Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond

Vaclav Papez
University College London – Institute of Health informatics
Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond

Vaclav Papez 1,2, Maxim Moinat 3,4, Erica A. Voss 5, Sofia Bazakou 3, Anne Van Winzum 3, Alessia Peviani 6, Stefan Payralbe 3, Michael Kalilfelz 6, Folkert W. Asselbergs 1,2,7, Daniel Prieto-Alhambra 8,9, Richard J.B. Dobson 1,2,9, and Spiros Denaxas 1,2,10,11

1Institute of Health Informatics, University College London, London, UK, 2Health Data Research UK, London, UK, 3The Hyve, Utrecht, The Netherlands, 4Erasmus Medical Center Rotterdam, Rotterdam, The Netherlands, 5Department of Epidemiology, Janssen Research & Development LLC, Raritan, New Jersey, USA, 6Odysseus Data Services GmbH, Berlin, Germany, 7Amsterdam University Medical Centers, Department of Cardiology, University of Amsterdam, Amsterdam, The Netherlands, 8Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, UK, 9Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), King's College London, London, UK, 10British Heart Foundation Data Science Centre, London, UK and 11UCL Hospitals, NIHR Biomedical Research Centre (BRC), London, UK

Vaclav Papez and Maxim Moinat contributed equally to this work.

Corresponding Author: Spiros Denaxas, PhD, Institute of Health Informatics, University College London, London NW12DA, UK; s.denaxas@ucl.ac.uk

Received 13 July 2022; Revised 3 October 2022; Editorial Decision 5 October 2022; Accepted 12 October 2022
Background

• EHDEN Rapid Collaboration Call
• UK Biobank (~500K)
  • Baseline data (~8k data fields, proprietary dictionaries)
  • EHR from primary care (SNOMED CT, CTV3, EMIS and TPP proprietary codes, dm+d)
  • EHR from hospital care (ICD-10, ICD-9, OPCS4, OPCS3)
  • Mortality register (ICD-10, ICD-9)
  • Cancer register (ICD-O)
  • Covid-19 measurements (EMIS and TPP proprietary codes)
  • Genomic data
• OMOP Common Data Model (v5.3)
Methods

• ETL
  • Syntactic mapping
  • Semantic mapping
    • Athena Vocabulary repository
    • Bespoke mappings (8 in total)

• Testing and validation
  • Manually written test cases and automated tests on synthetic data
  • OHDSI Achilles, OHDSI DQD and EHDEN CDMInspection
  • Comparing a series of metrics between the raw data and the OMOP converted data
ETL Workflow

1. White Rabbit Scan
2. Delphyne
3. ETL runs on UKB, Data Validation
   - Create synthetic data
   - Develop ETL
   - Refine scripts
   - OMOPed data
Semantic mapping example
| Results                        | Source UK Biobank data | OMOP-Transformed UK Biobank data | Transformed UK Biobank COVID-19 positive sub population |
|-------------------------------|------------------------|----------------------------------|---------------------------------------------------------|
| Patients                      | 502,505                | 502,504                          | 3,086                                                   |
| % Female                      | 54.4                   | 54.4                             | 48.76                                                   |
| Median age (IQR)              | 58 (13)                | 58 (13)                          | 58 (15)                                                  |
| Median Townsend deprivation index (IQR) | -2.135 (4.18)         | -2.135 (4.18)                    | -1.111 (5.19)                                            |
| BMI median - baseline (IQR)   | 26.652 (5.72)          | 26.65 (5.70)                     | 27.7 (6.21)                                              |
| BMI median - GP EMIS (IQR)    | 27.2 (6.9)             | 27.3 (6.84)                      | 28.89 (8)                                                |
| SBP median - baseline (IQR)   | 136 (26)               | 136 (26)                         | 136 (25)                                                 |
| DBP median - Baseline (IQR)   | 81 (14)                | 81 (14)                          | 82 (14)                                                  |
### Results

|                  | Source UK Biobank data | OMOP-Transformed UK Biobank data | Transformed UK Biobank COVID-19 positive sub population |
|------------------|------------------------|---------------------------------|-------------------------------------------------------|
| **Smoking status** |                        |                                 |                                                       |
| Not answered     | 2,276                  | Not mapped                      | Not mapped                                             |
| Never            | 317,891                | 317,891                         | 1,676                                                  |
| Previous         | 197,949                | 197,949                         | 1,323                                                  |
| Current          | 55,676                 | 55,676                          | 395                                                    |
| **Comorbidities**|                        |                                 |                                                       |
| T2DM             | 40,433 (8.04%)         | 40,476 (8.05%)                  | 453 (14.67%)                                           |
| HF               | 8,068 (1.60%)          | 8,053 (1.6%)                    | 140 (4.53%)                                            |
| AMI              | 10,593 (2.10%)         | 10,749 (2.13%)                  | 110 (3.56%)                                            |
| COPD             | 22,364 (4.45%)         | 22,367 (4.45%)                  | 328 (10.62%)                                           |
| HT               | 175,449 (34.91%)       | 175,539 (34.93%)                | 1,571 (50.9%)                                          |
## Results

- 690 baseline datafields with 2898 values encoded by proprietary coding system mapped

| Source Vocab                              | Used source terms # | Mapped used terms # (%) | Events #   | Mapped event # (%) |
|-------------------------------------------|---------------------|-------------------------|------------|-------------------|
| Baseline ethnic status                    | 22                  | 10 (45.45%)             | 533,612    | 512,158 (95.97%)  |
| Self-reported non-cancer illness          | 446                 | 351 (78.69%)            | 1,127,434  | 946,053 (83.91%)  |
| Self-reported cancer                      | 82                  | 48 (58.53%)             | 53,384     | 37,802 (70.81%)   |
| Self-reported medication                  | 3,737               | 1,100 (29.43%)          | 1,381,148  | 1,218,935 (88.25%)|
| Self-reported procedures                  | 254                 | 128 (50.39%)            | 994,355    | 864,788 (86.96%)  |
| Haematology samples                       | 124                 | 93 (75%)                | 61,119,731 | 45,629,849 (74.65%)|
| Hospital EHR admission source             | 86                  | 44 (51.16%)             | 3,541,594  | 282,505 (7.97%)   |
| Hospital EHR admission method             | 63                  | 58 (92.06%)             | 3,541,610  | 3,540,046 (99.95%)|
| Hospital EHR discharge destination        | 91                  | 56 (61.53%)             | 3,484,435  | 3,189,509 (91.53%)|
Results

- A small number of patients identified in converted data only
- Successfully transformed
  - Hospital care
    - 99.9% ICD-10; 91% ICD-9
    - 89.32% OPCS4; 77% OPCS3
  - 99.95% Death events
  - Primary care
    - 97.67% SNOMED CT; 97.78% CTV3
    - 98.74% dm+d
    - 0.19% TPP and EMIS
- DQD
  - 3399 checks passed
  - 18 failed
Discussion

Phenotype Algorithms for the Identification and Characterization of Vaccine-Induced Thrombotic Thrombocytopenia in Real World Data: A Multinational Network Cohort Study

Azza Shoaiib1,2, Gowtham A Rao3,4, Erica A Voss3,4, Anna Astropolets4,5, Miguel Angel Mayer6, Juan Manuel Ramirez-Anguita6, Filip Majlakovic7, Biljana Carevic8, Scott Horban9, Daniel R Morales9, Talita Duarte-Salles10, Clement Fraboulet11, Tanguy Le Carrou12, Spiros Denaxas13, Vaclav Papez13, Luis H John14, Peter R Rijnbeek14, Evan Minty15, Thamir M Alshammarri16, Rupa Makadia3,4, Clair Blacketer17, Frank DeFalco3,4, Anthony G Sena3,4, Marc A Suchard17, Daniel Prieto-Alhambra18, Patrick B Ryan1,3.

Affiliations + expand

PMID: 35653017 PMCID: PMC9160850 DOI: 10.1007/s40264-022-01187-y Free PMC article

Contextualizing adverse events of special interest: A multinational cohort study to characterize the baseline incidence rates in 24 million COVID-19 infected subjects across 26 databases

Erica A. Voss MPH1,2,3, Azza Shoaiib PhD1,2, Anna Yin Hui Lai PhD1,4, Claire Blacketer MPH1,2,3, Thamir Alshammarri PhD1,2, Rupa Makadia PhD1,3, Kevin Haynes PharmD1, Anthony G. Sena BA1,2,3, Gowtham Rao MD1,3, Sebastiaan van Sandijk MSC1,4, Clement Fraboulet MS1, Laurent Boyer1, Tanguy Le Carrou1, Scott Horban BSc Hon1, Daniel R. Morales PhD1,3, Jordi Martínez Roldán MD1,2, Juan Manuel Ramirez-Anguita PhD1,2,4, Miguel A. Mayer MD1,2,14, Marcel de Wilde2,3,2,2, Luis H. John MS1,2, Talita Duarte-Salles PhD1,3, Elena Roel MD1,5, Andrea Pistillo MSC1, Raivo Kolde PhD1,6, Filip Majlakovic MSC1,7, Spiros Denaxas PhD1,2,18, Vaclav Papez PhD1,4,19, Michael G. Kahn MD1,2,19, Karthik Natarajan PhD20,21, Christian Reich MD21, Alex Secora PhD1, Evan P. Minto MD21, Nimag H. Shah MBBS, PhD21,22, Jose D. Posada PhD21,22, Maria Teresa Garcia Morales MSC21, Diego Bosca PhD21, Honorio Cadenas Juanino21, Antonio Diaz Holgado22, Miguel Pedrera Jimenez21, Pablo Serrano Balazote21, Noelia Garcia Barrio21, Selyuk Shen MD21, Ali Yagiz Uresin MD21, Baris Erdogan PhD21, Luc Belmans MD21, Geert Byttebier MSC21, Manu L.N.G. Malbrain MD21,22, Daniel J Dedman MPH21, Zara Cuccu21, Rohit Vashisth PhD21, Atul J. Butte MD21,3,22, Ayan Patel MS21,3, Lisa Dahm PhD21,22, Cora Han JD21,22, Fan Bu PhD21, Faizah Arshad21, Anna Astropolets MSC21,22, Fredrik Nyberg MD21, George Hripcsak MD21,22, Marc A. Suchard MD21,3,22, Dani Prieto-Alhambra MD21,3,22, Peter R Rijnbeek PhD21, Martijn J. Schuemie PhD21,3,22, Patrick B Ryan PhD1,2,3.
Acknoledgement

• The Hyve team
  • Maxim Moinat
  • Sofia Bazakou
  • Anne Van Winzum
  • Alessia Peviani
• Spiros Denaxas
• Erica A Voss
• Daniel Prieto-Alhambra
• Folkert Asselbergs
• Richard Dobson
• Michael Kallfelz
• IMI BigData@Heart
• European Health Data & Evidence Network (EHDEN) project grant
• UCLH NIHR Biomedical Research Centre (BRC)
Thank you!