Mapping UK Biobank to the Experimental Factor Ontology.

Zoë May Pendlington[1,2], Paola Roncaglia[1], Edward Mountjoy[2,3], Gautier Koscielny[2,4], Helen Parkinson[1] and Simon Jupp[1,2].*

[1] European Molecular Biological Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [2] Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [3] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [4] GSK, Medicines Research Center, Gunnels Wood Road, Stevenage, SG1 2NY, UK.

ABSTRACT

UK Biobank is a massive datasource for population health data used extensively in research. Some of the clinical data is mapped to ICD-10, but coverage is incomplete, as are mappings from existing ICD-10 codes to public ontologies. Here, we describe a pipeline to map 1,552 UK Biobank traits to public ontology terms via the Experimental Factor Ontology. Our approach uses ontology mapping services and manual curation to provide almost complete coverage (97%), thus increasing interoperability of UK Biobank with public datasets. Both mappings and services described are freely available for use and the mapping pipeline represents a typical curation workflow that can be adopted for other domains.

1 INTRODUCTION

UK Biobank (UKB; http://www.ukbiobank.ac.uk/) is a large-scale prospective population study, providing data for general health, disease and associated phenotype research. Its aim is to understand life-threatening and disabling conditions brought on by environmental and genomic factors (Ollier, Sprosen & Peakman, 2005). Between 2006 and 2010, 500,000 volunteers aged 40-69 were enrolled in the UKB study (Collins, 2012). Questionnaire data were collected to track lifestyle, socioeconomic and environmental factors alongside cognitive ability, medical and physical examination and genotyping (Sudlow et al., 2015). Data for activity, heart rate and sleep was obtained by 24-hour activity monitoring 20% of participants.

UKB data are used in health applications and link electronic health records containing diseases, death circumstances, hospital admissions and general health. There is growing interest in integrating UKB with other public resources; e.g. Open Targets (OTAR) projects integrates UKB data into a Genetics portal[1] (Carvalho-Silva et al., 2018) which uses variant-based statistical evidence to identify drug targets and trait-associated loci. Genome-wide association studies have been using UKB data, leading the GWAS Catalog to include UKB data (Buniello et al., 2019). As UKB data are integrated with existing public resources, there is an increasing need to harmonise traits, phenotypes and diseases, to support cross-querying and visualisation.

Health-related outcome variables in UKB are mapped to ICD-10 codes (10th revision of the International Statistical Classification of Diseases and Related Health Problems; World Health Organization., 2004). However, ICD-10 is not fully interoperable with other open biomedical ontologies (OBO) commonly used to annotate public data (Sollie et al., 2013). To facilitate UKB data interoperability with other resources, like the GWAS Catalog and OTAR, we have mapped existing ICD-10 codes in UKB to public ontologies, and provided ontology mappings also for other traits in UKB, thereby covering the entire dataset. We used ontology services from EMBL-EBI to provide these mappings, including to OBO Foundry Ontologies (Smith et al., 2007).

UKB includes information on diseases, phenotypes, measurements and biomarkers. As no single ontology exists to cover this diversity, we used the Experimental Factor Ontology (EFO), an application ontology which provides an interoperable and orthogonal set of terms and acts as a bridge to several other public ontologies. EFO pulls together terms from various OBO Foundry ontologies and provides many cross-references (x-refs) to other medical coding systems such as ICD and SNOMED. EFO further enriches the OBO ontologies by providing cross-bridging relations (axioms) to create logical definitions for terms, e.g. linking phenotypes to diseases (Sarntivijai et al., 2016) and diseases to anatomy. It is the primary ontology for annotating public GWAS experiments in the GWAS Catalog (Buniello et al., 2019; Carvalho-Silva et al., 2018), and contains extensive clinical measurement and biomarker representation; therefore, it is a good fit for UKB data and is well integrated with a many public data sources (Malone et al., 2010).

2 METHODS

UKB data comprises concepts mapped to ICD-10 traits (64%) obtained from clinical records, and participant self-reported traits obtained by questionnaire (36%), these had no ontology mapping. Traits were mapped to ontologies

* To whom correspondence should be addressed.

[1] https://genetics.opentargets.org/

using ontology services from EMBL-EBI[2]. 70% (389) of the self reported trait questions were rewritten to assist automated prediction of ontology mappings, e.g. in regards to body size, "when you were 10 years old, compared to average would you describe yourself as:" was given the query label "comparative body size at age 10".

The ontology services aid in data mapping to ontologies (Zooma[3]), translation of vocabularies to one another (Ontology X-ref Service; OxO[4]) and a search engine to over 200 biomedical ontologies (Ontology Lookup Service; OLS[5]).

Zooma predicts ontology mappings for free-text queries, using background knowledge collected from previously mapped data, manually curated by other databases at EMBL-EBI. Zooma awards each mapping with a confidence score: high (automatic mapping to previously curated Zooma datasource), good, medium and none, the scores are used by curators to prioritise curation efforts.

OxO provides access to ontology x-refs extracted from public ontologies and the Unified Medical Language System (UMLS). It can be used to find mappings between two vocabularies, e.g. IDC10 codes to other public ontologies[6].

Mapping types were assigned after the automated process during manual curation (Table 1). After the first phase, 974 (63%) of inadequate mappings (Table 1) and 521 (34%) of traits remaining with no mapping were searched for using OLS in a three-stage process: searching EFO only to determine if existing term mappings were already available, searching ontologies currently imported by EFO - including the Human Phenotype Ontology (HPO), and, finally, searching all ontologies in OLS which could be used as a source of new terms representing unmapped traits.

Traits with no adequate match after manual curation were reviewed, and split into terms that needed new ontology classes to be created and terms that were considered "currently unmappable". A master file of all traits was compiled and published online[7], containing trait labels extracted from UKB data, mapped-to term labels, mapped term URIs, mapping scores (Table 1) and ICD-10 or self-reported trait field codes.

3 RESULTS

UKB traits consisted of 1,552 traits: 999 ICD-10 and 553 self-reported traits. ICD-10 traits were curated by two groups using their preferred methodology: 395 traits were mapped using Zooma only and 604 using OxO only (Table 2). All 553 self-reported traits were mapped using Zooma only (Table 2). After the automated mapping processes, all traits were manually verified.

Table 1. Mapping score criteria

Mapping	Criteria definition
Exact	Trait and mapped-to term are deemed fully equivalent to one another in definition
Broad	Mapped-to term is a larger concept than the trait itself. The mapped-to trait may be an umbrella concept covering the query trait.
Narrow	Mappted-to term is a smaller concept than the trait itself. The query trait may be an umbrella concept covering the mapped-to term.
Inadequate	Suggested mapping is either considered too broad, too narrow or incorrect.
None	Trait with no mapping.

Mapping criteria considered for results of mapping pipeline.

All 948 traits mapped by Zooma were awarded a confidence score (see Methods). Before manual curation, 826 mappings (87%) had good or medium score. After curation, 291 (35%) mappings were approved, while 535 (65%) were replaced by manually curated mappings.

The 604 ICD-10 traits mapped using OxO were dictated by the shortest paths to terms from EFO or from external ontologies imported into EFO[8]. These mappings were then manually curated; 213 (35%) were approved, 74 (12%) were replaced with manually mappings, and 317 (53%) were not mapped to any term.

As of EFO 2.106[9], we successfully mapped 97% (1,508) of UKB traits to EFO. All mappings were assigned a mapping score based on criteria (Table 1). 93% of all mappings are either exact - for example, 'typhus fever' was mapped to EFO:0009117 'typhus' - or mapped to a broader category - for example, "comparative height size at age 10" was mapped to EFO:0004339 'body height' (Table 3).

The trait "secondary malignant neoplasm of respiratory and digestive organs" was considered a dual concept and mapped to EFO:0003853 "respiratory system neoplasm" and EFO:0008549 "digestive system neoplasm". Similarly, 48 traits were broken into multiple concepts and mapped to two or three EFO terms. In some cases, multiple related UKB traits were mapped to the same EFO term; e.g., ICD-10:S65 'injury of blood vessels at wrist and hand level' and ICD-10:S20 'superficial injury of thorax' were broadly mapped to EFO:0000546 'injury', and would therefore resolve to the same entity in applications such as the OTAR Genetics portal. Going forward, higher resolution could be obtained by creating new subclasses of the broader term.

[2] https://www.ebi.ac.uk/spot/ontology

[3] https://www.ebi.ac.uk/spot/zooma/

[4] https://www.ebi.ac.uk/spot/oxo/

[5] https://www.ebi.ac.uk/ols/

[6] https://www.ebi.ac.uk/spot/oxo/datasources/ICD-10CM

[7] https://raw.githubusercontent.com/EBISPOT/EFO-UKB-mappings/master/UK_Biobank_master_file.tsv

[8] https://github.com/opentargets/ontology-utils

[9] https://github.com/EBISPOT/efo/releases/tag/v2019-03-18

Table 2. Examples of mapping results

Tool	UKB trait	Tool mapping result	Manual curation	Action	Mapping type
Zooma	"Do you often feel lonely?"	EFO:0007865 'Loneliness measurement'	EFO:0007865 'Loneliness measurement'	Add to mapping file.	Exact
OxO	ICD10:H58 'Other disorders of eye and adnexa diseases classified elsewhere'	ICD10:H58†	MONDO:0000462 'Eye adnexa disease'	New EFO term'eye adnexa disease', x-reffing ICD10:H58 and MONDO:0000462. Add to mapping file.	Broad
Zooma	"Has a doctor ever told you that you have had any of the conditions below?" "lung cancer (not mesothelioma)" was an option.	EFO:0002934 'Lung cancer cell line'†	EFO:0001071 'Lung carcinoma'	Add to mapping file.	Narrow

The mapping pipeline for both ICD-10 and self-reported traits is shown via examples of mappings using Zooma and/or OxO tools, with mapping outcomes (Table 1) and any action we took in making changes to EFO, including the approved mapping in the master mapping file. († meaning mapping rejected)

Table 3. Break-down of UK Biobank mappings

Mappings	ICD10-mapped traits (%)	Self-reported traits (%)
Total	999 (65% of all traits)	553 (35% of all traits)
Exact	494 (49%)	404 (73%)
Narrow	27 (3%)	25 (5%)
Broad	447 (45%)	111 (20%)
None	31 (3%)	13 (2%)

All UKB mappings were assigned a mapping type based on criteria (Table 1).

227 terms were added to EFO (Figure 1); 176 were new EFO terms and 73 were terms imported from HPO, e.g. HP:0000952 'jaundice' and HP:0031703 'hearing loss'.

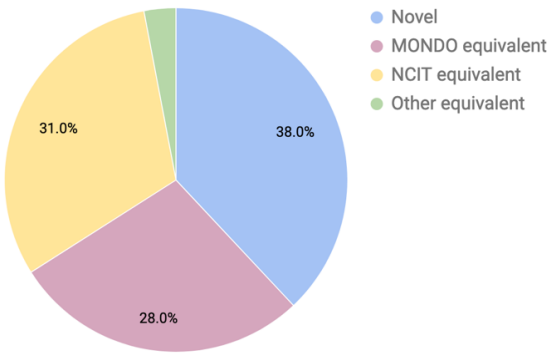

Figure 1. Representation of new terms added for UKB traits in EFO.

Of the new terms in EFO, 28% were equivalent to MONDO terms, 31% were equivalent to NCI Thesaurus terms and 3% referenced other ontologies including the Ontology for miRNA Target (Huang et al., 2011), the Symptom Ontology[10] and The Mammalian Phenotype Ontology (Smith and Eppig, 2012). X-refs were made to these ontologies in the new EFO terms. 38% were novel EFO terms and did not exist in other ontologies.

From the 1,011 ICD-10 traits, 95% (964) were added to the corresponding EFO mapped terms as a cross reference. The remaining 5% are currently unmapped and have not been added to EFO.

4 DISCUSSION

Our approach to mapping UKB to public resources shows how EMBL-EBI ontology services are used to map a large dataset, improving interoperability for several high accesses data resources. UKB is ongoing, with data collection from aging participants (Sudlow et al., 2015). This will provide data on disease progression and aging, alongside imaging (of brain, heart, abdomen, bones and carotid arteries) and exome sequencing (Allen et al., 2012). Thus, we expect more data to be generated requiring reapplication of our mapping pipeline.

Currently, the pipeline is semi automated as curation enables maximum coverage of the dataset. Use of Zooma and OxO reduced manual curation time, with 32% of all automated mappings retained in the final outcome. Most curator time was spent validating automated mappings in order to increase their confidence. Manual validation results will be fed back into the Zooma knowledgebase, to further enhance Zooma's ability to predict mappings for similar datasets in the future. Our work has resulted in ICD-10 mappings being added to EFO and also made available via the OxO service

[10] http://symptomontologywiki.igs.umaryland.edu/wiki/index.php

with appropriate provenance. This reduces the need for manual curation of these mappings in the future. All future mappings will also be fed back to Zooma and OxO, continuing a loop that will make our pipeline less reliant on manual curation and validation. Present and future results of our pipeline will be available to others. For example, the Expression Atlas use EFO to provide efficient search and visualisation of gene and protein expression in many areas including diseases and phenotypes (Petryszak et al., 2015). And the OTAR Genetics portal will now be able to produce efficient search and visualisation of newly-derived genetic associations derived from UKB data. Links between these associations made possible by ontology inferences will further aid in discovery of novel drug targets and insight into causal genes. The unmapped entries (3% of UKB traits) have no obvious ontology mapping. Such problematic traits include the self-reported "Seen doctor (GP) for nerves, anxiety, tension or depression", where our manual mapping of EFO:0000677 'mental or behavioural disorder' is too broad. We are currently addressing these unmapped entries, and considering addition of novel categories in EFO, such as 'self-reported trait' and appropriate children, cross-referencing ICD-10 codes where appropriate.

ACKNOWLEDGEMENTS

Thanks to Danielle Welter, Patricia L. Whetzel, Christopher J. Mungall and Nicole Vasilevsky for their helpful advice on some mappings.

FUNDING

This project has received funding from Open Targets (OTAR005), the European Union's Horizon 2020 research and innovation programme under grant agreement No 654248 (CORBEL) and 676559 (ELIXIR Excelerate) and EMBL-EBI institutional funding.

REFERENCES

Allen, N., Sudlow, C., Downey, P., Peakman, T., Danesh, J., Elliott, P., Gallacher, J., Green, J., Matthews, P., Pell, J., Sprosen, T. and Collins, R. (2012). UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, **1**(3), 123-126.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P.L., Amode, R., Guillen, J.A., Riat, H.S., Trevanion, S.J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L.A., Cunningham, F. and Parkinson, H. (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research.*, **47**, D1005-D1012.

Carvalho-Silva, D., Pierleoni, A., Pignatelli, M., Ong, C., Fumis, L., Karamanis, N., Carmona, M., Faulconbridge, A., Hercules, A., McAuley, E., Miranda, A., Peat, G., Spitzer, M., Barrett, J., Hulcoop, D., Papa, E., Koscielny, G. and Dunham, I. (2018). Open Targets Platform: new developments and updates two years on. *Nucleic Acids Research.*, **47**(D1), D1056-D1065.

Collins, R. (2012). What makes UK Biobank special? *The Lancet.*, **379**(9822),1173-1174.

Huang, J., Townsend, C., Dou, D., Liu, H. and Tan, M. (2011). OMIT: A Domain-Specific Knowledge Base for MicroRNA Target Prediction. *Pharmaceutical Research*, **28**(12), 3101-3104.

Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A. and Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**(8), 1112-1118.

Ollier, W., Sprosen, T. and Peakman, T. (2005). UK Biobank: from concept to reality. *Pharmacogenomics.*, **6**(6), 639-646.

Petryszak, R., Keays, M., Tang, Y., Fonseca, N., Barrera, E., Burdett, T., Füllgrabe, A., Fuentes, A., Jupp, S., Koskinen, S., Mannion, O., Huerta, L., Megy, K., Snow, C., Williams, E., Barzine, M., Hastings, E., Weisser, H., Wright, J., Jaiswal, P., Huber, W., Choudhary, J., Parkinson, H. and Brazma, A. (2015). Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Research, **44**(D1), D746-D752.

Sarntivijai, S., Vasant, D., Jupp, S., Saunders, G., Bento, A., Gonzalez, D., Betts, J., Hasan, S., Koscielny, G., Dunham, I., Parkinson, H. and Malone, J. (2016). Linking rare and common disease: mapping clinical disease-phenotypes to ontologies in therapeutic target validation. *Journal of Biomedical Semantics*, **7**(1).

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L., Eilbeck, K., Ireland, A., Mungall, C., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S., Scheuermann, R., Shah, N., Whetzel, P. and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology.*, **25**(11), 1251-1255.

Smith, C. and Eppig, J. (2012). The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. *Mammalian Genome*, **23**(9-10), 653-668.

Sollie, A., Sijmons, R., Lindhout, D., van der Ploeg, A., Rubio Gozalbo, M., Smit, G., Verheijen, F., Waterham, H., van Weely, S., Wijburg, F., Wijburg, R. and Visser, G. (2013). A New Coding System for Metabolic Disorders Demonstrates Gaps in the International Disease Classifications ICD-10 and SNOMED-CT, Which Can Be Barriers to Genotype-Phenotype Data Sharing. *Human Mutation*, **34**(7), 967-973.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T. and Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine.*, **12**(3), e1001779.

World Health Organization. (2004). *International statistical classification of diseases and related health problems*. Geneva: World Health Organization.