹, Ben Lacey¹, Naomi Allen¹, and Yangmei Li¹

¹ UK Biobank, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom

Objectives

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. Our approach minimised the loss of clinical detail, a common limitation in such efforts, to enhance its utility for health research.

Methods

UK Biobank is a prospective cohort study of ~500,000 adults, recruited between 2006-10, with follow up for health outcomes through linkage with administrative health data. Clinical coding schemas 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 secondary care, cancer registries and death records; self-reported conditions were also reported at recruitment. We reviewed existing mapping resources and, with clinical support, mapped clinical codes in different schemas to 4-digit ICD-10 to provide detailed clinical information using a single internationally-recognised schema.

Results

We processed data from 230,096 participants with primary care records, 442,267 with secondary care records, 40,447 with death records, and 397,063 with self-reported data. We successfully mapped to 81% of Read2 codes ($N = 12,448$), 93% of CTV3 (24,188), 92% of ICD-9 (3,060), and 100% of self-reported (509) to ICD-10 codes. Although existing resources frequently allowed a single code to be mapped to a single ICD-10 code (94% of the mapped codes for Read2, 58% of CTV3, and 79% of ICD-9), the remaining codes require extensive clinical review, which is ongoing. The conversion increased the granularity of health outcomes by 5.8 times from 2,006 3-digit ICD-10 codes to 11,625 4-digit ICD-10 codes. The most common ICD-10 codes included those related to musculoskeletal diseases (24%).

Conclusion

The increased granularity of ICD coding enhances the research potential of UK Biobank data, enabling precise outcome definitions and detailed comparisons with other healthcare datasets. The enhanced mappings revealed underrepresented and nuanced outcomes, improving subtyping of conditions, and supporting robust comparisons with external datasets using internationally recognised coding standards.

