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

Journal of the American Medical Informatics Association, 00(0), 2022, 1–9
https://doi.org/10.1093/jamia/ocac203
Advance Access Publication Date: 13 October 2022
Research and Applications

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 [3], Stefan Payralbe[3], Michael Kallfelz[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]

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

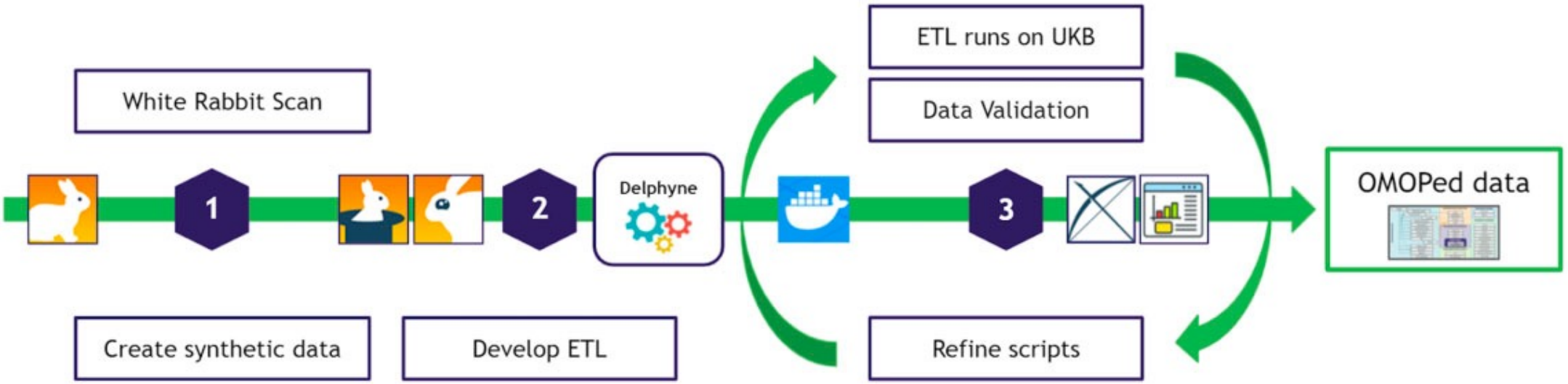
Background

- 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)

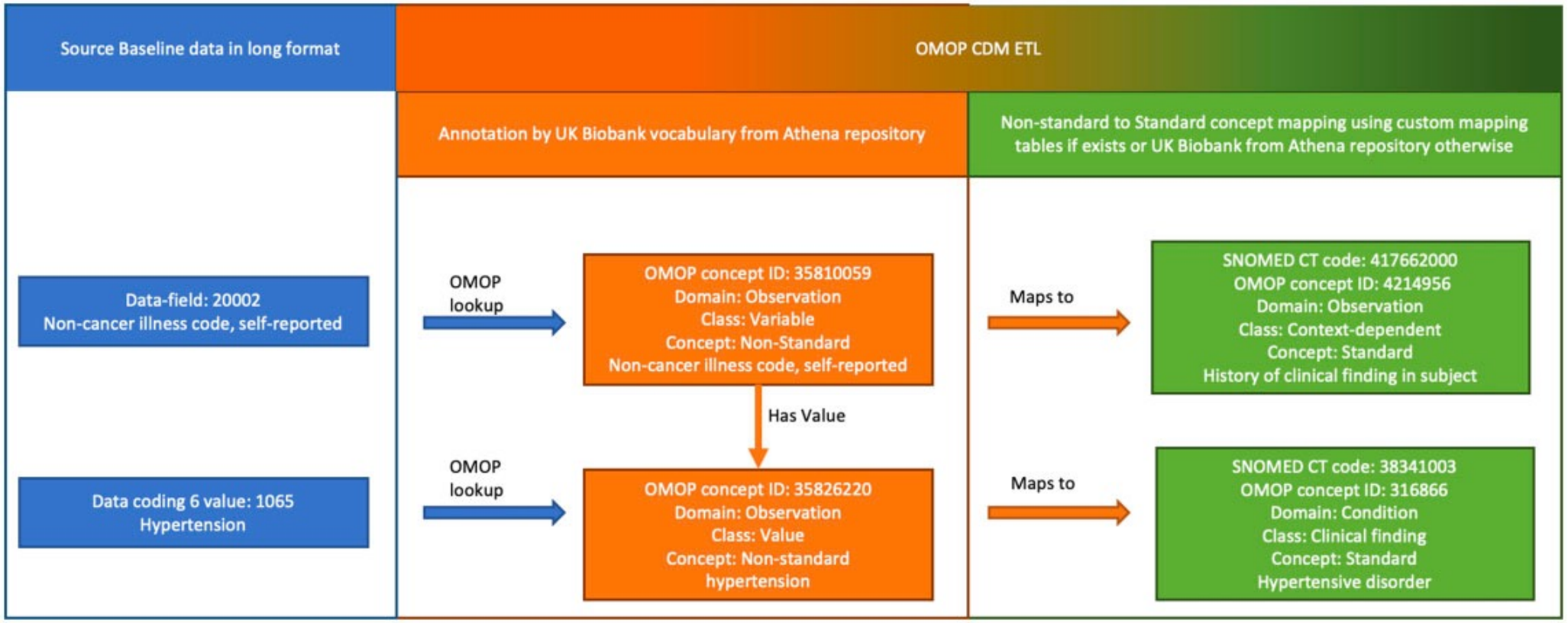
Methods

- 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



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

Published

Observational Study > Drug Saf. 2022 Jun;45(6):685-698. doi: 10.1007/s40264-022-01187-y. Epub 2022 Jun 2.

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

Azza Shoaibi [1 2], Gowtham A Rao [3 4], Erica A Voss [3 4], Anna Ostropolets [4 5], Miguel Angel Mayer [6], Juan Manuel Ramírez-Anguita [6], Filip Maljković [7], Biljana Carević [8], Scott Horban [9], Daniel R Morales [9], Talita Duarte-Salles [10], Clement Fraboulet [11], Tanguy Le Carrour [12], Spiros Denaxas [13], Vaclav Papez [13], Luis H John [14], Peter R Rijneek [14], Evan Minty [15], Thamir M Alshammari [4 16], Rupa Makadia [3 4], Clair Blacketer [3 4], Frank DeFalco [3 4], Anthony G Sena [3 4], Marc A Suchard [4 17], Daniel Prieto-Alhambra [18], Patrick B Ryan [3 4]

Affiliations + expand

PMID: 35653017 PMCID: PMC9160850 DOI: 10.1007/s40264-022-01187-y Paperpile

Free PMC article

In preparation

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

Rupa Makadia[1,2], Azza Shoaibi[1,2], Gowtham Rao[1,2], Anna Ostropolets[1,3], Peter R. Rijnbeek[1,4], Erica A Voss[1,2], Talita Duarte-Salles[1,5], Juan Manuel Ramírez-Anguita[6], Miguel A. Mayer[7], Daniel Morales[8], Filip Maljković[9], Spiros Denaxas[10], Fredrik Nyberg[11], Vaclav Papez[10], Clement Fraboulet[12] ,Tanguy Le Carrour[13], Anthony G. Sena[1,2], Thamir M Alshammari[1,14], Lana YH Lai[1,15], Kevin Haynes[2], Marc A. Suchard[1,16], George Hripcsak[1,3], Patrick B. Ryan[1,2,3]

OHDSI

1 List of Abbreviations
2 Responsible Parties
3 Amendments and Updates
4 Milestones
5 Rationale and Background

RESEARCH PROTOCOL

Adverse Events of Special Interest within COVID-19 Subjects

Version: 1.0.1

Under review

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 MPH[1, 2, 3], Azza Shoaibi PhD[1, 3], Lana Yin Hui Lai PhD[1, 4], Clair Blacketer MPH[1, 2, 3], Thamir Alshammari PhD[1,5], Rupa Makadia PhD[1, 3], Kevin Haynes PharmD[3], Anthony G. Sena BA[1, 2, 3], Gowtham Rao MD[1, 3], Sebastiaan van Sandijk MSc[1,6], Clement Fraboulet MS[7], Laurent Boyer[7], Tanguy Le Carrour[8], Scott Horban BSc Hons[9], Daniel R. Morales PhD[10, 11], Jordi Martínez Roldán MD[12], Juan Manuel Ramírez-Anguita PhD[13, 14], Miguel A. Mayer MD[13, 14], Marcel de Wilde[1, 2], Luis H. John MS[1, 2], Talita Duarte-Salles PhD[1,15], Elena Roel MD[15], Andrea Pistillo MSc[15], Raivo Kolde PhD[16], Filip Maljković MSc[17], Spiros Denaxas PhD[18, 19, 20], Vaclav Papez PhD[18, 19], Michael G. Kahn MD[1, 21], Karthik Natarajan PhD[1, 22, 23], Christian Reich MD[1, 24], Alex Secora PhD[24], Evan P. Minty MD[1, 25], Nigam H. Shah MBBS, PhD[1, 26], Jose D. Posada PhD[1, 27], Maria Teresa Garcia Morales MSc[28], Diego Bosca PhD[41], Honorio Cadenas Juanino[29], Antonio Diaz Holgado[29], Miguel Pedrera Jiménez[42], Pablo Serrano Balazote[42], Noelia García Barrio[42], Selçuk Şen MD[30], Ali Yağız Üresin MD[30], Baris Erdogan PhD[31], Luc Belmans MD[32], Geert Byttebier MSc[32], Manu L.N.G. Malbrain MD[32, 33], Daniel J Dedman MPhil[34], Zara Cuccu[34], Rohit Vashisht PhD[1, 35], Atul J. Butte MD[1, 35, 36], Ayan Patel MS[1, 35], Lisa Dahm PhD[1, 36], Cora Han JD[1, 36], Fan Bu PhD[37], Faaizah Arshad[1, 37], Anna Ostropolets MD[1, 22], Fredrik Nyberg MD[38], George Hripcsak MD[1, 22, 23], Marc A. Suchard MD[1, 37, 39], Dani Prieto-Alhambra MD[1, 2, 40], Peter R Rijnbeek PhD[1, 2], Martijn J. Schuemie PhD[1, 3, 37], Patrick B Ryan PhD[1, 3,22]

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!