Pharmacoepidemiology and Drug Safety

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Standardised and Reproducible Phenotyping Using Distributed Analytics and Tools in the Data Analysis and Real World Interrogation Network (DARWIN EU)

Francesco Dernie[1,2] | George Corby[1,2] | Abigail Robinson[1,2] | James Bezer[1,2] | Nuria Mercade-Besora[2] | Romain Griffier[3] | Guillaume Verdy[3] | Angela Leis[4] | Juan Manuel Ramirez-Anguita[5] | Miguel A. Mayer[4,5] | James T. Brash[6] | Sarah Seager[6] | Rowan Parry[7] | Annika Jodicke[2] | Talita Duarte-Salles[7,8] | Peter R. Rijnbeek[7] | Katia Verhamme[7] | Alexandra Pacurariu[9] | Daniel Morales[9,10] | Luis Pinheiro[9] | Daniel Prieto-Alhambra[2,7] | Albert Prats-Uribe[2]

[1]Medical Sciences Division, University of Oxford, Oxford, UK | [2]Pharmaco- and Device Epidemiology, Centre for Statistics in Medicines, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK | [3]Public Health Department, Medical Information Service, Medical Informatics and Archiving Unit (IAM), University Hospital of Bordeaux, Bordeaux, France | [4]Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Research Institute, Barcelona, Spain | [5]Management and Control Department, Consorci Mar Parc de Salut de Barcelona, Barcelona, Spain | [6]Real World Solutions, IQVIA, Brighton, UK | [7]Department of Medical Informatics, Erasmus University Medical Centre, Rotterdam, The Netherlands | [8]Fundació Institut Universitari per a la Recerca a l'Atencio Primaria de Salut Jordi Gol I Gurina (IDIAPJGol), Universitat Autonoma de Barcelona, Barcelona, Spain | [9]Real World Evidence Workstream, European Medicines Agency, Amsterdam, The Netherlands | [10]Division of Population Health and Genomics, University of Dundee, Dundee, UK

Correspondence: Daniel Prieto-Alhambra (d.prietoalhambra@darwin-eu.org)

Received: 10 January 2024 | **Revised:** 2 October 2024 | **Accepted:** 6 October 2024

Funding: This work is part of the DARWIN EU initiative, funded by the European Medicines Agency. Francesco Dernie, Annika Jodicke, Daniel Prieto-Alhambra and Albert Prats-Uribe receive partial support from the National Institute for Health and Care Research (NIHR) in the form of the Oxford NIHR Biomedical Research Centre.

Keywords: pancreatic cancer | phenotyping | systemic lupus erythematosus

ABSTRACT

Purpose: The generation of representative disease phenotypes is important for ensuring the reliability of the findings of observational studies. The aim of this manuscript is to outline a reproducible framework for reliable and traceable phenotype generation based on real world data for use in the Data Analysis and Real-World Interrogation Network (DARWIN EU). We illustrate the use of this framework by generating phenotypes for two diseases: pancreatic cancer and systemic lupus erythematosus (SLE).

Methods: The phenotyping process involves a 14-steps process based on a standard operating procedure co-created by the DARWIN EU Coordination Centre in collaboration with the European Medicines Agency. A number of bespoke R packages were utilised to generate and review codelists for two phenotypes based on real world data mapped to the OMOP Common Data Model.

Results: Codelists were generated for both pancreatic cancer and SLE, and cohorts were generated in six OMOP-mapped databases. Diagnostic checks were performed, which showed these cohorts had broadly similar incidence and prevalence figures to previously published literature, despite significant inter-database variability. Co-occurrent symptoms, conditions, and medication use were in keeping with pre-specified clinical descriptions based on previous knowledge.

Conclusions: Our detailed phenotyping process makes use of bespoke tools and allows for comprehensive codelist generation and review, as well as large-scale exploration of the characteristics of the resulting cohorts. Wider use of structured and reproducible phenotyping methods will be important in ensuring the reliability of observational studies for regulatory purposes.

Summary

- Routinely collected health data including computerised medical records, health claims, and registries among others, contain increasing numbers of patients worldwide and gather valuable data for research. Creating reliable cohorts of patients with a disease of interest is dependent on the quality of the process of identifying diseases in data form (their 'phenotype')
- We outline a framework co-created with the European Medicines Agency in the context of the Data Analysis and Real-World Interrogation Network (DARWIN EU).
- Our process aims to produce reliable disease phenotypes through a standardised, reproducible, and transparent framework, spanning code list generation, clinical reviews, and quality processes to evaluate the face validity of the resulting phenotypes when compared to a pre-specified clinical description based on current clinical knowledge.
- We illustrate the use of our framework by phenotyping systemic lupus erythematosus and pancreatic cancer and testing the resulting computable phenotypes in six large databases of primary care records, hospital records, and health claims data.

1 | Background

A phenotype is the set of observable and measurable characteristics of an individual, and can be caused or modified by disease. In the context of data science, Hripcsak and Albers defined 'a phenotype [as] a specification of an observable, potentially changing state of an organism, [...]. Phenotype algorithms—that is, algorithms that identify or characterise phenotypes—may be generated by domain experts and knowledge engineers [...] to generate novel representations of the data' [1].

Increasingly available rich data sources provide many opportunities to investigate relationships between diseases, medicinal products, and outcomes. Common data models (CDM) like the Observational Health Data Sciences and Informatics (OHDSI) [2] maintained Observational Medical Outcomes Partnership (OMOP) CDM [3] harmonise diverse data sources to enable federated analytics.

When using routinely collected healthcare data, multiple data domains and codes may be used to define a disease phenotype in data form. For example, diabetes mellitus [4] may be represented as a diagnostic code, or a combination of HbA1c levels, antidiabetic medicine/s use, and/or complications. The process of creating and validating a disease phenotype in data and generating a cohort of patients with that disease, or 'phenotyping', requires a systematic and transparent process to ensure the reproducibility and reliability of results [5].

In this paper we outline a step-by-step fully reproducible and traceable process for phenotyping based on OMOP-mapped real world data, and demonstrate how this can be used to generate computable phenotypes (from now on, phenotypes) for two illustrative examples: systemic lupus erythematosus (SLE) and pancreatic cancer. The process defined in this manuscript relates to the standard operating procedure developed by the Data Analysis and Real-World Interrogation Network (DARWIN EU) [6] Coordination Centre in collaboration with the European Medicines Agency (EMA).

2 | Methods

2.1 | Data Source and Study Populations

We illustrated the proposed process step-by-step using two examples and six distinct databases mapped to the OMOP CDM (Table 1). A flowchart summary of the whole process can be found in Figure S1.

TABLE 1 | Data sources used for cohort generation following phenotype generation.

Name of database	Country	Type of data	Time span	Number of active subjects
Clinical Practice Research Datalink (CPRD) GOLD—January 2023 version [7]	United Kingdom	Primary care general population data	09/1987 to 06/2023	17.2 million people
IQVIA PharMetrics Plus for Academics [8]	United States of America	Claims-based data	01/2017 to 06/2022	34.8 million people
Institut Municipal Assistència Sanitària Hospital del Mar Information System (IMASIS)	Spain	Hospital	01/1990 to 09/2023	1.0 million people
IQVIA Disease Analyzer Germany (IQVIA DA Germany)	Germany	Primary care general population	01/1992 to 09/2023	43.0 million people
Clinical Data Warehouse of Bordeaux University Hospital (CDWBordeaux)	France	Hospital	01/2005 to 02/2024	2.2 million people
Integrated Primary Care Information (IPCI) database [9]	The Netherlands	Primary care general population	01/2006 to 06/2023	2.8 million people

Note: Full metadata and further information on these databases is available on https://darwin-eu.org/index.php/data/data-network.

2.2 | Steps 1–3: Proposing a Phenotype, Assessing Feasibility and Optimisation

2.2.1 | Step 1: Phenotype Proposal and Clinical Description

A Phenotype Proposal Form (PPF) is prepared by the principal investigator (PI), in collaboration with an expert phenotype lead (PL) and/or phenotype assistants (PA) (Figure S2). The PPF includes information on the requestor, a summary of the phenotype, the intended use (e.g., for incidence/prevalence studies, or drug utilisation, and the databases it will be used on) and timelines, and a clinical description, summarising the disease's epidemiology, presenting symptoms, treatments, and potential strengthening or disqualifying factors (e.g., particularly sensitive/specific blood tests which increase the likelihood of an individual having the condition of interest).

2.2.2 | Step 2: Feasibility

The PL assesses the feasibility of generating the phenotype by ascertaining the possibility of using current vocabularies (code dictionaries) to generate it, and the real-world data sources available. If unfeasible, the reasons are documented in the PPF to avoid similar future efforts. This could be done using prior knowledge or via a quick search in the vocabulary repository (athena.ohdsi.org).

2.2.3 | Step 3: Optimisation Requirements

The phenotyping team can determine what further information may optimise the phenotype, including variability in disease presentation and its intended use. Different phenotype 'variants' (i.e., to increase specificity or sensitivity or represent different disease sub-types) may be required. Also, one phenotype may need more than one concept sets and different complex algorithms; for example, we may want to define a diabetes phenotype based on certain requirements and temporality on diagnosis (conditions), use of antidiabetic medication and insulin (drugs) and HbA1c levels (measurements).

It is important to note, that, although in this paper we demonstrate the process for outcomes and with a single concept set for simplicity, in practise, different variants may be needed if the intended use of the phenotypes is as an eligibility criterion or as an exposure; and different variants with multiple criteria and more complex algorithms could also be created for these phenotypes.

2.3 | Steps 4 and 5: Evaluating Existing Phenotypes

Steps 4 and 5 are conducted using a repository of phenotype definitions and lists of codes. This repository, known as DECK, is a custom web application, which allows code list reviews, versioning, and storage of completed phenotypes. All exchanges between the phenotyping team are conducted on the discussion tab of the clinical definition in DECK for traceability (Figure 1).

2.3.1 | Step 4: Search for a Suitable Phenotype

PLs and PAs check whether a suitable phenotype already exists within DECK or in other repositories of phenotypes based on OMOP data, such the OHDSI phenotype library, [10] that can be reused or optimised for a given study.

2.3.2 | Step 5: Suitability and Relevance to Study of Interest

If the PL identifies a potentially suitable phenotype in DECK, they will assess its relevance to the proposed study, documenting the reasoning behind this decision. Although not yet fully specified, some example considerations regarding suitability are shown in Box 1.

a. If the decision is that the existing phenotypes are not suitable, then the PL will document and jump to Step 6.

b. If the decision is that a phenotype is suitable for the proposed use, then the PL will document and jump to Step 13.

c. If the decision is that the phenotype may be suitable, the PL would jump to Step 10.

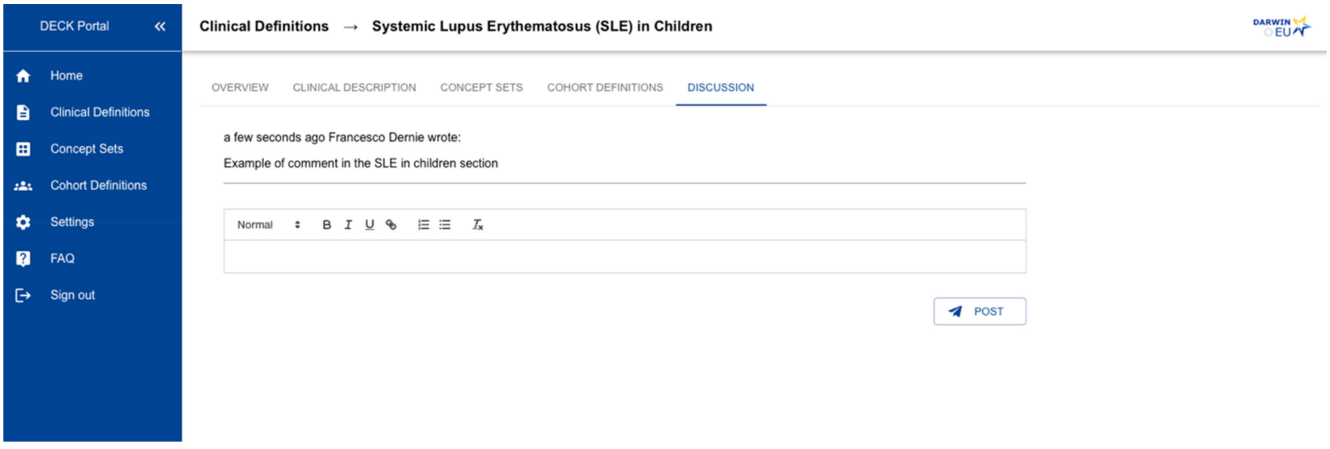

FIGURE 1 | Screenshot of the discussion tab of the SLE in children clinical definition on the DECK Portal software.

BOX 1 | Criteria used to assess suitability of an existing/stored phenotype.

Criteria to determine its suitability:

- Date when the phenotype were generated
- Vocabulary version used for the development (e.g., OMOP Version 5.0)
- Previous use and previous optimization strategy
- Previous publication/s that can be used as reference
- Previous validation efforts
- Cohort diagnostics (see Step 10): date and data partners involved. Link to results and/or to that phenotype diagnostics evaluation if available
- Other/s

2.4 | Steps 6–9: New Phenotype Generation

2.4.1 | Step 6: Evaluate Existing Concept Sets or Code Lists

The PL will query the DECK to search for the list of diagnostic codes—for example 'diabetes mellitus' or 'metformin' needed for the phenotype and evaluate its suitability for the current phenotype. These can be in a form of plain lists of codes, or of concept sets: simplified code lists that make use of the hierarchical structure of the vocabularies, for example, to describe a certain code plus all 70 codes below in the hierarchy but one, one can define it as a list of two codes and two attributes instead of 70 codes: the parent code (plus the including descendants attribute) and the code of the excluded descendant (plus the excluded attribute). If new concept sets are needed, they proceed to Step 7, otherwise jump to Step 10.

2.4.2 | Step 7: Generate the Concept Set

The PL generates a first provisional list of diagnostic codes (aka concept IDs) using relevant tools like CodeListGenerator (https://github.com/darwin-eu/CodelistGenerator) or PHOEBE [11]. In some instances, the PL will consult clinicians, pharmacists, and/or topic experts. The omission of this consultation will be logged with reasoning/justification. The search strategy used is logged and stored for future reproducibility. The resulting output will be a list of potential concept IDs to be evaluated in the following steps.

2.4.3 | Step 8: Refine the Concept Set

The PL will identify at least two reviewers with a clinical/pharmacy background (PAs) and/or topic expertise and present them with the initial concept set from Step 7 together with the provided PPF and clinical description (Steps 1–3). DECK will be used to facilitate the transparent, user-friendly, and fully traceable review of the codes by both reviewers, with a full log of the decisions made and discussions. The reviewers will review the initial concept set and include or exclude codes from the concept set based on information from the clinical description and their own expertise, determining whether a particular code fits the proposed disease or not. Codes may be included in one variant (Step 3) but not another. Any discrepancies will be resolved in discussion with the PL. The reviewer's choices and the final list are stored as versions inside DECK and the choices are documented.

2.4.4 | Step 9: Generating the Cohort From the Phenotype

Once the concept set/s have been finalised after Step 9, the PL will construct the algorithm that defines the intended phenotype and identifies a set of people with the disease in each dataset, the cohorts. This phenotype algorithm, or 'cohort definition' includes the concept set/s generated as inclusion criteria and additional logic, if necessary, based on the PPF and clinical description (Steps 1–3). This can be done in different tools, for example: ATLAS software [12] or programmatically using the CapR R package [13] or Cohort. In both cases, the result is stored in a json file, and it is saved in the DECK with its metadata for future reference.

2.5 | Steps 10 and 11: New Phenotype Evaluation

In Steps 10 and 11, we ensure the generated cohorts are trustworthy representations of the intended phenotypes, running diagnostic checks and documenting these.

2.5.1 | Step 10: Performing Phenotype Diagnostics

The PL will decide which diagnostics are necessary to run and in which contributing data sources on a per phenotype-study basis. Diagnostic packages like CohortDiagnostics [14] allow exploration of generated cohorts, and comparison of these with clinical knowledge and published literature. Examples of measurements examined during the diagnostic process may include, but are not limited to, cohort counts (number of subjects by phenotype variant), code counts (number of subjects for each included diagnostic code within the phenotype), orphan codes (suggestions of codes for inclusion), age/sex distribution, population incidence and prevalence, and large-scale characterisation (medication use, clinical conditions, and laboratory measurement codes at various time periods prior to and following their first diagnostic code of the phenotype of interest). Ideally, characterisation of a background population of similar age-sex should be performed to allow for meaningful benchmarking.

For each database, we calculated incidence and prevalence for a sample of 1 million patients, and other large-scale characterisations in a more limited sample of 1000 individuals with pancreatic cancer or SLE, comparing these to an age and sex-matched comparison cohort using existing DARWIN EU analytical packages detailed below.

2.5.2 | Step 11: Evaluating the Phenotype Diagnostics

Diagnostics will be reviewed by the phenotyping team and/or clinical experts. A standardised form will be filled in by the reviewers (Figure S2), including examination of the concept IDs included from each data source, and review of the parameters outlined in Step 10. If some of the diagnostics are considered not passed, reviewers will add recommendations for improvement. PL will then decide to iterate through Steps 8–11 until a satisfactory phenotype has been developed or/and recommend some data partners as not suitable for identifying that phenotype.

2.6 | Steps 12 and 13: Signing Off for Study Execution

2.6.1 | Step 12: Phenotype Versioning and Storage

The PL will create a new not editable version on the DECK of the phenotype and mark it as ready for use once all previous steps are completed.

2.6.2 | Step 13: Sign Off for Study Execution

Once the process is completed, the PL will sign off the phenotype definition as ready for study execution, mark the version as study version and notify the PI of the study.

2.7 | Step 14: Review and Log of Study Learnings

2.7.1 | Step 14: Review of Study Learnings

Comments on challenges from running the study, the data sources the phenotype were used in, and the study it was used for, will be added by the PI to the cohort comments in the DECK.

2.7.2 | Software

R Version 4.2.3 was used for all analyses. Several bespoke open software packages developed by DARWIN EU were utilised including CodeListGenerator [15], PatientProfiles [16], IncidencePrevalence [17] and the DECK web portal, currently in development.

3 | Results

We used our phenotyping process to generate new phenotypes for pancreatic cancer and SLE. The key decisions taken at each stage and results from our phenotype diagnostics are presented below.

3.1 | Pancreatic Cancer Phenotype

3.1.1 | Pancreatic Cancer: Steps 1–3

The phenotype was deemed feasible and adequately represented within the vocabulary of diagnostic codes (OMOP V5.0). A clinical description was generated by a PA (Figure S4) and variants were determined after initial exploration of the topic area between the PA and PL. Six phenotype variants that could be defined with condition codes were defined (Table 2).

3.1.2 | Pancreatic Cancer: Steps 4 and 5

No suitable phenotype existed already in DECK.

3.1.3 | Pancreatic Cancer: Steps 6–9

A provisional list of concept IDs was generated using CodeListGenerator (Figure S4). An additional search was done, using the Medical Subject Headings (MeSH) Descriptor D010190 (Pancreatic Neoplasms), but revealed no further terms appropriate for inclusion. The PA performed the initial review of this concept set, followed by second review by the PL using the DECK Portal (Figure 2). There was 100% agreement between the reviewers regarding code inclusion across the six variants. A total of 129 codes were initially identified, of which 94 were included in the concept set (Table S1).

Figure 3 illustrates the inclusion of codes into each variant as per the rationale in Table 2. The full list of included and excluded codes are presented in Table S1.

TABLE 2 | Pancreatic cancer phenotype variants and their rationale.

Variant name	Rationale
Broad	Aim to detect all records for pancreatic cancer, with very high sensitivity but lower specificity
Narrow	Aim to detect records for pancreatic cancer, with very high specificity, but lower sensitivity
Broad endocrine	Aims to detect all records deemed to affect the endocrine pancreas therefore excluding for instance 'neoplasm of exocrine pancreas'
Broad exocrine	As above but for exocrine cancers
Narrow endocrine	Aims to only detect records that had a high likelihood of endocrine cancer—so for instance 'neoplasm of endocrine pancreas' but not neoplasm of pancreas
Narrow exocrine	As above but for exocrine cancers

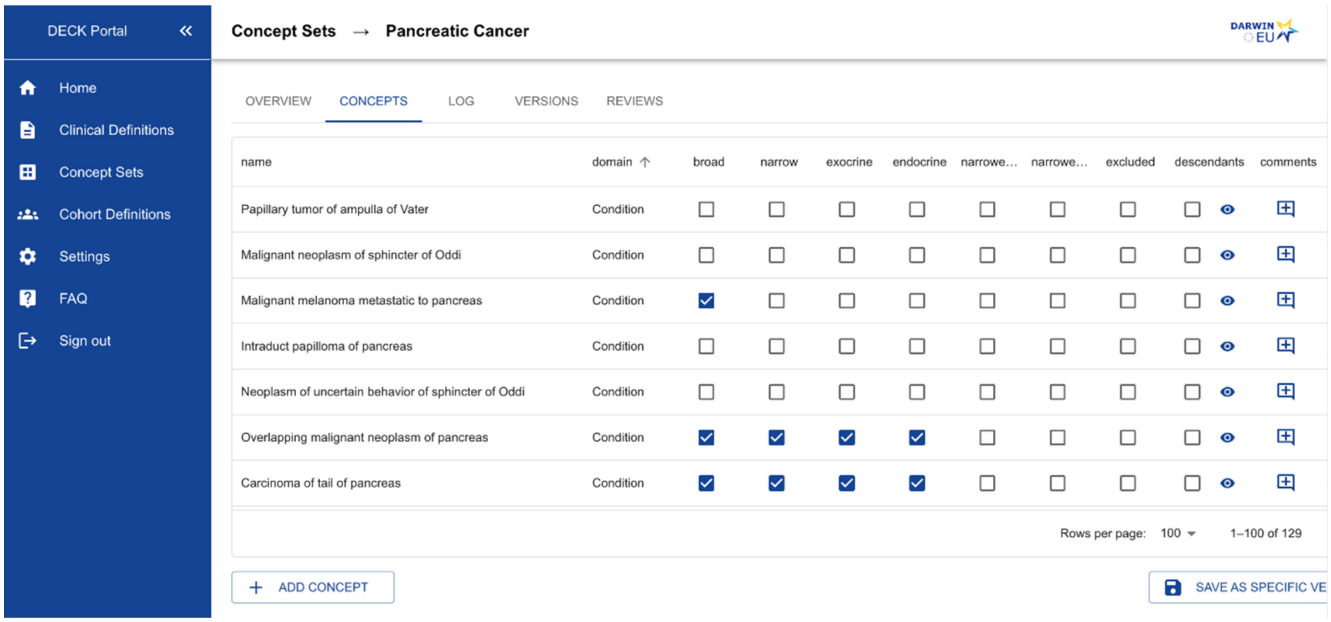

FIGURE 2 | Screenshot of the DECK Portal software for the review of codes for different variants (in this case broad, narrow, exocrine, endocrine, narrow endocrine and narrow exocrine) of pancreatic cancer phenotypes. Each Concept ID can be included in none, some or all variants. The comments tab allows investigators to leave a comment about a particular Concept ID, for instance explaining their reasoning for including it in a particular variant if necessary.

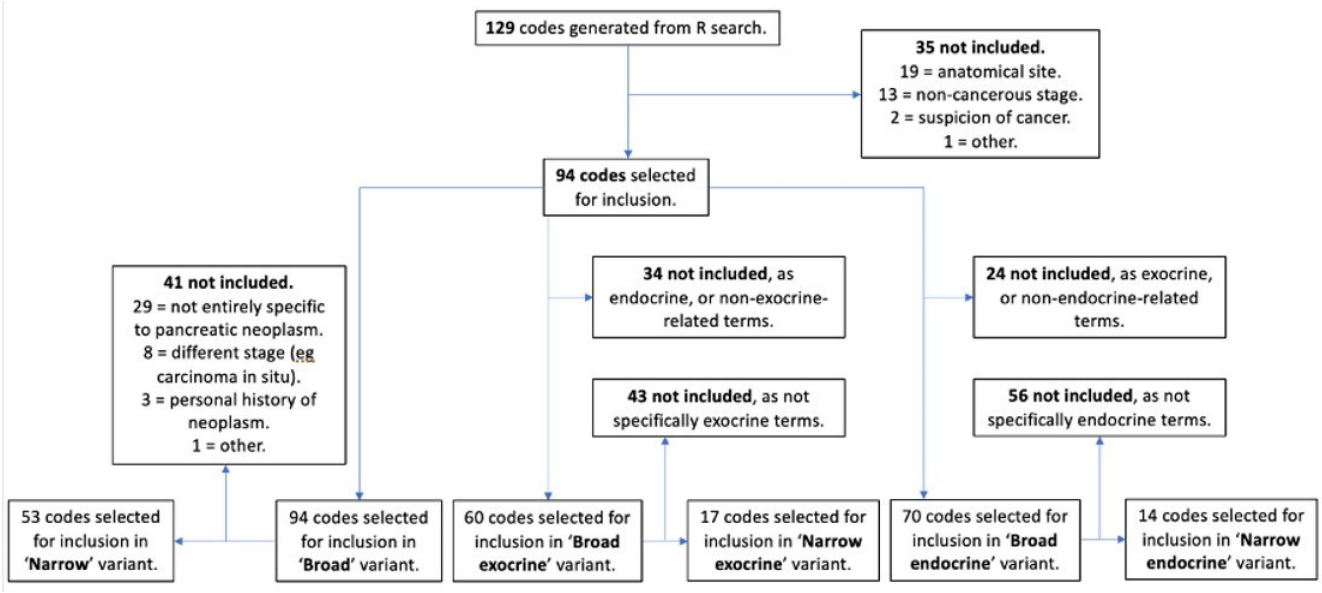

FIGURE 3 | Flow chart to illustrate code selection for pancreatic cancer variants.

3.1.4 | Pancreatic Cancer: Steps 10 and 11

The number of diagnostic codes and subjects identified for each of the variant phenotypes agreed in Step 3 are shown in Tables 3 and S2. The number of codes present in each database was fewer than the 'selected' codes identified through our codelist and phenotype generation.

None of the additional codes suggested by the diagnostic package (22 for narrow cohort, 27 for broad) were deemed suitable for inclusion.

Initially there were 5920 subjects, all related to one included code, in the 'narrow exocrine' variant for PharMetrics Plus for Academics. Upon review, it was clear that the code 'History of malignant neoplasm of pancreas', was included in error in this variant. This was subsequently removed, and Steps 8–11 repeated, resulting in zero subjects in this variant. As expected, 'narrow' cohorts contained fewer patients than corresponding 'broad' cohorts.

Individual estimates of incidence (Figure 4) and prevalence (Figure 5) differed between the six databases. In 2023, for the >65 year old age group, the highest estimates of incidence were seen in IMASIS (187.9 per 100 000 person-years) and CDW Bordeaux (169.7 per 100 000 person-years), with the lowest estimates in CPRD GOLD (60.7 per 100 000 person-years) and IPCI (76.2 per 100 000 person-years). The highest most-recent

TABLE 3 | The number of codes and subjects in each 'variant' of the pancreatic cancer phenotype, by database.

Variant	Number of selected codes/129	Number of codes present in each database (number of subjects included)					
		CPRD	Pharmetrics	IMASIS	IPCI	IQVIA Germany	CDW Bordeaux
Broad	94 (72.9%)	11 (14095)	8 (21445)	7 (1750)	1 (3472)	7 (26177)	24 (5184)
Narrow	53 (41.1%)	7 (13002)	7 (20057)	7 (1750)	1 (3472)	7 (26177)	23 (5180)
Broad exocrine	60 (46.5%)	6 (13818)	7 (21025)	6 (1734)	1 (3472)	6 (26012)	19 (4875)
Narrow exocrine	17 (13.2%)	0 (0)	0 (0)	0 (0)	0	0	6 (2558)
Broad endocrine	70 (54.3%)	10 (13920)	7 (21203)	6 (1732)	1 (3472)	6 (26012)	17 (4413)
Narrow endocrine	14 (10.9%)	2 (70)	1 (1293)	1 (30)	0	1 (192)	5 (525)

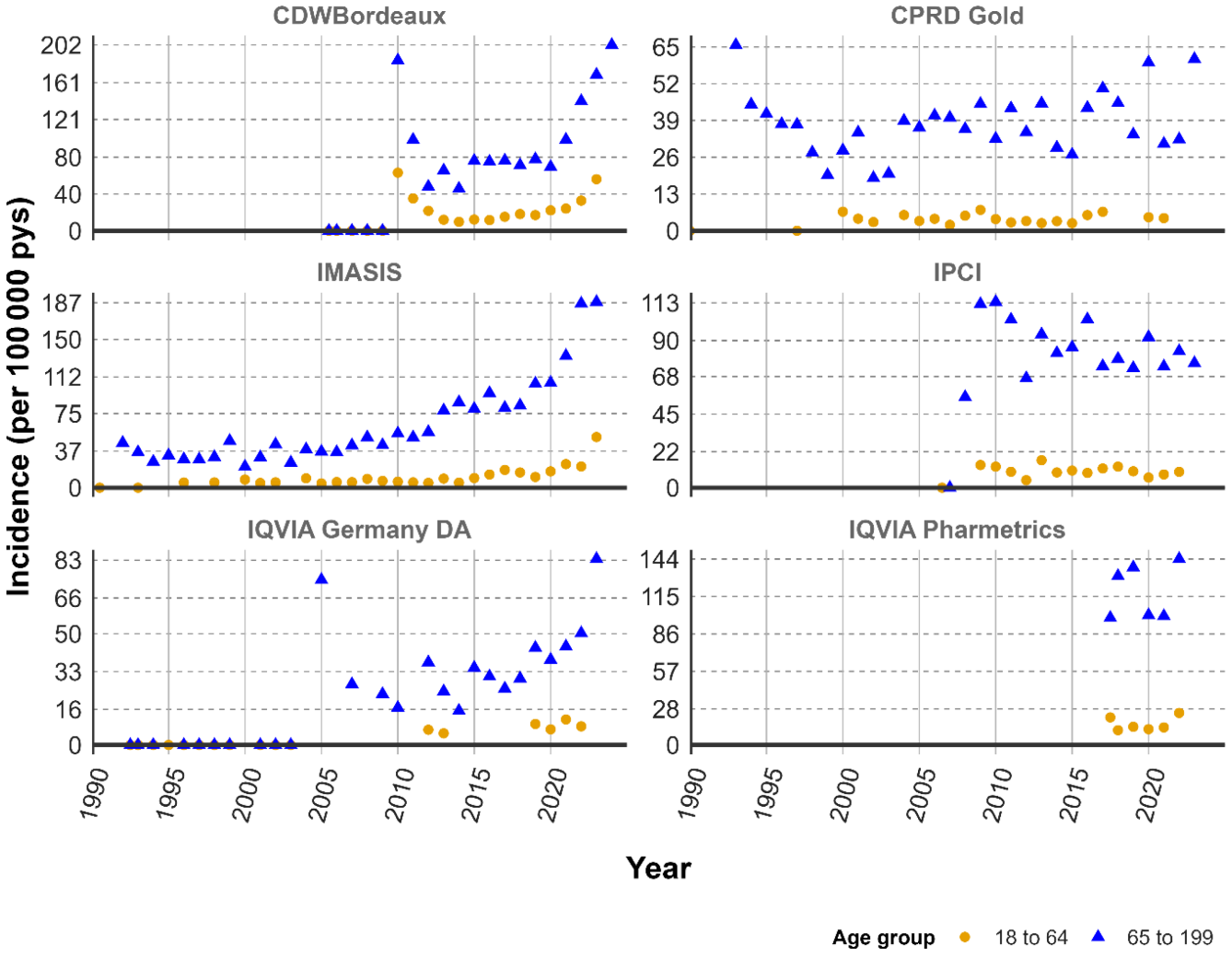

FIGURE 4 | Incidence per 100 000 person-years of pancreatic cancer (narrow variant) among ≥ 65 (blue) and 18–64 (orange) year-olds, grouped by database.

prevalence was seen in CDW Bordeaux (0.75%) and IMASIS (0.42%), and lowest in CPRD GOLD (0.057%) and IQVIA Germany (0.16%). Both incidence and prevalence were consistently higher in the older (> 65 years) age group than the adult (18–64) age group. For comparison, data from the Global Cancer Observatory (GLOBOCAN) reports age-standardised rates of pancreatic cancer in 2022 of 6.7 per 100 000 in the United Kingdom, 7.5 in Spain, 8.1 in the Netherlands, 8.6 in the USA, 8.7 in Germany and 9.4 in France [18].

Age and sex distribution at time of diagnosis of pancreatic cancer was broadly comparable between databases, with most cases

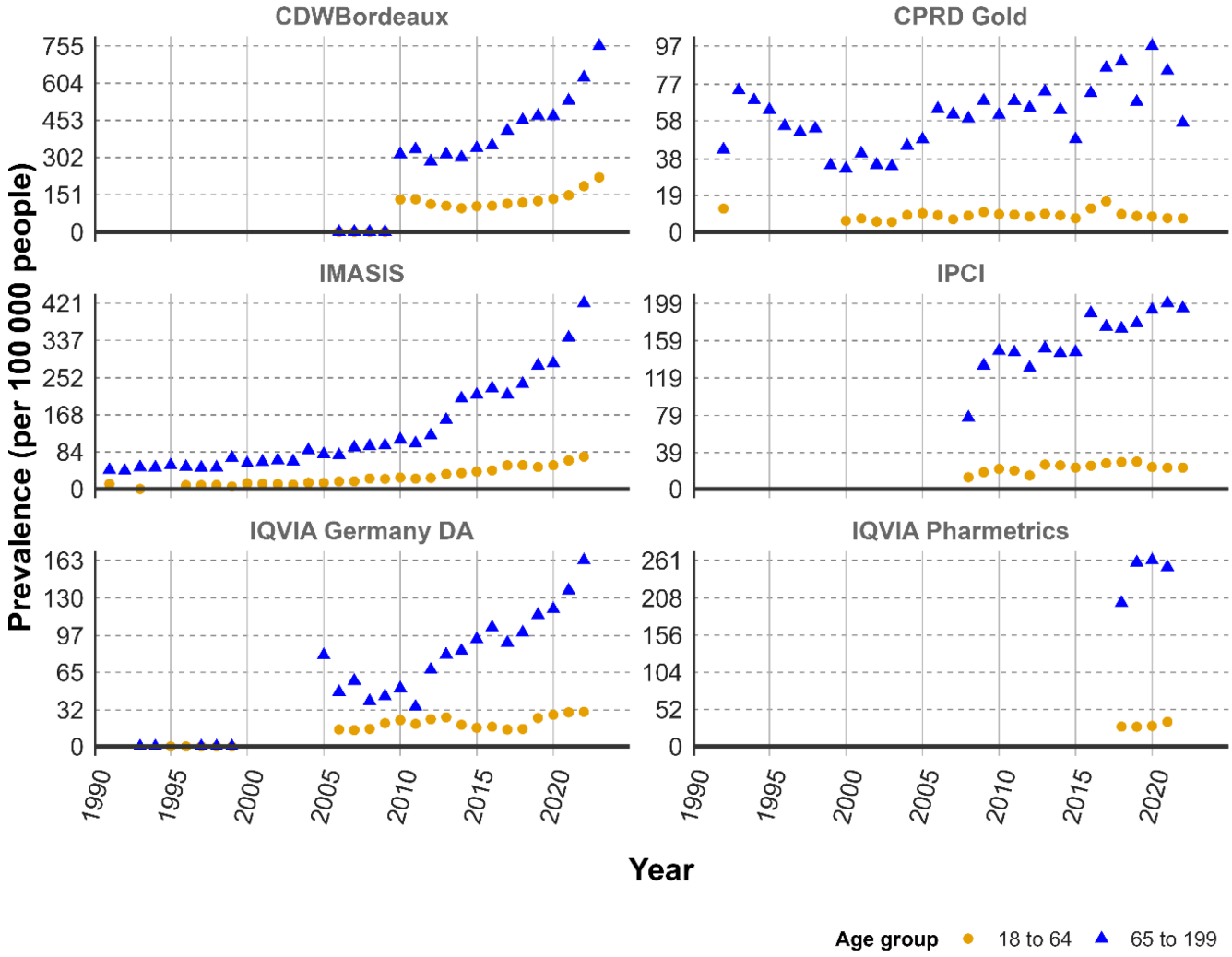

FIGURE 5 | Prevalence (per 100 000 people) of pancreatic cancer (narrow variant) among ≥65 (blue) and 18–64 (orange) year-olds, grouped by database.

occurring in people aged between 70 and 80 years old (Figure 6). This is in line with data from Cancer Research UK [19] and the US National Cancer Institute [20].

Comparison of the co-occurrence of selected conditions to an age-matched cohort showed large percentage differences in expected phenotypic conditions and observations including jaundice and upper abdominal pain, which are hallmark symptoms of the condition (Table 4).

Similarly, medications with extremely high percentage differences are strongly in line with what would be expected for a cohort of people with pancreatic cancer, including opioid pain relief and antiemetics (Table 5).

3.1.5 | Pancreatic Cancer: Steps 12–14

The cohort was versioned and finalised. The cohorts were discussed among the phenotyping team and the final phenotype is now available for future use.

3.2 | SLE Phenotype

3.2.1 | SLE: Steps 1–3

The phenotype generation was deemed feasible. A clinical description was generated by the PAs (Figure S5), and variants were determined after initial exploration of the topic area between the PA and PL (Table 6).

3.2.2 | SLE: Steps 4 and 5

No suitable cohort or phenotype was found to exist already in DECK.

3.2.3 | SLE: Steps 6–9

A provisional list of concept IDs was generated using CodeListGenerator (Figure S6). The PA performed the initial concept set review. This was subsequently reviewed by three further

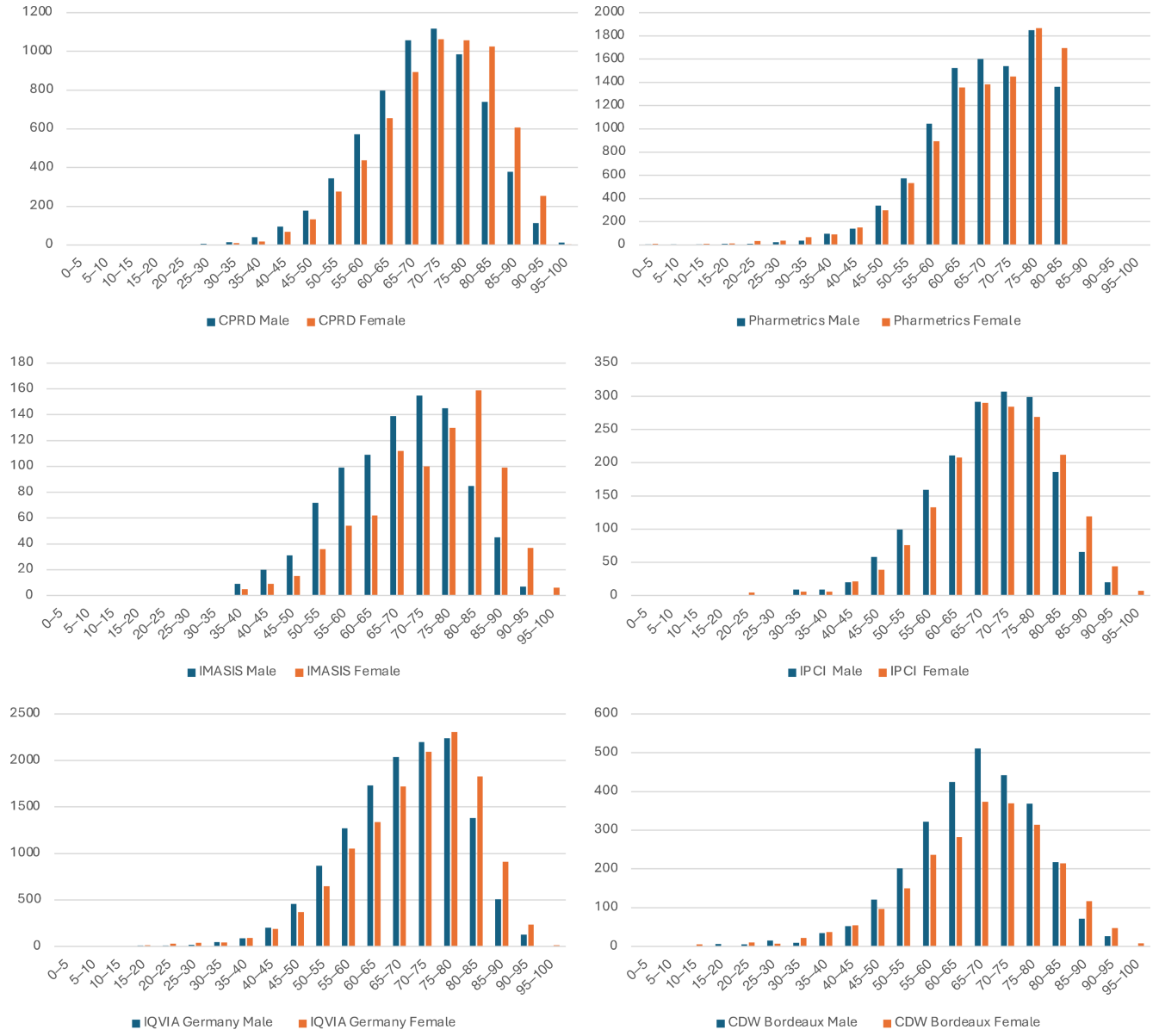

FIGURE 6 | Age distribution of pancreatic cancer (narrow variant phenotype) across the six databases.

TABLE 4 | Percentage difference in selected condition occurrence, for 'narrow' pancreatic cancer variant, before or after diagnosis, with an age and sex matched cohort, in the six databases.

Concept name	Before (−31 to −365 days) or after (+31 to +365 days)	Difference in proportions (%)					
		CPRD GOLD	Pharmetrics	IMASIS	IPCI	IQVIA Germany	CDW Bordeaux
Epigastric pain	Before	+ 2145	+ 590	+ 349			
Jaundice	Before		+ 2591		+ 922		+ 364
Upper/abdominal pain	Before	+ 986	+ 1837	+ 157	+ 270	+ 532	+ 867
Type 2 diabetes mellitus	After	+ 50	+ 79	+ 454	+ 55	+ 276	+ 197
Jaundice	After	+ 2173	+ 5211				+ 3645
Pulmonary embolism	After	+ 1226	+ 1878		+ 522	+ 123	
Death	After	+ 2627			+ 1281		
Terminal illness/terminal care/hospice care	After	+ 4951	+ 837				

TABLE 5 | Percentage difference in selected drug prescription, for 'narrow' pancreatic cancer variant, days (+31 to +365) versus age matched population, in the six databases.

		Relative difference in proportions (%)				
Concept name	**CPRD GOLD**	**Pharmetrics**	**IMASIS**	**IPCI**	**IQVIA Germany**	**CDW Bordeaux**
Fentanyl	+ 1689	+ 407	+ 567	+ 2151	+ Inf	
Morphine	+ 3616	+ 564	+ 1652	+ 2196	+ Inf	+ 562
Ondansetron	+ 2777	+ 531	+ 1141	+ 13 517	+ Inf	+ 2264
Picosulfurate	+ 2652			+ 175	+ Inf	
Methotrimeprazine	+ 7729		+ 626			
Haloperidol	+ 1304	+ 421	+ 656	+ 2591	+ Inf	+ 386
Metoclopramide	+ 3843	+ 1488	+ 2436	+ 1767	+ Inf	+ 2499
C Methotrimeprazine resin	+ 2426		+ 692			
(Butyl)-Scopolamine	+ 1429		+ 3639	+ 257	+ Inf	+ 727

Note: + Inf—if only found counts on the pancreatic cancer cohort and < 5 in the matched sample.

TABLE 6 | SLE phenotype variants and their rationale.

Variant name	**Rationale**
Broad	Aims to include all current and likely SLE records (higher sensitivity, lower specificity)
Narrow	Aims to include highly likely SLE records (higher specificity)
Prevalent	To include records indicating a past history of SLE in addition to the above

TABLE 7 | The number of codes, and subjects, in each 'variant' of the SLE phenotype, across the five databases.

		Number of codes present in each database (number of subjects included)				
Variant	**Number of selected codes/92**	**CPRD**	**Pharmetrics**	**IMASIS**	**IQVIA Germany**	**CDW Bordeaux**
Broad (adult)	49 (53.2%)	10 (9066)	8 (67642)	5 (1041)	3 (15030)	3 (1556)
Narrow (adult)	41 (44.6%)	8 (7381)	7 (65310)	4 (826)	2 (12785)	2 (1478)
Prevalent (adult)	42 (45.7%)	8 (7381)	7 (65310)	4 (826)	2 (12785)	2 (1478)
Broad (juvenile)	49 (53.2%)	10 (255)	8 (876)	5 (11)	3 (294)	3 (59)
Narrow (juvenile)	41 (44.6%)	8 (206)	7 (823)	4 (10)	2 (234)	2 (59)
Prevalent (juvenile)	42 (45.7%)	8 (206)	7 (823)	4 (10)	2 (234)	2 (59)

individuals including the PL and PI. A total of 92 codes were initially included of which 50 were selected for inclusion (Table S3). Figure 7 illustrates the rationale for code selection into each variant.

It was decided to include the code 'neonatal lupus erythematosus' in the broad group but not the narrow or prevalent group given it is not a primary disease but driven by maternal antibody transfer.

3.2.4 | SLE: Steps 10 and 11

Phenotype diagnostics were run in and reviewed by the PA, PL, PI and other PAs on the team.

The numbers of codes present in each database was fewer than the original 'selected' codes (Tables 7 and S4). We were not able to assess SLE in IPCI.

None of the additional codes suggested by the diagnostics package were deemed suitable for inclusion, and the cohort overlap between variants was as expected.

As expected, 'narrow' cohorts contained fewer patients than the 'broad' cohorts, driven by the inclusion of 'lupus erythematosus' (not necessarily systemic), and 'neonatal lupus erythematosus' in the broad cohort. There were no differences in numbers of subjects included in the narrow and prevalent cohorts because

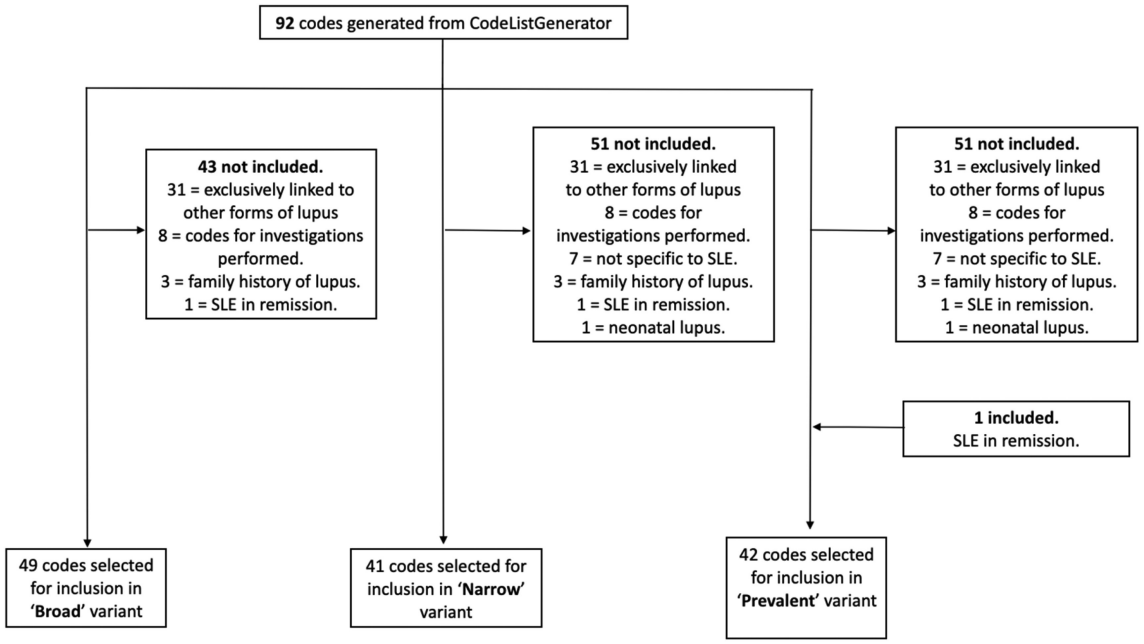

FIGURE 7 | Flow chart to illustrate code selection for SLE variants.

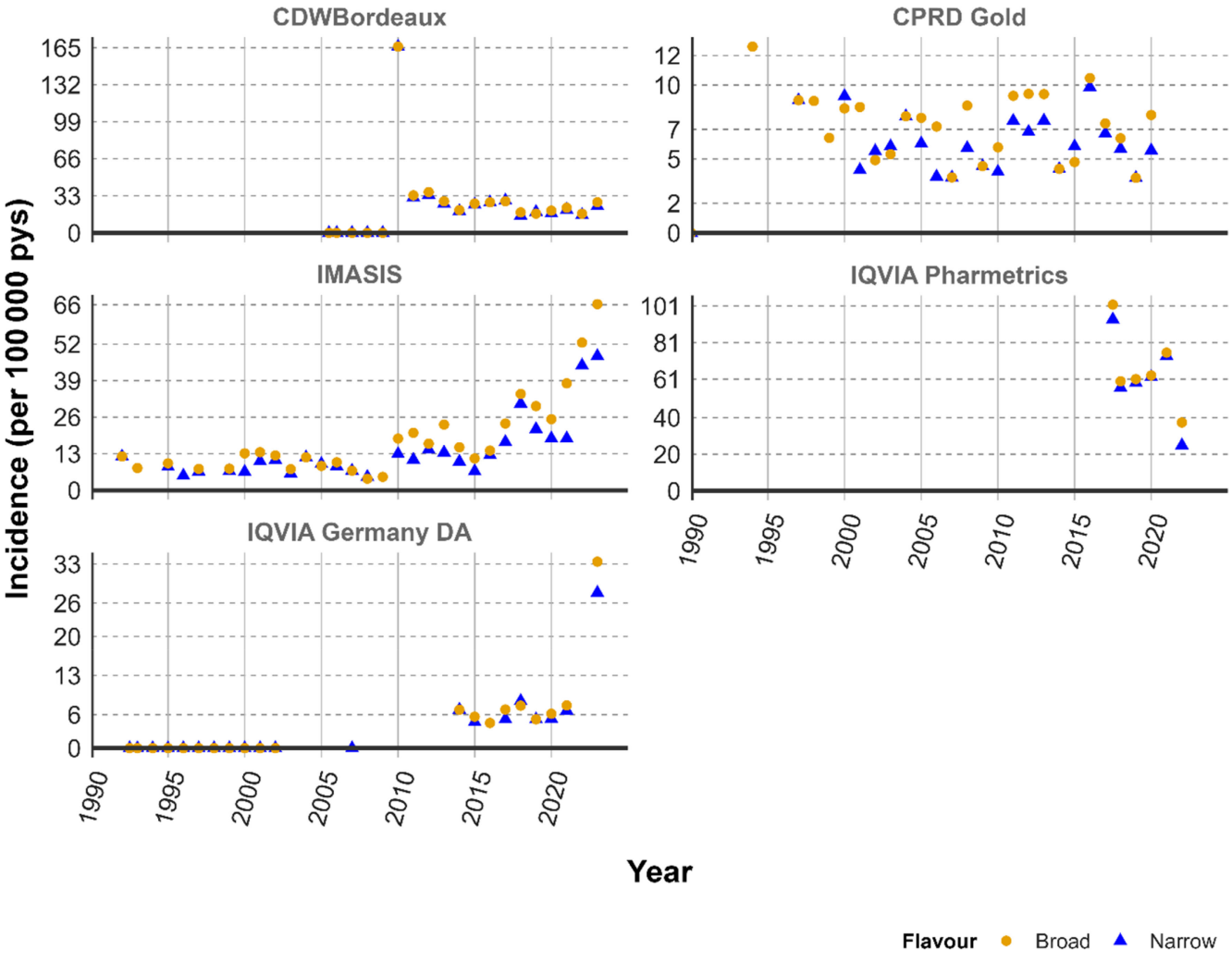

FIGURE 8 | Incidence per 100 000 person-years of broad (orange) and blue (narrow) adult SLE phenotypes in the six databases. Results for the 18–64 years age group is shown for illustrative purposes.

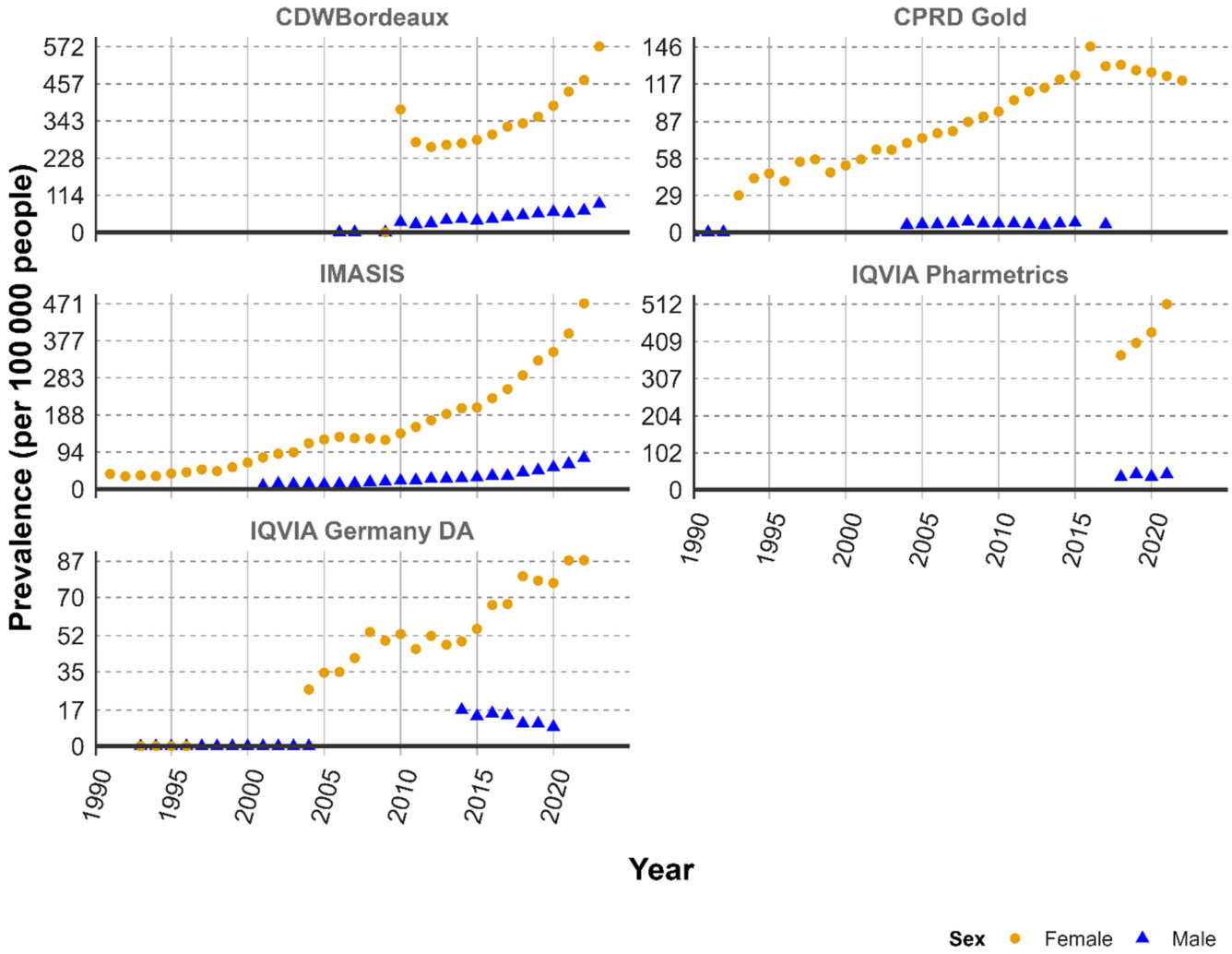

FIGURE 9 | Prevalence of 'narrow' adult SLE phenotype among female (orange) and male (blue) patients in the six databases. Results for the 18–64 years age group shown for illustrative purposes.

the extra 'systemic lupus erythematosus in remission' code present in the prevalent concept set yielded no subjects.

In the adult SLE populations, the broad and narrow phenotypes had broadly comparable incidence rates (Figure 8) within each database, although there were considerable differences in incidence estimates between databases. Higher estimates of incidence were seen generally in CDW Bordeaux and Pharmetrics, with lower estimates in CPRD GOLD, IQVIA Germany and IMASIS. A recent meta-analysis and Bayesian hierarchal mixed model estimated relatively similar incidence rates of SLE among the European countries in our study (4.04 per 100 000 person-years in the UK, 4.85 in France, 5.95 in the Netherlands, 3.96 in Germany and 4.07 in Spain), with higher estimates for the USA (12.13 per 100 000 person-years) [21]. The prevalence of SLE was markedly higher among females compared to males in all databases (Figure 9). Juvenile incidence are prevalence rates are omitted as they occurred in less than 1 per million.

Adult SLE was most commonly diagnosed in the middle-age across all databases (Figure 10), and a clear difference in number of cases by sex (broadly 9:1 female:male). This is in line with previous studies in CPRD [22], and established global trends [23].

Prior to diagnosis, subjects frequently had joint pain, a key presenting feature of SLE (Table 8). They also had high rates of measurements related to suspected SLE (immunology laboratory testing) and muscular pain (creatine kinase). Other musculoskeletal or autoimmune conditions were seen in most databases (hypothyroidism, carpal tunnel syndrome and Raynaud's disease), alongside specialist clinic reviews, likely due to the multisystem nature of SLE and diagnostic difficulties.

Additionally, we observed high rates of immunosuppressant drugs, non-steroidal anti-inflammatory medications, and antihistamines (Table 9).

3.2.5 | SLE: Steps 12–14

The cohort was versioned and stored and is now available for future use.

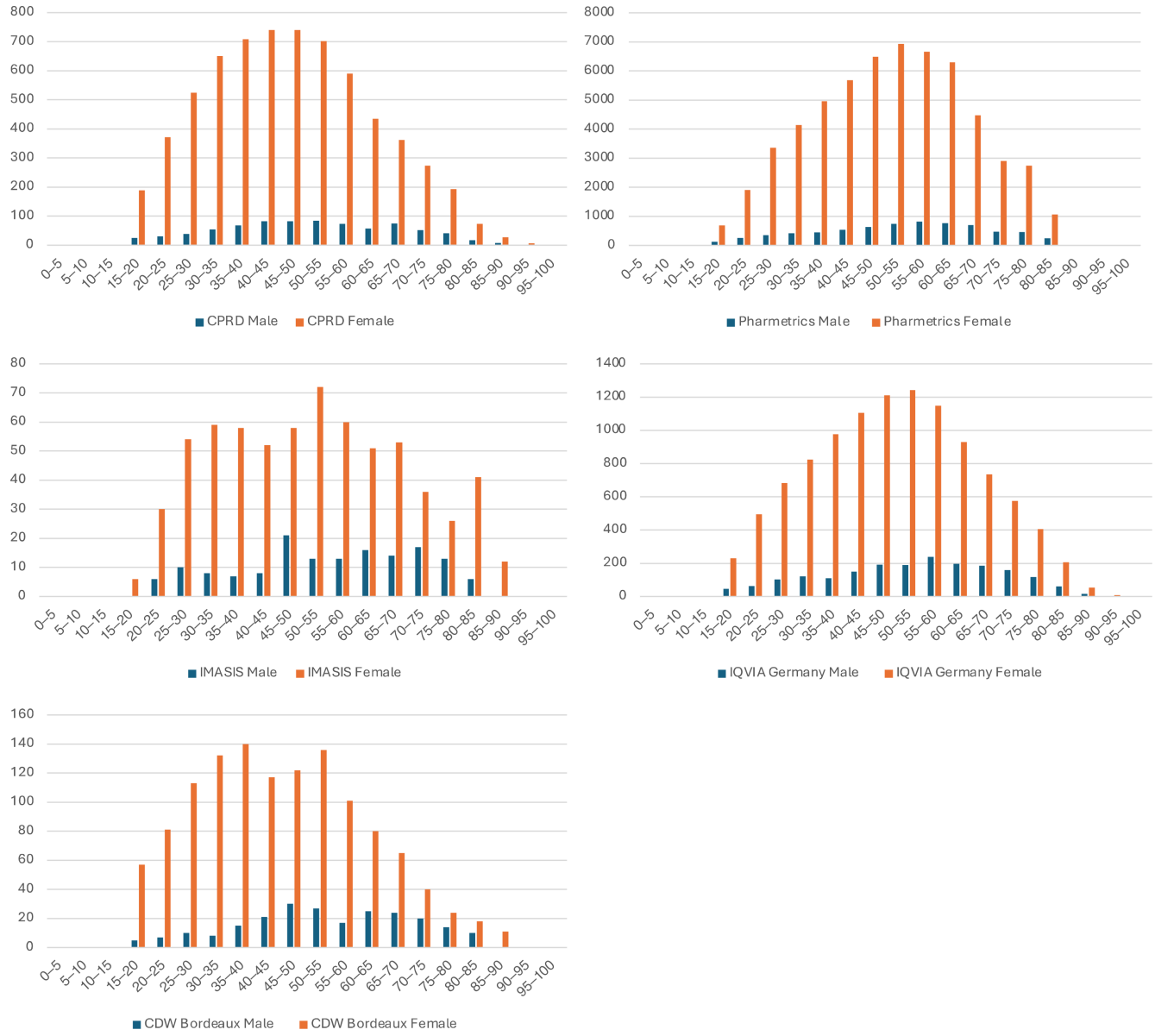

FIGURE 10 | Age distribution of SLE cases (adult narrow phenotype) across the six databases.

4 | Discussion

We describe a standardised, reproducible framework for developing disease phenotypes for use in OMOP-mapped real world data. Additionally, we illustrate its potential by generating phenotypes for complex diseases, namely pancreatic cancer and SLE. We generated multiple 'variants' of these diseases, allowing us to compare the influence of varying diagnostic code inclusion on key parameters such as incidence and prevalence. These phenotypes have been versioned, stored, and are available for future use.

Several techniques have been used for phenotype generation, including rule-based systems, machine learning, and natural language processing methods [24]. More recently, the PHenotype Observed Entity Baseline Endorsements (PHOEBE) tool has been developed to facilitate code list generation within the OHDSI CDM [11] and has been used as a part of the diagnostics to identify possible missed codes.

We compared our codelists with previous reports. One recent study has examined the incidence of SLE in CPRD. Our careful review of codes (Step 8) and phenotype diagnostics (Step 10) led to the exclusion of codes with semantic similarities but without clinical equivalence to SLE, which were however included in this recent study [22]. Examples include drug-induced lupus, 'SLE—Saint Louis Encephalitis' (a different diseases with the same acronym), or neonatal lupus (only included in our broad variant). This led to marginally higher rates in the previous manuscript compared to our estimates.

Our codes for pancreatic cancer were largely similar to those used in prior studies [25–27]. However, while we excluded cancers of the ampulla of Vater and the sphincter of Oddi and logged anatomical distinction as reasons for exclusion, previous studies included one [26, 27] or the other without a clear rationale [25]. This highlights the value of our fully documented and traceable process to maximise reproducibility.

TABLE 8 | Percentage difference in selected condition occurrence, for 'narrow' SLE variant, before or after diagnosis, with an age matched cohort, in CPRD GOLD.

Concept name	Before (−31 to −365 days) or after (+31 to +365 days)	Difference in proportions (%)				
		CPRD GOLD	Pharmetrics	IMASIS	IQVIA Germany	CDW Bordeaux
Joint pain	Before	+ 2014	+ 712	+ 127	+ 230	+ 614
Immunology laboratory test	Before	+ 2416	+ 917			
Double stranded DNA antibody test	Before			+ 4345		+ 3259
Creatine kinase measurement	Before	+ 241	+ 561	+ 533		+ 355
Carpal tunnel syndrome	Before	+ 194	+ 80	+ 127		+ 16
Seen in rheumatology clinic	Before	+ 1545				
Seen in dermatology clinic	Before	+ 797				
Raynaud's disease	Before	+ 971	+ 1252			+ 479
Hypothyroidism	After	+ 83	+ 48	− 62.3	+ 210	− 32

TABLE 9 | Percentage difference in selected drug prescriptions, for 'narrow' SLE variant after diagnosis (+31 to +365 days) versus age matched population, in CPRD GOLD.

Concept name	CPRD GOLD	Difference in proportions (%)			
		Pharmetrics	IMASIS	IQVIA Germany	CDW Bordeaux
Hydroxychloroquine	+ 11 396	+ 3402	+ 6082	+ Inf	
Methotrexate	+ 1305	+ 1068	+ 876	+ Inf	+ 387
Desloratadine	+ 285				+ 241
Hydroxyzine	+ 436	+ 165	+ 225	+ Inf	+ 192
Celecoxib	+ 262	+ 248			
Azathioprine	+ 2310	+ 1341		+ Inf	
Levothyroxine	+ 220	+ 45	+ 129	+ 80	+ 290
Prednisolone/prednisone	+ 724	+ 148/+ 279	/+ 1057	/+ 182	+ 1556/+ 5874

4.1 | Strengths and Limitations of the Current Study

The proposed 14-step process outlined above has three main strengths. First, the process of assessing the suitability of existing phenotypes helps minimise duplication of effort, and maximise the use of previous validated algorithms. Second, the DECK portal facilitates versioning, commenting, and tracking of disagreements between reviewers, improving overall transparency. Third, the phenotype diagnostics stage allows the assessment of whether a cohort, derived from a phenotype, has similar characteristics to those expected based on pre-specified clinical descriptions. This step also allows for the inclusion of previously missed relevant codes. Our framework also allows for the generation of different 'variants' of a phenotype to represent different sub-types of disease, to increase sensitivity or specificity or to differentiate exposures or outcomes, which is an important consideration depending on the intended use of a phenotype. Our process is flexible, and although we used single code lists based phenotypes to construct and compare cohorts in six separate databases from six countries, the process can accommodate more complex phenotypes that combine different concept sets and logics.

Limitations of the framework include the increased human resources required to complete a high-quality phenotype. Second, while the framework aims to improve the ability to create a reliable phenotype, it cannot eliminate the risk that subjects have been mis-coded at source. In line with this, a key limitation of the suggested process is the inability to identify subjects with the disease but with missing diagnoses, that is, incomplete data. This misclassification is a key limitation of most real world data sources, and alternative mitigation strategies like quantitative bias analyses should be explored where suspected.

Our proposed framework does not routinely involve medical record review to directly ascertain that included subjects do indeed have the disease of interest (separately to the diagnostic codes they may have been assigned) but relies instead on assessment by investigators of other population-based characteristics coded in their medical records, including other medical conditions and drug prescriptions. This kind of review is usually not feasible in our environment due to the number of cases and data partners involved, and due to local data privacy legislation prohibiting this practise. Novel methods including NLP or large language models are promising and should be evaluated in the future [29].

4.2 | Implications for Policy and Practise

Structured phenotyping methods have the potential to increase the reliability of the findings of observational studies, which are frequently integrated into guidelines used in clinical practise [30].

High-quality phenotypes using standard vocabularies allow these to be used in studies across multiple databases, if properly evaluated in each of the databases, a benefit which has already been employed in studies conducted by DARWIN EU. Moving forward, further co-creation and optimisation of phenotypes with regulatory partners such as the EMA will allow these to be implemented in studies investigating a wide range of population level outcomes.

Some challenges remain. Our case examples show that some aspects of our generated phenotypes, such as age, sex, and comorbidities, were relatively consistent across the six diverse databases we used, but that other estimates, specifically for incidence and prevalence, did differ considerably. This is usually due to differences in the underlying data source. For example, routine primary care data is usually better suited for this, where we capture in the denominator all population in the catchment area of a GP practise, compared to hospital or claims data where we may capture only those that have had any contact with the health service, being potentially sicker. Differences in rates, especially in secondary care, may not be due to real differences in incidence and prevalence of SLE and pancreatic cancer in the local populations. Differences in incidence rates of several adverse events and medical conditions based on the data source have been described previously in European data [31]. This highlights the need to record these differences for future studies as part of the phenotyping process, involve data partners to leverage existing data knowledge, have readily accessible metadata on the dataset used that describes all dimensions of the data, [28, 32] run phenotype diagnostics if they are to be used in new datasets, and search the wider literature for comparison of established disease trends in each population of interest.

5 | Conclusions

In conclusion, we report and illustrate a phenotyping framework that provides a reproducible and transparent process for generating reliable disease phenotypes for use in the context of RWE generation for regulatory decision-making in Europe. Wider use of this and other phenotyping methods will improve the value and reliability of findings from observational studies.

5.1 | Plain Language Summary

When people visit their GP or local hospital, the reason for their visit, including their symptoms (such as cough), diagnosis (lung infection) and medications (antibiotics) are recorded, or 'coded' into the electronic health records system. The exact format of this varies between countries, but these databases are often used to answer questions we might be interested in, such as how common diseases are or what medications are used to treat them. However, coding systems are not always perfect, and there may be many different ways of coding the same disease. Deciding on what a disease 'looks like' in different data, for example, from primary care or hospital records involves picking and choosing all the different codes which may represent the disease you are interested in. This process is called 'phenotyping' and is not always done consistently. In this paper, we lay out a framework to perform this phenotyping process, and we show how we have used it to create reliable groups of patients with two complex diseases: cancer of the pancreas, an organ involved in digestion and regulation of blood sugar levels, and systemic lupus erythematosus, a disease where the body's own immune system attacks its tissues and organs.

Acknowledgements

This document expresses the opinion of the authors of the paper and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties. We acknowledge contributions from Prof Asieh Golozar (NorthEastern University, Odysseus), who played a role in the conceptualisation of our phenotyping process.

Disclosure

The work presented in this manuscript has not been posted or presented elsewhere at the time of submission.

Ethics Statement

Permissions were obtained to use CPRD GOLD for these analyses (protocol numbers 23_003556 and 23_003326). The use of PharMetrics Plus for Academics and IQVIA Germany DA are exempt from specific ethic approval submissions. Approval obtained to use IMASIS for these analyses was granted with protocol numbers CEIm 2023/11262 and 2024/11513.

Conflicts of Interest

Daniel Prieto-Alhambra's department has received grant/s from Amgen, Chiesi-Taylor, Lilly, Janssen, Novartis and UCB Biopharma. His research group has received consultancy fees from Astra Zeneca and UCB Biopharma. Amgen, Astellas, Janssen, Synapse Management Partners and UCB Biopharma have funded or supported training programmes organised by Daniel Prieto-Alhambra's department.

References

1. G. Hripcsak and D. J. Albers, "High-Fidelity Phenotyping: Richness and Freedom From Bias," *Journal of the American Medical Informatics Association* 25, no. 3 (2017): 289–294.

2. G. Hripcsak, J. D. Duke, N. H. Shah, et al., "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational

Researchers," *Studies in Health Technology and Informatics* 216 (2015): 574–578.

3. J. M. Overhage, P. B. Ryan, C. G. Reich, A. G. Hartzema, and P. E. Stang, "Validation of a Common Data Model for Active Safety Surveillance Research," *Journal of the American Medical Informatics Association* 19, no. 1 (2012): 54–60.

4. R. A. DeFronzo, E. Ferrannini, L. Groop, et al., "Type 2 Diabetes Mellitus," *Nature Reviews Disease Primers* 1, no. 1 (2015): 15019.

5. S. Lanes, J. S. Brown, K. Haynes, M. F. Pollack, and A. M. Walker, "Identifying Health Outcomes in Healthcare Databases," *Pharmacoepidemiology and Drug Safety* 24, no. 10 (2015): 1009–1016.

6. DARWIN EU, "Data Analysis and Real World Interrogation Network," https://www.darwin-eu.org/index.php.

7. https://cprd.com/primary-care-data-public-health-research.

8. https://www.iqvia.com/locations/united-states/library/fact-sheets/iqvia-pharmetrics-plus.

9. M. A. J. de Ridder, M. de Wilde, C. de Ben, et al., "Data Resource Profile: The Integrated Primary Care Information (IPCI) Database, The Netherlands," *International Journal of Epidemiology* 51, no. 6 (2022): e314–e323.

10. G. Rao, "PhenotypeLibrary: The OHDSI Phenotype Library," 2024.

11. A. Ostropolets, P. Ryan, and G. Hripcsak, "Phenotyping in Distributed Data Networks: Selecting the Right Codes for the Right Patients," *American Medical Informatics Association Annual Symposium Proceedings* 2022 (2022): 826–835.

12. OHDSI, "Software Tools," https://www.ohdsi.org/software-tools/.

13. OHDSI, "Capr package," https://github.com/OHDSI/Capr.

14. https://ohdsi.github.io/CohortDiagnostics/.

15. https://cran.rstudio.com/web/packages/CodelistGenerator/index.html.

16. https://cran.r-project.org/web/packages/PatientProfiles/index.html.

17. https://cran.r-project.org/web/packages/IncidencePrevalence/index.html.

18. https://gco.iarc.fr/today/en/dataviz/maps-heatmap?mode=population&cancers=13.

19. Cancer Research UK, accessed November, 2023, https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/pancreatic-cancer/incidence#heading-One.

20. https://seer.cancer.gov/statfacts/html/pancreas.html.

21. J. Tian, D. Zhang, X. Yao, Y. Huang, and Q. Lu, "Global Epidemiology of Systemic Lupus Erythematosus: A Comprehensive Systematic Analysis and Modelling Study," *Annals of the Rheumatic Diseases* 82, no. 3 (2023): 351–356.

22. N. Conrad, S. Misra, J. Y. Verbakel, et al., "Incidence, Prevalence, and Co-Occurrence of Autoimmune Disorders Over Time and by Age, Sex, and Socioeconomic Status: A Population-Based Cohort Study of 22 Million Individuals in the UK," *Lancet* 401, no. 10391 (2023): 1878–1890.

23. F. Rees, M. Doherty, M. J. Grainge, P. Lanyon, and W. Zhang, "The Worldwide Incidence and Prevalence of Systemic Lupus Erythematosus: A Systematic Review of Epidemiological Studies," *Rheumatology* 56, no. 11 (2017): 1945–1961.

24. C. Shivade, P. Raghavan, E. Fosler-Lussier, et al., "A Review of Approaches to Identifying Patient Phenotype Cohorts Using Electronic Health Records," *Journal of the American Medical Informatics Association* 21, no. 2 (2014): 221–230.

25. S. J. Price, S. A. Stapley, E. Shephard, K. Barraclough, and W. T. Hamilton, "Is Omission of Free Text Records a Possible Source of Data Loss and Bias in Clinical Practice Research Datalink Studies? A Case–Control Study," *BMJ Open* 6, no. 5 (2016): e011664.

26. N. U. Din, O. C. Ukoumunne, G. Rubin, et al., "Age and Gender Variations in Cancer Diagnostic Intervals in 15 Cancers: Analysis of Data From the UK Clinical Practice Research Datalink," *PLoS One* 10, no. 5 (2015): e0127717.

27. R. D. Neal, N. U. Din, W. Hamilton, et al., "Comparison of Cancer Diagnostic Intervals Before and After Implementation of NICE Guidelines: Analysis of Data From the UK General Practice Research Database," *British Journal of Cancer* 110, no. 3 (2014): 584–592.

28. M. Swertz, E. van Enckevort, J. L. Oliveira, et al., "Towards an Interoperable Ecosystem of Research Cohort and Real-World Data Catalogues Enabling Multi-Center Studies," *Yearbook of Medical Informatics* 31, no. 1 (2022): 262–272.

29. K. Yoshida, T. Cai, L. G. Bessette, et al., "Improving the Accuracy of Automated Gout Flare Ascertainment Using Natural Language Processing of Electronic Health Records and Linked Medicare Claims Data," *Pharmacoepidemiology and Drug Safety* 33, no. 1 (2024): e5684.

30. N. A. Dreyer, S. R. Tunis, M. Berger, D. Ollendorf, P. Mattox, and R. Gliklich, "Why Observational Studies Should Be Among the Tools Used in Comparative Effectiveness Research," *Health Affairs* 29, no. 10 (2010): 1818–1825.

31. C. Willame, C. Dodd, C. E. Durán, et al., "Background Rates of 41 Adverse Events of Special Interest for COVID-19 Vaccines in 10 European Healthcare Databases – An ACCESS Cohort Study," *Vaccine* 41, no. 1 (2023): 251–262.

32. R. Gini, R. Pajouheshnia, H. Gardarsdottir, et al., "Describing Diversity of Real World Data Sources in Pharmacoepidemiologic Studies: The DIVERSE Scoping Review," *Pharmacoepidemiology and Drug Safety* 33, no. 5 (2024): e5787.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.