*Journal of the American Medical Informatics Association*, 29(3), 2022, 546–552
https://doi.org/10.1093/jamia/ocab260
Advance Access Publication Date: 13 December 2021
Case Report

Case Report

# Curating a longitudinal research resource using linked primary care EHR data—a UK Biobank case study

**Philip Darke [1], Sophie Cassidy [2], Michael Catt[3], Roy Taylor[4], Paolo Missier[1,†], and Jaume Bacardit [1,†]**

[1]School of Computing, Newcastle University, Newcastle upon Tyne, UK, [2]Central Clinical School, The University of Sydney, Sydney, Australia, [3]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK, and [4]Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, UK

Corresponding Author: Philip Darke, FIA, EPSRC Centre for Doctoral Training in Cloud Computing for Big Data, School of Computing, Newcastle University, Urban Sciences Building, 1 Science Square, Newcastle Helix, Newcastle upon Tyne NE4 5TG, UK; p.a.darke2@newcastle.ac.uk
[†]These authors contributed equally to this work.

Received 23 December 2020; Revised 3 November 2021; Editorial Decision 4 November 2021; Accepted 23 November 2021

**ABSTRACT**

Primary care EHR data are often of clinical importance to cohort studies however they require careful handling. Challenges include determining the periods during which EHR data were collected. Participants are typically censored when they deregister from a medical practice, however, cohort studies wish to follow participants longitudinally including those that change practice. Using UK Biobank as an exemplar, we developed methodology to infer continuous periods of data collection and maximize follow-up in longitudinal studies. This resulted in longer follow-up for around 40% of participants with multiple registration records (mean increase of 3.8 years from the first study visit). The approach did not sacrifice phenotyping accuracy when comparing agreement between self-reported and EHR data. A diabetes mellitus case study illustrates how the algorithm supports longitudinal study design and provides further validation. We use UK Biobank data, however, the tools provided can be used for other conditions and studies with minimal alteration.

**Key words:** electronic health records, medical record linkage, longitudinal studies, phenotype, diabetes mellitus

## INTRODUCTION

Access to nonemergency healthcare in the UK is overseen by General Practitioners (GPs). Electronic Health Records (EHRs) maintained by GPs have been used in observational research for over 30 years[1] and are often used in longitudinal cohort studies, for example, to follow participants after the end of data collection.[2,3] However, in contrast with the highly curated data collected under a study protocol, EHR data collection is unstandardized, driven by patient need, and subject to a range of biases.[4] The successful integration of linked EHR data in cohort studies is therefore challenging but increasingly important in a wide variety of research fields.

Using UK Biobank[5] as an exemplar, we developed methodology to incorporate linked primary care EHR data from multiple data providers within a cohort study. We infer periods of data collection in contrast to typical approaches that censor participants when they deregister from a medical practice. Around 40% of participants with multiple registration periods have longer follow-up under this approach without sacrificing phenotyping accuracy when comparing agreement between self-reported data for a range of conditions and medications.

A diabetes mellitus case study was used to demonstrate how the algorithm supports longitudinal study design. Two NHS-approved

diabetes prediction tools performed in line with previous validation studies, further validating the approach. We also contribute extensive supplementary material examining the quality of linked EHR data in UK Biobank. R code is provided to enable researchers to apply the approach to UK Biobank and other cohort studies with linked EHR data.

## STUDY DATA

UK Biobank is a large prospective study of serious illness in middle and old age with longitudinal follow-up achieved primarily through linkage to national data sets.[6] Interim primary care EHR data were released in September 2019 covering around 230 000 participants with subsequent updates available for COVID-19-related research. These were obtained from intermediaries including the suppliers of GP practice management systems and were linked and de-identified by UK Biobank.[7] Participants provided written consent.

Data were recorded by healthcare professionals working at GP practices in England, Scotland, and Wales as part of routine patient care. The interim release included GP practice registration periods, coded diagnoses, test results, drug prescriptions, and administrative data recorded prior to 2016/17. UK Biobank purposefully carried out minimal data cleaning prior to release[7] and the volume of data available varied considerably across individuals.

## CURATING A LONGITUDINAL RESEARCH RESOURCE

The successful integration of linked EHR data required: 1) cleaning and validating the raw data, 2) identifying periods of data collection relative to study visits, and 3) extracting clinically relevant diagnoses, observations, and test results.

### Initial data cleaning

Data were provided from the TPP SystmOne, EMIS Web, and Vision practice management systems, in contrast with large UK EHR repositories which typically use data from a single system. Registration period, clinical event, and prescription record quality were assessed against standards developed from the Clinical Practice Research Datalink "acceptable patient flag"[8] (Supplementary Table S1). Data quality varied by provider (Supplementary Tables S2–S5). Records with missing dates or codes were excluded.

### Identifying periods of EHR data collection

Understanding when data have been collected is essential for longitudinal studies. Existing primary care EHR research often focusses on data recorded after the introduction of the National Health Service (NHS) Quality and Outcomes Framework (QOF) in 2004 which encouraged consistent recording practices across a range of conditions. Practice registration records are typically used determine periods of data collection, for example, by selecting from participants registered with a GP at study start and censoring at practice deregistration. There are limitations when applying this approach to linked EHR data. GP practices began to adopt EHR systems in the 1980s and data recorded prior to the QOF may have clinical importance; while participants may only register with a single NHS practice at a time, records may follow individuals that transfer between practices resulting in the presence of data outside of registration periods; and censoring participants at the first practice deregistration may curtail follow-up (Figure 1). The latter is of particular concern for cohort studies, where a natural objective is to leverage EHR data to follow participants over time.

To address these limitations, an algorithm was developed to identify periods of EHR data collection across practice registration periods (exemplified in Figure 2). Full details are included in Section 2 in the Supplementary Material. A high-level description of the algorithm follows:

1. The period of data collection starts at the first observation/prescription record accompanied by a diagnosis/clinical event record, for example, to capture the recording of a BMI when joining a medical practice.
2. Collection was assumed to be complete until practice deregistration and during any subsequent registration periods.
3. Periods of record collection outside of registration periods were included if they contained at least one nonprescription record. Collection was assumed to have taken place during unregistered periods shorter than 1 year.
4. Participants were censored at the earlier of the inferred end of data collection, the data extract date, or the date of death in linked death registry data.

The algorithm is applied separately for each data provider and the resultant periods combined.

### Extracting clinically relevant data from linked EHRs

UK Biobank data feature a range of coding classifications and multiple data fields for observations and biomarkers[7,9] that must be handled. Existing rule-based phenotyping algorithms aim to replicate diagnostic criteria[10] using clinical code sets to identify relevant exposures and outcomes,[11] however, code set repositories for UK EHR research[2,12,13] typically only cover Read v2 diagnostic codes and limited prescription coding. Comprehensive Read v2, Clinical Terms Version 3 (CTV3), and British National Formulary code sets covering a range of conditions, observations, biomarkers, and drugs were developed (Section 7 in Supplementary Material). Units of measurement were rarely provided in the data, and code descriptions and data dictionaries[14] were often unreliable. An approach was therefore developed to harmonize units (Section 3.2 in Supplementary Material).

## VALIDATING THE PROCESSED EHR DATA

### Assessing the algorithm against GP registration records

A typical approach is to assume full data collection during periods of GP registration. For example, 191 878 participants (83.4% of participants with clinical event data) were registered with a GP at the first UK Biobank study visit with a mean period of 6.9 years to practice deregistration. In contrast, our algorithm maximized study population and follow-up, identifying 196 901 (85.6%) participants with active data collection at the first visit and a mean follow-up of 7.4 years. By design, no participant had a shorter follow-up period under the algorithm. The impact varied by participant, however, the synthetic examples in Figure 1 represent common scenarios:

**Participant 1**

GP registration occurs before the inferred start of data collection for 67% of participants. The mean period of registration before data collection starts is 11.2 years (median 5.8 years). The date of GP registration is often a poor indicator of the start of data collection.

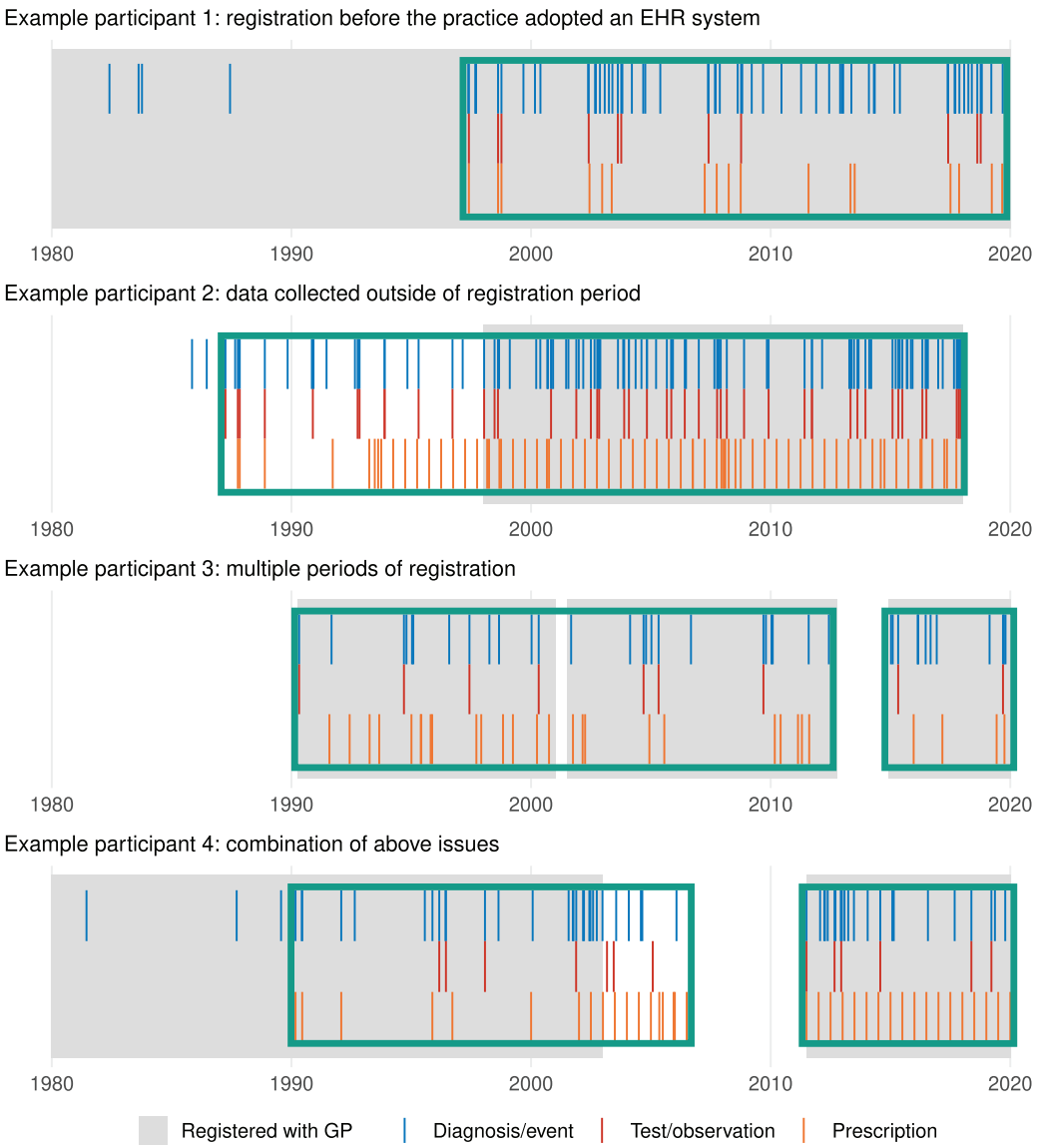


**Figure 1.** Common issues in EHR data collection illustrated with synthetic participant data. These resemble realistic participant types, for example, around 70% of UK Biobank participants have data outside of periods of practice registration. Example 1—individual registered with a practice at birth that subsequently adopted an EHR system in the 1990s (prior records are paper-based). Example 2—individual registered with a practice in 1999 but records are also held from a previous period of registration with another practice. Example 3—multiple periods of registration are available from different practices and/or data providers. Example 4—a combination of the above issues. The boxed areas illustrate the inferred periods of data collection using our algorithm.

### Participant 2

Conversely, 24% of participants show evidence of data collection before the first GP registration (mean 5.8 years of additional data). This may be the result of data transfer when participants move between GP practices. Studies that identify this additional data may be able to use earlier study start dates for example.

### Participant 3

About 31% of participants have multiple registration periods (Supplementary Table S5). Around 40% of these participants have a longer follow-up under our algorithm (mean 3.8 years). As additional linked EHR data are published, the number of participants with multiple periods of registration will increase and methods that follow participants across registrations will be required to maintain follow-up.

## Agreement with self-reported medical conditions and medication

While "ground truth" medical state is typically unavailable, results can be compared with self-reported health in UK Biobank. Participants were phenotyped for selected conditions using EHR data and the results compared with self-reported health at the first study visit (Table 1 and Section 4 in Supplementary Material). The comparison was made for participants with at least 1 year of continuous data collection determined using: 1) our algorithm and 2) assuming data collection only during periods of GP registration. The algorithm generally showed better sensitivity for conditions however the difference between approaches was small. The algorithm therefore maximized study population and follow-up without sacrificing phenotyping accuracy.

The metrics in Table 1 are driven by provider 3 (England TPP) which supplied the majority of linked EHR data. Performance by

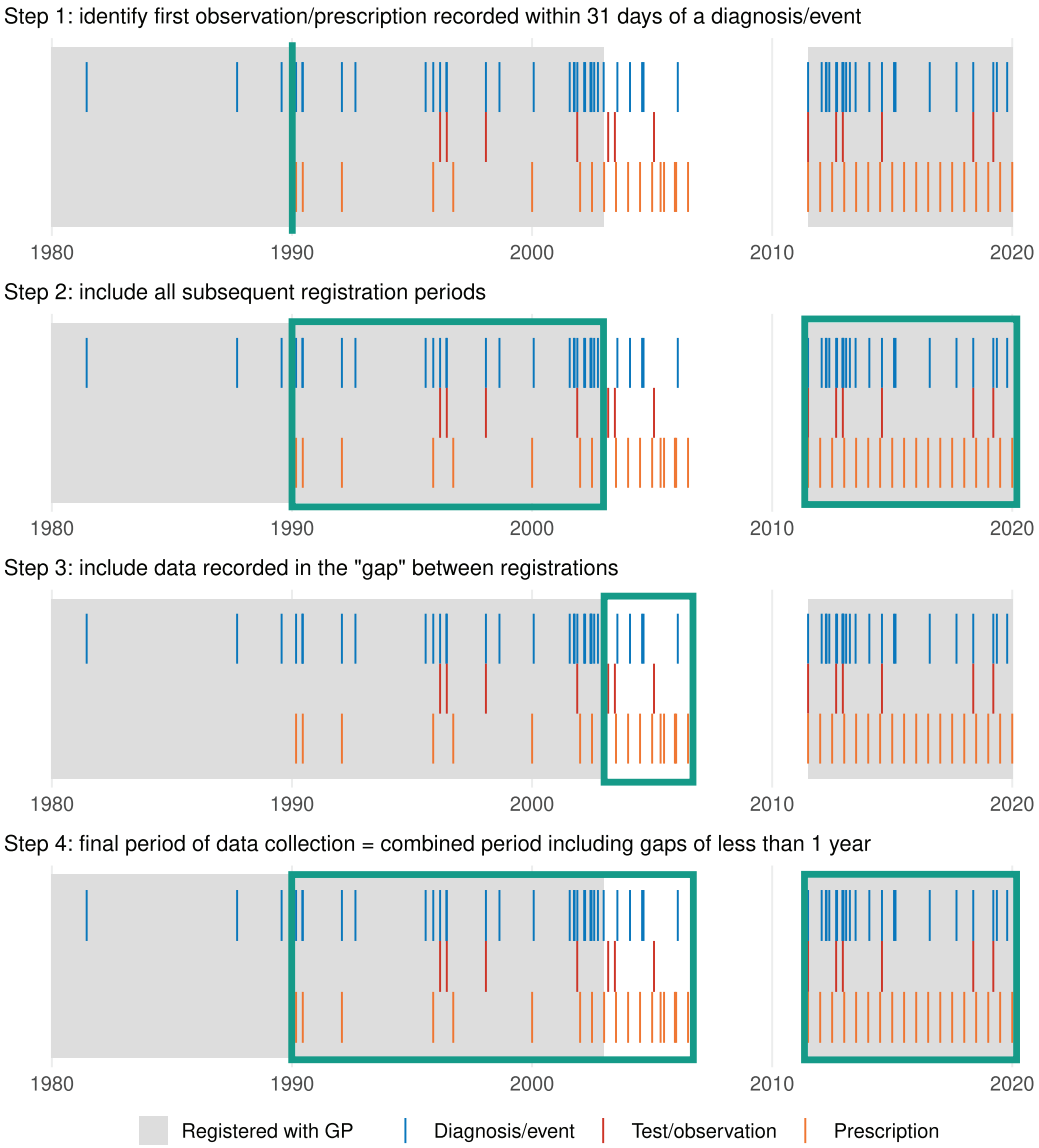


**Figure 2.** Application of our algorithm to determine periods of complete EHR data collection. The example participant has multiple periods of registration and data outside of registration periods. The boxed areas are the inferred periods of data collection. Further details are included in the Supplementary Materials (Algorithm A1 and Supplementary Figure S1).

data provider is provided in Supplementary Tables S8–S11. The algorithm outperformed the use of registration records for provider 1 (England Vision) where the primary issue is identifying the start of data collection. Performance was similar for the remaining providers which featured registration periods that conflict or have gaps suggesting that the algorithm handles these cases well.

Agreement between self-reported and EHR data varied by condition (Section 4 in Supplementary Material). Prescription data appeared to be of lower quality, with evidence of missing or truncated prescription histories beyond the "system-wide block of missing [provider 2 (Scotland) prescription] records prior to 2012".[7] UK Biobank prescription data also features nonstandard coding complicating its use (Section 3.3 in Supplementary Material). Accordingly, agreement with self-reported data was generally lower, however, this may also be the result of prescriptions made outside of primary care (eg, emergency corticosteroids) or inconsistent self-reporting.

## CASE STUDY: LONGITUDINAL DIABETES PHENOTYPING

To demonstrate how our approach supports study design, participants were longitudinally phenotyped for health states associated with diabetes mellitus (diabetes). Diabetes was selected as its diagnosis and management typically takes place in a primary care setting. Diabetes subtypes differ markedly in clinical features but are not widely self-reported in UK Biobank. Previous approaches estimated subtype from self-reported data,[16] however, the newly available linked EHR data potentially offer more objective supporting information. The phenotyping approach was developed with clinical experts (Section 5 in Supplementary Material) but can be readily adapted for other conditions. Figure 3 shows phenotyping tool output for a synthetic participant.

Longitudinal phenotyping is challenging for chronic conditions as multiple diagnosis codes are often recorded over time, for exam-

**Table 1.** Agreement between self-reported and EHR data at the first UK Biobank visit for the conditions and medications used in the QDiabetes-2018 model[15]

| Active data collection at first UK Biobank visit determined using: | Our algorithm (Algorithm A1 in Supplementary Materials) | | | GP registration records | | |
|---|---|---|---|---|---|---|
| | Sensitivity | Specificity | Precision | Sensitivity | Specificity | Precision |
| Presence of a previous diagnostic record | | | | | | |
| Diabetes | **94.4** | 99.8 | 95.8 | 94.3 | 99.8 | **95.9** |
| Hypertension | **72.2** | 98.1 | 93.2 | 72.1 | 98.1 | **93.3** |
| MI/heart attack | 70.6 | 99.9 | 94.7 | **70.7** | 99.9 | **94.8** |
| Angina | 59.8 | 99.4 | 76.4 | **59.9** | 99.4 | **76.5** |
| Stroke | **55.8** | 99.6 | **65.2** | 55.4 | 99.6 | 64.9 |
| Transient ischemic attack | **56.0** | 99.4 | 23.7 | 55.8 | 99.4 | 23.7 |
| Bipolar disorder | **67.2** | 99.7 | 41.3 | 67.1 | **99.8** | **41.8** |
| Schizophrenia | **87.4** | 99.8 | 28.4 | 87.0 | 99.8 | **29.1** |
| Polycystic ovarian syndrome | **57.3** | 99.8 | 22.4 | 57.0 | 99.8 | **23.0** |
| Presence of a prescription record in previous 90 days | | | | | | |
| Antihypertensives | 86.0 | 98.2 | 93.6 | **86.2** | 98.2 | **93.7** |
| Statins | 88.1 | 97.9 | 89.0 | **88.2** | 97.9 | 89.0 |
| Corticosteroids | 49.6 | 99.3 | 45.1 | **49.8** | 99.3 | **45.2** |
| Atypical antipsychotics | 79.7 | 100.0 | 85.6 | **80.4** | 100.0 | 85.6 |

*Note:* Agreement was defined as the presence of a diagnostic record prior to the visit for medical conditions, or the presence of a prescription record in the 90 days prior to the visit for current medication. *Sensitivity* is the proportion of self-reporting participants that have a confirmatory EHR record. *Specificity* is the proportion of participants that do not self-report that also do not have an EHR record. *Precision* is the proportion of participants with an EHR record that also self-report. Overall agreement was similar under each approach, indicating that the algorithm did not sacrifice phenotyping accuracy. Bold indicates higher value.

ple, at annual care reviews. The first instance of a code may not correspond to the date of diagnosis if data collection is incomplete. We aimed to minimize the risk of misidentifying the date of incidence by restricting phenotyping to periods of complete data collection as inferred by our algorithm.

To further validate our algorithm, the performance of two NHS-approved diabetes prediction tools was evaluated on the processed data. QDiabetes-2018[15] and the Leicester Risk Assessment score[17] are used to triage individuals at risk of diabetes.[18] QDiabetes was developed using primary care EHR data and results are shown in Table 2. Leicester score results are presented in Section 6.2 of the Supplementary Materials. Both scores performed broadly in line with previous validation studies, indicating that the data processed is suitable for use in longitudinal diabetes studies. This reinforces recent work suggesting that risk factor associations in UK Biobank are generalizable across a range of conditions[19] despite the healthier, less-deprived, and less diverse ethnic make-up relative to the UK population.[20]

## CONCLUSION

Linked EHR data can be a valuable source of data to cohort studies. An approach was presented to integrate linked primary care EHR data within a cohort study. This maximizes study populations and follow-up using a rule-based approach to determine periods of EHR data collection for each participant. The processed UK Biobank EHR data showed good agreement with self-reported health status and NHS-approved diabetes prediction tools performed well in a longitudinal study, validating the approach and demonstrating how linked EHR data can be used in study design.

We provide extensive supplementary material examining the quality of linked EHR data in UK Biobank. Tools are also provided to implement our approach. These are designed to be general-purpose and support a range of study designs. The approach generalizes to other medical conditions and studies using linked EHR data.

## FUNDING

This work was supported by the Engineering and Physical Sciences Research Council, Centre for Doctoral Training in Cloud Computing for Big Data, Newcastle University (grant number EP/L015358/1).

## AUTHOR CONTRIBUTIONS

PD conceptualized the work, generated the code sets, processed the data, analyzed the results, drafted the manuscript, and is the guarantor. SC and RT provided diabetes-specific input and reviewed the clinical code sets. JB and PM supervised the research. All authors critically revised the manuscript and approved the final version.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online. R code including participant plotting tools are available at https://github.com/philipdarke/ukbb-ehr-data.

## ACKNOWLEDGMENTS

The data used were provided under UK Biobank application 12184. The authors are grateful to all UK Biobank participants who generously contributed their time to the study. PD would like to thank Dr Peter Philipson at Newcastle University, Dr Sam Hodgson at University of Southampton, Dr Sarah Finer at Queen Mary University of London, and Professor Naomi Allen and Dr Rishi Caleyachetty at Oxford University/UK Biobank for helpful discussions.

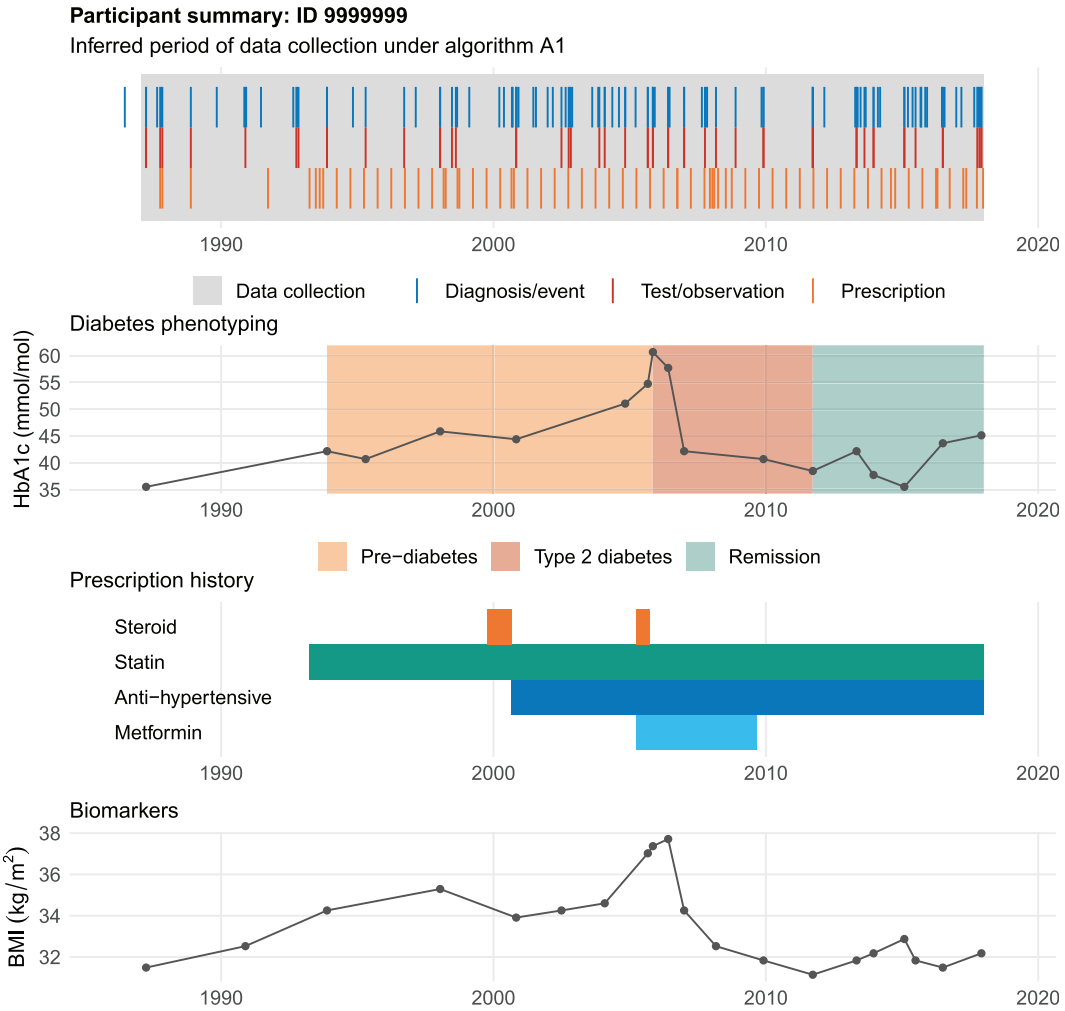


**Figure 3.** Example output from the longitudinal phenotyping tool for a synthetic participant. Our algorithm was used to identify periods of complete data collection (top panel). Periods of nondiabetic hyperglycemia (prediabetes), type 2 diabetes, and remission were identified. Periods of medication and biomarkers are also shown. We phenotyped periods of complete data collection to reduce the risk of inaccurately identifying the date of incidence of diabetes. Similar phenotyping approaches using linked EHR data can be used to enforce study criteria or identify more complex endpoints.

**Table 2.** QDiabetes-2018 model performance (concordance index) for the 10-year incidence of diabetes using UK Biobank EHR data

| | Model A (demographic data, medical history, and BMI) | Model B (A plus current fasting plasma glucose result) | Model C (A plus current HbA1c result) |
|---|---|---|---|
| Male | | | |
| UK Biobank | 0.781 | 0.831 | 0.882 |
| QResearch[15] | 0.814 | 0.866 | 0.855 |
| Female | | | |
| UK Biobank | 0.832 | 0.877 | 0.904 |
| QResearch[15] | 0.834 | 0.889 | 0.878 |

*Note:* Performance on the integrated linked EHR data is broadly in line with Hippisley-Cox et al.[15] (shown as QResearch).

## CONFLICT OF INTEREST STATEMENT

None declared.

## DATA AVAILABILITY

The data used are available from the UK Biobank under the arrangements detailed at https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access.

## REFERENCES

1. Wolf A, Dedman D, Campbell J, *et al.* Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. *Int J Epidemiol* 2019; 48 (6): 1740.
2. Finer S, Martin HC, Khan A, *et al.* Cohort profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *Int J Epidemiol* 2020; 49 (1): 20–21i.
3. Koivula RW, Forgie IM, Kurbasic A, *et al.* Discovery of biomarkers for glycaemic deterioration before and after the onset of type 2 diabetes: de-

scriptive characteristics of the epidemiological studies within the IMI DIRECT Consortium. *Diabetologia* 2019; 62 (9): 1601–15.
4. Goldstein BA, Navar AM, Pencina MJ, *et al.* Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017; 24 (1): 198–208.
5. Sudlow C, Gallacher J, Allen N, *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* 2015; 12 (3): e1001779.
6. Allen N, Sudlow C, Downey P, *et al.* UK Biobank: current status and what it means for epidemiology. *Health Policy Technol* 2012; 1 (3): 123–6.
7. UK Biobank. Resource 591: primary care data. 2020. http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=591 Accessed 15 May 2020.
8. Booth H, Dedman D. Achim Wolf. CPRD Aurum: frequently asked questions. 2019. https://cprd.com/sites/default/files/CPRD%20Aurum%20FAQs%20v2.0_2.pdf Accessed 18 May 2020.
9. Denaxas S, Shah AD, Mateen BA, *et al.* A semi-supervised approach for rapidly creating clinical biomarker phenotypes in the UK Biobank using different primary care EHR and clinical terminology systems. *JAMIA Open* 2020; 3 (4): 545–56.
10. Spratt SE, Pereira K, Granger BB, *et al.* Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. *J Am Med Inform Assoc* 2017; 24 (e1): e121–8.
11. Williams R, Kontopantelis E, Buchan I, *et al.* Clinical code set engineering for reusing EHR data for research: a review. *J Biomed Inform* 2017; 70: 1–13.
12. CALIBER Research. CALIBER codelists. https://www.caliberresearch.org/portal/codelists Accessed 14 July 2020.
13. Springate DA, Kontopantelis E, Ashcroft DM, *et al.* ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PLoS One* 2014; 9 (6): e99825.
14. UK Biobank. Resource 951: COVID numeric reference codes. 2021. http://biobank.ndph.ox.ac.uk/showcase/refer.cgi?id=951 Accessed 23 November 2021.
15. Hippisley-Cox J, Coupland C. Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. *BMJ* 2017; 359: j5019.
16. Eastwood SV, Mathur R, Atkinson M, *et al.* Algorithms for the capture and adjudication of prevalent and incident diabetes in UK Biobank. *PLoS One* 2016; 11 (9): e0162388.
17. Gray LJ, Taub NA, Khunti K, *et al.* The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diab Med* 2010; 27 (8): 887–95.
18. National Institute for Health and Clinical Excellence. Type 2 diabetes: prevention in people at high risk (PH38). 2017. https://www.nice.org.uk/guidance/ph38 Accessed 30 January 2020.
19. Batty GD, Gale CR, Kivimäki M, *et al.* Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ* 2020; 368: m131.
20. Fry A, Littlejohns TJ, Sudlow C, *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol* 2017; 186 (9): 1026–34.