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
Research and Applications

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 Kalffelz 6, Folkert W. Asselbergs 1,2,7, Daniel Prieto-Alhambra 4,8, 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 NW1 2DA, 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
4. OMOPed data

- Create synthetic data
- Develop ETL
- Refine scripts
Semantic mapping example
<table>
<thead>
<tr>
<th>Results</th>
<th>Source UK Biobank data</th>
<th>OMOP-Transformed UK Biobank data</th>
<th>Transformed UK Biobank COVID-19 positive sub population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>502,505</td>
<td>502,504</td>
<td>3,086</td>
</tr>
<tr>
<td>% Female</td>
<td>54.4</td>
<td>54.4</td>
<td>48.76</td>
</tr>
<tr>
<td>Median age (IQR)</td>
<td>58 (13)</td>
<td>58 (13)</td>
<td>58 (15)</td>
</tr>
<tr>
<td>Median Townsend deprivation index (IQR)</td>
<td>-2.135 (4.18)</td>
<td>-2.135 (4.18)</td>
<td>-1.111 (5.19)</td>
</tr>
<tr>
<td>BMI median - baseline (IQR)</td>
<td>26.652 (5.72)</td>
<td>26.65 (5.70)</td>
<td>27.7 (6.21)</td>
</tr>
<tr>
<td>BMI median - GP EMIS (IQR)</td>
<td>27.2 (6.9)</td>
<td>27.3 (6.84)</td>
<td>28.89 (8)</td>
</tr>
<tr>
<td>SBP median - baseline (IQR)</td>
<td>136 (26)</td>
<td>136 (26)</td>
<td>136 (25)</td>
</tr>
<tr>
<td>DBP median - Baseline (IQR)</td>
<td>81 (14)</td>
<td>81 (14)</td>
<td>82 (14)</td>
</tr>
</tbody>
</table>
# Results

<table>
<thead>
<tr>
<th>Smoking status</th>
<th>Source UK Biobank data</th>
<th>OMOP-Transformed UK Biobank data</th>
<th>Transformed UK Biobank COVID-19 positive sub population</th>
</tr>
</thead>
<tbody>
<tr>
<td>– Not answered</td>
<td>2,276</td>
<td>Not mapped</td>
<td>Not mapped</td>
</tr>
<tr>
<td>– Never</td>
<td>317,891</td>
<td>317,891</td>
<td>1,676</td>
</tr>
<tr>
<td>– Previous</td>
<td>197,949</td>
<td>197,949</td>
<td>1,323</td>
</tr>
<tr>
<td>– Current</td>
<td>55,676</td>
<td>55,676</td>
<td>395</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comorbidities</th>
<th>Source UK Biobank data</th>
<th>OMOP-Transformed UK Biobank data</th>
<th>Transformed UK Biobank COVID-19 positive sub population</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2DM</td>
<td>40,433 (8.04%)</td>
<td>40,476 (8.05%)</td>
<td>453 (14.67%)</td>
</tr>
<tr>
<td>HF</td>
<td>8,068 (1.60%)</td>
<td>8,053 (1.6%)</td>
<td>140 (4.53%)</td>
</tr>
<tr>
<td>AMI</td>
<td>10,593 (2.10%)</td>
<td>10,749 (2.13%)</td>
<td>110 (3.56%)</td>
</tr>
<tr>
<td>COPD</td>
<td>22,364 (4.45%)</td>
<td>22,367 (4.45%)</td>
<td>328 (10.62%)</td>
</tr>
<tr>
<td>HT</td>
<td>175,449 (34.91%)</td>
<td>175,539 (34.93%)</td>
<td>1,571 (50.9%)</td>
</tr>
</tbody>
</table>
Results

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

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

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 PhD4,5, Anna Yin Hui Lai PhD6,7, Claire Blacketer MPH1,2,3, Thamir Alshammari PhD1,2,3, Rupa Makadia PhD1,2,3, Kevin Haynes PharmD3, Gowtham Rao MD2,3, Sebastiaan van Sandijk MSC1,6, Clement Fraboulet MS1,6, Laurence Boyer7,8, Tanguy Le Carrou9,4, Scott Horban BSc Hon5, Daniel R. Morales PhD12,13, Jordi Martinez Roldan MD9, Juan Manuel Ramirez-Anguita PhD9,14, Miguel A. Mayer MD13,14, Marcel de Wildt2,5, Luis H. John MS12,13, Talita Duarte-Salles PhD12,13, Elena Roel MD13, Andrea Pistillo MSC15, Raivo Kolde PhD16, Filip Maljkovic MSC17, Spiros Denaxas PhD18,19,20, Vaclav Papez PhD14,15, Michael G. Kahn MD16,21, Karthik Natarajan PhD22,23, Christian Reich MD24, Alex Secora PhD24, Evan P. Minty MD25, Nigam H. Shah MBBS, PhD26, Jose D. Posada PhD27, Maria Teresa Garcia Morales MS28, Diego Bosca PhD29, Honorio Cadenas Juanino30, Antonio Diaz Holgado30, Miguel Pedrera Jimenez42, Pablo Serrano Balazote30, Noelia Garcia Barrio30, Selcuk Shen MD30, Ali Yagiz Uresin MD30, Baris Erdogan PhD30, Luc Belmans MD30, Geert Byttebier MSC31, Manu L.N.G. Malbrain MD32,33, Daniel J Dedman MPhil34, Zara Cuccu34, Rohit Vashisht PhD35, Atul J. Butte MD36,37, Ayan Patel MS12,37, Lisa Dahm PhD31,32, Cora Han JD3, Fan Bu PhD32, Faaizah Arshad31, Anna Ostropolets MD12,23, Fredrik Nyberg MD10, George Hripcsak MD1,2,3, Marc A. Suchard1,2, Patrick B. Ryan1,2,3, In preparation

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 Rao3,4, Erica A Voss3,4, Anna Ostropolets4,5, Miguel Angel Mayer6, Juan Manuel Ramirez-Anguita7, Filip Maljkovic8, 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 Alshammari4,16, Rupa Makadia3,4, Clair Blacketer3,4, Frank DeFalco3,4, Anthony G Sena3,4, Marc A Suchard4,17, Daniel Prieto-Alhambra18, Patrick B Ryan3,4

Affiliations + expand

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

Evaluating the impact of alternative phenotype definitions on incidence rates across a global data network

Rupa Makadia1,2, Azza Shoaiib1,2, Gowtham Rao1,2, Anna Ostropolets1,3, Peter R. Rijnbeek1,4, Erica A Voss1,2, Talita Duarte-Salles1,5, Juan Manuel Ramirez-Anguita6, Miguel A. Mayer7, Daniel Morales8, Filip Maljkovic9, Spiros Denaxas10, Fredrik Nyberg11, Vaclav Papez12, Clement Fraboulet12, Tanguy Le Carrou13, Anthony G. Sena14, Thamir M Alshammari15, Lana YH Lai16, Kevin Haynes16, Marc A. Suchard1,6, George Hripcsak1,3, Patrick B. Ryan1,2,3
Acknowledgement

• 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!