Deriving research-quality phenotypes from national electronic health records to advance precision medicine: a UK Biobank case-study

Spiros C. Denaxas*, Ghazaleh Fatemifar*, Riyaz S. Patel, Harry Hemingway

***Abstract*— High-throughput genotyping and increased availability of electronic health records (EHR) are giving scientists the unprecedented opportunity to exploit routinely generated clinical data to advance precision medicine. The extent to which national structured EHR in the United Kingdom can be utilized in genome-wide association studies (GWAS) has not been systematically examined. In this study, we evaluate the performance of an EHR-derived acute myocardial infarction phenotype (AMI) for performing GWAS in the UK Biobank.**

I. Introduction

GWAS traditionally leverage longitudinal studies to manually derive disease cases/controls. Consequently, they often examine a small set of broad phenotypes. With decreasing genotyping costs and increased availability of computational resources, there is an unmet need for larger, phenotypically-rich datasets to drive precision medicine at scale [1]. EHR contain a wealth of information (e.g. diagnoses, laboratory measurements, prescriptions) that can potentially compose high-resolution phenotypes. While significant progress has been made by eMERGE [2] in the US, the extent to which national, structured UK EHR can be utlized has not been systematically investigated. In this study, we perform a GWAS in the UK Biobank using an EHR-derived AMI phenotype and evaluate our findings.

II. Methods

The UK Biobank [3] is an extensively-phenotyped and genotyped (Affymetrix, 820,967 SNPs) cohort of 502,640 UK participants. We performed sample and SNP quality control procedures in participants with genetic data (n=120,286) and derived the final cohort (n=112,142). We applied a previously-validated AMI EHR phenotype from CALIBER [4] which links structured national UK EHR sources from primary/secondary care and mortality in ~15 million patients. We utilized diagnostic (ICD-10 I21-I23, I24.1, I25.2, ICD-9 410-414, 4297) and procedure codes (OPCS-4: K50.2-3) to define cases - non-cases were coded as "unaffected". We used a logistic regression to test the association between 10 million expected allelic dosages and AMI controlling for sex, batch, genotyping chip, recruitment centre and PC 1-15. Evaluation was through comparison with CARDIoGRAMplusC4D [5].

S. C. Denaxas, Institute of Health Informatics, University College London, UK (phone: +4420354955324; e-mail: s.denaxas@ucl.ac.uk)

G. Fatemifar, Institute of Health Informatics, University College London, UK (e-mail: g.fatemifar@ucl.ac.uk).

R. Patel , Institute of Health Informatics, University College London, UK (e-mail: riyaz.patel@ucl.ac.uk).

H. Hemingway, Institute of Health Informatics, University College London, UK (e-mail: h.hemingway@ucl.ac.uk).

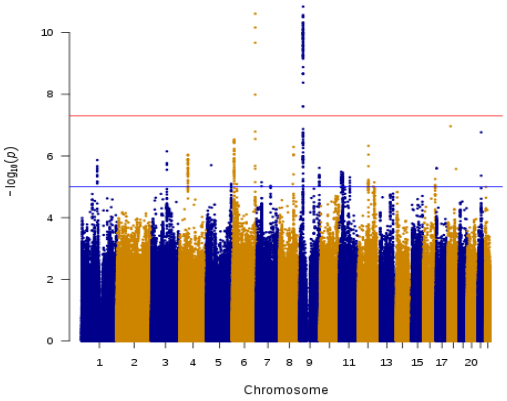

Figure 1. Genome-wide Manhattan plot for AMI

III. Results

Using an EHR AMI phenotype, we identified 3,408 affected and 108,734 unaffected participants. We discovered 69 non-independent genetic variants spanning (Fig 1.) regions on chromosomes 6 and 9 (e.g. 9p21 loci) showing genome-wide significance ($p<5x10^{-8}$, $\lambda=1.02$). Consistent direction and magnitude of associations were replicated in 67 (97%) of previously reported genetic variants [5].

IV. Discussion

Using an EHR-derived AMI phenotype, we identified positively-associated variants across previously reported loci from conventional GWAS. Transforming raw EHR to research-ready phenotypes however is challenging since data are collected for various purposes and robust high-throughput phenotyping methods are required. EHR phenotypes can potentially be used to increase sample and phenotypic resolution and drive the discovery of clinically-actionable associations.

References

[1] J. Denny, *et* al., "Phenome-wide Association Studies as a Tool to Advance Precision Medicine", *Annu Rev Genomics Hum Genet*, vol. 17, pp. 353-73, 2016.

[2] O. Gottesman, *et* al., "The eMERGE Network", *Genetics in Medicine*, vol. 15, pp. 761-771, 2013.

[3] Sudlow C., *et* al., "UK Biobank: An Open Access Resource for identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age", *PLOS Med"*, vol. 12. e1001779, 2015.

[4] Morley K., *et* al., "Defining Disease Phenotypes Using National Linked Electronic Health Records: A Case Study of Atrial Fibrillation", *PLOS ONE*, vol. 9, e110900, 2014.

[5] CARDIoGRAMplusC4D, *et* al., "A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease", *Nat Genet*, vol. 47, pp. 1121-1130, 2015.