METHOD ARTICLE

REVISED Checklist and guidance on creating codelists for routinely collected health data research [version 2; peer review: 3 approved]

Previous Title 'Checklist and guidance on creating codelists for electronic health records research'

Julian Matthewman[1*], Kirsty Andresen[1*], Anne Suffel[1*], Liang-Yu Lin[1], Anna Schultze[1], John Tazare[1], Krishnan Bhaskaran[1], Elizabeth Williamson[1], Ruth Costello[1], Jennifer Quint[2], Helen Strongman[1]

[1]London School of Hygiene & Tropical Medicine, London, England, WC1E 7HT, UK
[2]Imperial College London, London, England, UK

[*] Equal contributors

v2 **First published:** 17 Apr 2024, **4**:20
https://doi.org/10.3310/nihropenres.13550.1
Latest published: 18 Sep 2024, **4**:20
https://doi.org/10.3310/nihropenres.13550.2

Abstract

Background

Codelists are required to extract meaningful information on characteristics and events from routinely collected health data such as electronic health records. Research using routinely collected health data relies on codelists to define study populations and variables, thus, trustworthy codelists are important. Here, we provide a checklist, in the style of commonly used reporting guidelines, to help researchers adhere to best practice in codelist development and sharing.

Methods

Based on a literature search and a workshop with researchers experienced in the use of routinely collected health data, we created a set of recommendations that are 1. broadly applicable to different datasets, research questions, and methods of codelist creation; 2. easy to follow, implement and document by an individual researcher, and 3. fit within a step-by-step process. We then formatted these recommendations into a checklist.

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 2 (revision) 18 Sep 2024	✓ view	✓ view	✓ view
version 1 17 Apr 2024	? view	? view	? view

1. **Shirley Wang**, Howard Hughes Medical Institute - Harvard Medical School, Boston, USA
2. **Elizabeth Ford**, University of Sussex, Brighton, UK
3. **Duncan Edwards**, University of Cambridge, Cambridge, UK

Any reports and responses or comments on the article can be found at the end of the article.

Results

We have created a 10-step checklist, comprising 28 items, with accompanying guidance on each step. The checklist advises on which metadata to provide, how to define a clinical concept, how to identify and evaluate existing codelists, how to create new codelists, and how to review, check, finalise, and publish a created codelist.

Conclusions

Use of the checklist can reassure researchers that best practice was followed during the development of their codelists, increasing trust in research that relies on these codelists and facilitating wider re-use and adaptation by other researchers.

Plain english summary

When a person receives many types of health care, such as a doctor registering a diagnosis or prescribing a drug, information is collected in their computer system. This information is often organised in a structured way, so that each piece of information can be assigned a "code". For example, if a person was diagnosed with type 1 diabetes, this could be recorded with the code E10 from the International classification of diseases, which contains codes on all possible diseases. For type 2 diabetes the code would be E11. To use this information for research, researchers need to define which people they want to study by making a list of all the relevant codes (a "codelist"). For example, to study people with type 1 and 2 diabetes they would need to include E10 and E11 in their codelist. The international classification of diseases coding system includes over 70,000 codes, and other medical dictionaries can include hundreds of thousands of codes. These lists can therefore be long and complex to create. While they are very important in ensuring that research using this data is correct, no step-by-step guidelines exist to help researchers create codelists. To tackle this, we created a checklist and guidance document which researchers can now use to make sure they don't miss important steps and checks while creating their codelists, and to help them share their codelists so they can be re-used by other researchers. We collected recommendations that other authors have made before us, and developed detailed guidance together with experts in using these types of data for research.

Keywords

codelists, clinical codes, codesets, valuesets, electronic health records, checklist, reporting guidance, reproducibility

Corresponding author: Julian Matthewman (julian.matthewman1@lshtm.ac.uk)

Author roles: Matthewman J: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Andresen K**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Suffel A**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Lin LY**: Investigation, Resources, Validation, Writing – Review & Editing; **Schultze A**: Resources, Validation, Writing – Review & Editing; **Tazare J**: Resources, Validation, Writing – Review & Editing; **Bhaskaran K**: Resources, Validation, Writing – Review & Editing; **Williamson E**: Resources, Validation, Writing – Review & Editing; **Costello R**: Resources, Validation, Writing – Review & Editing; **Quint J**: Resources, Validation, Writing – Review & Editing; **Strongman H**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Helen Strongman is funded by the National Institute for Health and Care Research (NIHR) through an Advanced Fellowship (NIHR301730). Anne Suffel is funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Vaccines and Immunisation (NIHR200929), a partnership between the UK Health Security Agency and the London School of Hygiene and Tropical Medicine. Anna Schultze is employed by LSHTM on a fellowship sponsored by GlaxoSmithKline. Krishnan Bhaskaran and Elizabeth Williamson are funded by Wellcome Senior Research Fellowships (220283/Z/20/Z and 224485/Z/21/Z respectively). The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, NHS, UK Health Security Agency, or the UK Department of Health and Social Care.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Matthewman J, Andresen K, Suffel A *et al.* **Checklist and guidance on creating codelists for routinely collected health data research [version 2; peer review: 3 approved]** NIHR Open Research 2024, **4**:20 https://doi.org/10.3310/nihropenres.13550.2

First published: 17 Apr 2024, **4**:20 https://doi.org/10.3310/nihropenres.13550.1

REVISED **Amendments from Version 1**

Compared to the previously published version, the checklist contains one additional step (Checks), which prompts users to check their codelists against the database they were created for (internally) or other data sources (externally). Other changes include changing the title and clarifying that the checklist and guidance are relevant for research in routinely collected health data (not just in electronic health records), and additional discussion on codelist repositories and codelists used in clinical practice.

Any further responses from the reviewers can be found at the end of the article

Background

Routinely collected health data are commonly used for epidemiological research, bringing opportunities to address questions not easily answered with clinical trials or research-specific data collection[1]. Routinely collected health data are commonly structured and coded based on dictionary ontologies or clinical vocabularies. These vary widely in scope and specificity of coding; for example International Classification of Diseases[2] has traditionally been used for administrative purposes such as recording of deaths and hospital activity, whereas Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)[3] was developed for use in clinical practice and includes a more extensive range of codes.

To extract meaningful information on health-related characteristics and events (e.g., diagnoses, prescriptions, referrals, test results, lifestyle factors, etc.) from routinely collected health data, researchers create codelists (also referred to as clinical codelists, code sets, or value sets)[4]. This is done by identifying relevant codes from the dictionary vocabulary (e.g. all the diagnosis, treatment, referral, etc. codes in SNOMED-CT indicating that a person has diabetes). In studies using routinely collected health data, codelists define the study population, and other variables which researchers will use to answer the research question. Therefore, good practice in codelist development is an essential step in ensuring that codelists accurately capture the health-related characteristics or events of interest.

Checklists are increasingly being used in health research to promote adherence to recommended good practice[5], including research using routinely collected health data where the REporting of studies Conducted using Observational Routinely-collected Data (RECORD) statement requires "a complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers"[6]. While a number of articles already provide guidance on creating, sharing and managing codelists, these focus on specific scenarios (e.g. specific coding systems, or using specific codelist creation tools or methods), or pertain to higher level recommendations (e.g. for organisations, funders, or journals, rather than individual researchers)[4,7–11]. Thus, we created an easy to use checklist and step-by-step guidance that can be used by researchers using routinely collected health data to ensure good practice.

Methods

Patient and Public Involvement

The target audience for this methods paper is researchers who use, or are planning to use, electronic health records for research. Researchers at all stages of their academic careers were involved throughout the project, including in developing objectives. We will involve researchers from a wider group of institutions by encouraging them to participate in the open review process. Patients or the public were not involved in this project.

Checklist development

We formed a codelist task group including the following authors of this paper: JM, KA, AS, L-YL, and HS. All task group members were PhD students or academic staff members at LSHTM. The task group completed an initial literature search in PubMed to identify published papers describing methods and guidance for codelists. The most comprehensive review of the methodological literature on codelists was by Williams in 2017; this provides a set of best practice recommendations for future studies and software tools but did not aim to provide guidance for individual researchers on how to implement these recommendations[4]. We updated this review, using the published search strategy, to find new literature released since 2017 (for a description of this literature search process see Box 1: Updated literature search). We also reviewed recommendations in other pertinent publications identified during this process[8–11] and features of different codelist sharing websites and general purpose research repositories[12–15].

Box 1. Updated literature search

We performed a literature search based on, and using the same search strategy as, the existing review by Williams R, *et al.*, 2017[4] to find new literature released since 2017 on the topic. We did not intend to reevaluate recommendations proposed by Williams *et al.*, rather to identify important new literature on codelists that could be used to inform the creation of our checklist and guidance. We title-and-abstract-screened 427 papers published between June 2017 and December 2022 and indexed in PubMed, of which we full-text-screened 24. From these we excluded papers specifically discussing the transition in the US from ICD9 to ICD10, papers with a higher-level focus on terminologies such as mappings between them but no focus on codelists, and applied papers, including papers that use codelists but do not discuss construction, reuse, validation, or sharing of codelists (as was done in Williams R, *et al.*, 2017). There remained 9 papers from which we considered recommendations on codelist management. From these papers, we found 2 areas where additional recommendations we considered for inclusion in our checklist and guidance. The two identified topics are as follows:

1. When SNOMED CT is the available terminology, it may be preferrable to avoid "flat" codelists (i.e., a list of all codes to define a concept), in favour of using SNOMED CT concept hierarchies (i.e., a primary concept and its descendants optionally with additional relationships). These concept hierarchies may define more complex concepts (e.g. (Cerebrovascular accident OR History of Cerebrovascular accident) AND NOT Ruptured aneurysm)[16–18]. For drugs, it may be possible to use other terminologies such as MeSH, ATC, etc. to create similar concept hierarchies rather than creating "flat" codelists[19]. While a recommendation to make use of concept

hierarchies was already included in the Williams *et al.* 2017 review which was adapted for our checklist and guidance, we decided not to include guidance specific to the SNOMED-CT terminology, as this did not adhere to our criteria of being broadly applicable to different datasets, research questions, and methods of codelist creation.

2. If available, measures to check the quality of code sets should be made use of. The use of inter-terminology maps is recommended to check for codelists completeness when codelists exist in multiple terminologies (e.g. when creating a codelist in SNOMED CT, map an existing ICD-10 codelist to SNOMED and check for overlap and differences)[20]. However, caution is needed when mapping terms from different ontologies to each other as they may have been created for different purposes (e.g., documentation, billing, registries, referrals or information sharing) and are often used in different care settings (e.g., SNOMED CT in primary care in the UK and ICD-10 codes in secondary care). Some authors propose data centric natural-language processing methods to semi-automatically check codelists, however this will be dependent on the availability of such systems[21]. Within excluded papers, we found multiple recommendations for use of common data models which may address problems with codelists on a higher level, which we did not focus in this work. We mention the use of inter-terminology maps in the guidance section on searching for existing codelists.

Based on these publications and our expertise in using routinely collected health data, the task group drafted an initial checklist, encompassing a set of recommendations on codelist development and sharing that needed to fit the following criteria: 1. broadly applicable to different datasets, research questions, and methods of codelist creation; 2. easy to follow, implement and document by an individual researcher; 3. fit within a step-by-step process where some items should be completed before others. This draft checklist was presented to, and pilot tested on example codelists in a workshop with a wider group of researchers in the Electronic Health records research group at the London School of Hygiene and Tropical Medicine (EHR research group). From this we gathered feedback which was used to further refine recommendations (for a description of this process, see Box 2: Feedback from workshop). Finally, we circulated the checklist to be reviewed and approved by the EHR research group at LSHTM and other stakeholders.

Box 2. Feedback from workshop

The task group convened a small group workshop to understand current codelist reporting practices and improve the process of creation, management, storage and sharing of codelists. All academic staff and PhD student members of the LSHTM Electronic Health Records research group were invited to attend. The workshop was held at the workplace for approximately 3 hours and was facilitated by the task group. Each of 4 groups with 3 to 4 people was provided with an example codelist (that had been employed in previous research), a draft version of the codelist guidance document based on a review of existing literature, and a questionnaire. Each group used the questionnaire to assess the codelist against the provided draft guidelines. Attendees were then asked to provide input to the draft guidelines in a plenary session. The plenary session was structured in two main discussion topics: existing codelists and new codelists. The discussion centred on key themes contained within these discussion topics. The task group took notes during discussions and collated notes from the filled-in questionnaires. Key themes for existing codelists included identifying published codelists and updating existing codelists. Key themes for creating new codelists included defining the clinical concept, creating the codelist, finalising the codelist and sharing the codelist. Several key takeaways emerged from these discussions:

1. Existing codelists: Participants stressed the need to create precise instructions for using previous codelists and updating them effectively. This would involve documenting instances of "absence of" evidence, for example, where no relevant codelists were found.
2. New codelists: Defining the clinical concept: Need for clear processes around defining the clinical concept. Participants advocated for clearly documenting and versioning iterative searches for synonyms and consulting experts early when defining the clinical concept. The participants stressed that these components should be part of the core documentation provided with the codelist and metadata.
3. Creating codelists: A suggestion was made to provide a cover sheet template to facilitate the implementation of information from the guidance.
4. Sharing codelists: Recognition of authorship: Participants emphasized the need to establish guidelines for recognizing and crediting individuals involved in codelist creation.
5. Improve knowledge about codelists and coding systems: The group advocated for an overview of codelists and coding systems to provide context and clarity in their usage.

In summary, the small group workshop discussions yielded valuable insights for enhancing codelist creation, and documentation practices, ultimately aiming to improve the clarity and effectiveness of these processes for better healthcare data management and research.

Ethical consideration

Ethical approval was not required for this study as the current LSHTM policy is that only research activities involving human participants, their data, or their biological material must be submitted to and reviewed by the relevant LSHTM research ethics committee[22]. The workshop is considered a professional involvement activity, and not participation in a study; therefore no informed consent is required. We also confirmed these with the LSHTM ethics team in their response "The current LSHTM policy is that only research activities involving human participants, their data, or their biological material must be submitted to and reviewed by the relevant LSHTM research ethics committee. Approval must be in place before the research starts. We do not expect to review literature reviews as there are no human participants, individual level human data, or biological material. We also do not expect to review public/professional 'involvement' activities. Involvement in research means research that is done 'with' or 'by' the people involved, not 'to', 'for' or 'about' them. It just allows people with relevant experience contribute to how research is designed, conducted and disseminated."

Results

Below we provide a 10-step checklist (Table 1), comprising 28 items, with accompanying guidance on each step. We provide a filled-in example of the checklist in Table 2.

Table 1. Checklist.

	Step No	Item	Information to be provided
Metadata			
Metadata	0	a. Name	*What is the name of the codelist?*
		b. Author(s)	*Who created the codelist?*
		c. Date finalised	*When was the codelist finalised?*
		d. Target data source	*What data is the codelist designed to be used with?*
		e. Terminology	*What is the terminology? (e.g., SNOMED, ICD)*
Define a clinical concept			
Define	1	a. Concept	*What is the clinical concept (e.g., the disease, drug, test result, etc...) of interest?*
		b. Timeframe	*Should the codelist capture new, current, and/or previous events?*
		c. Accuracy	*Should the codelist capture probable or definite codes?*
		d. Setting	*What is the (health care) setting (e.g., primary care, hospital care)?*
Identify and evaluate existing codelists			
Search	2	a. Sources searched	*Which sources were searched (e.g., internet search, codelist repositories)?*
		b. Existing codelists found	*Which suitable codelists did you find?*
Verify	3	a. Verified by others	*Which information is available to verify the quality of suitable codelists?*
		b. Verified by yourself	*Which checks did you conduct to verify the quality of suitable codelists?*
Reference	4	a. Existing codelists used	*Are you making use of any existing codelists? If yes, reference these, and specify how they are being used.*
Create a new codelist			
Prepare	5	a. Synonyms	*What are synonyms and related words for the clinical concept (e.g., different names for a disease/drug) and how did you identify these (e.g., source of clinical knowledge)?*
		b. Exceptions	*What should not be included in the codelist?*
Create	6	a. Method used	*Which method (e.g., a script, a tool) did you use to create the draft codelist?*
		b. Search terms	*Which search terms, and if applicable, exclusion terms did you use?*
		c. Hierarchy used to extend search	*Did you use a dictionary hierarchy (e.g., ICD-10 chapters, SNOMED-CT concepts) to modify your search? If yes, specify.*
		d. Decisions made while iterating	*Which decisions did you make while iteratively refining the draft codelist?*
		e. (Optional) Categories	*Did you specify subcategories within the codelist? If yes, specify.*
Review, finalise and publish			
Review	7	a. Reviewers	*Who reviewed the codelist and what expertise did reviewers have?*
		b. Scope of review	*What was reviewed (Just the draft codelist or also the method, terms, etc..)?*
		c. Evidence of review	*Where is the review process documented?*
Checks	8	a. Internal checks	*What method(s) were used for internal checks, if any, and what are the findings?*
		b. External checks	*What method(s) were used for external checks, if any, and what are the findings?*
Publish	9	a. Codelist published	*Where is the codelist published?*
		b. Resources published	*Where are the resources used to create the codelist (e.g., scripts, list of terms)?*

Table 2. Example of filled in checklist.

	Step No	Item	Information to be provided
Metadata			
Metadata	0	a. Name	*Atopic eczema*
		b. Author(s)	*Julian Matthewman*
		c. Date finalised	*1st January 2023*
		d. Target data source	*CPRD Aurum January 2023 release*
		e. Terminology	*SNOMED CT (mapped to CPRD MedCodeId)*
Define a clinical concept			
Define	1	a. Concept	*Atopic dermatitis/atopic eczema*
		b. Timeframe	*Current and previous*
		c. Accuracy	*Also including codes for unspecified forms of eczema that may be atopic*
		d. Setting	*Clinical records from UK primary care*
Identify and evaluate existing codelists			
Search	2	a. Sources searched	*Internet search, HDR UK phenotype library, LSHTM datacompass, opencodelists*
		b. Existing codelists found	*Identified a number of codelists but none for CPRD Aurum; one study describing validation of eczema codelists was found: Abuabara et al. 2017 (10.1016/j.jid.2017.03.029)*
Verify	3	a. Verified by others	*See validation study above*
		b. Verified by yourself	*No further checks conducted as codelists could not be used directly*
Reference	4	a. Existing codelists used	*Medcodes from Abuabara et al. 2017 (10.1016/j.jid.2017.03.029) used to crosscheck new codelist*
Create a new codelist			
Prepare	5	a. Synonyms	*Identified from existing codelist, including Eczema, atopic dermatitis, Besnier's prurigo*
		b. Exceptions	Non-atopic forms of eczema as specified on the websites of the US (https://nationaleczema.org/eczema/types-of-eczema/) and UK (https://eczema.org/information-and-advice/types-of-eczema/) eczema societies

	Step No	Item	Information to be provided
Create	6	a. Method used	*Used search terms and exclusion terms in a script while iteratively refining terms*
		b. Search terms	*Search terms: eczema, atopic dermatitis, besnier's prurigo, allergic dermatitis* *Exclusion terms: fh, family history, contact, dyshidrotic, neurodermatitis, nummular, seborrheic, stasis, asteatotic, discoid, ear, otitis, auditory canal, eyes, eyelid, facial, female genital, vulval, hand, male genital, pompholyx, dyshidrotic, scalp, seborrhoeic, cradle cap, varicose, gravitational, pustular, erythrodermic, infectious, psoriasis, psoriasiform, immunodeficiency, vesicular, friction, hyperkeratotic, venous eczema, lip licking, desiccation, papular, drug eruption, infective, craquele*
		c. Hierarchy used to extend search	*Checked for codes with the same SnomedCTConceptId and codes with a descendant Read code*
		d. Decisions made while iterating	*In addition to non-atopic eczema from the eczema society website, also identified other non-atopic forms and other irrelevant codes, including erythrodermic eczema (erythroderma), infectious eczematoid dermatitis (which is likely non-atopic), psoriasis, immunodeficiency syndromes, friction eczema, lip licking eczema, desiccation eczema, papular eczema, drug eruptions*
		e. (Optional) Categories	*Symptom and diagnosis codes only (i.e., no codes for referrals, drugs, history of, etc..), definite atopic eczema (i.e., no codes for eczema that is possibly atopic)*
Review, finalise and publish			
Review	7	a. Reviewers	*Sinéad Langan (dermatologist and expert on atopic eczema research using electronic health records), Julian Matthewman (codelist author; clinician; conducted multiple studies on atopic eczema using UK primary care data)*
		b. Scope of review	*Both the draft codelist and search and exclusion terms were reviewed*
		c. Evidence of review	*The review process is documented in a GitHub issue thread at (...)*
Check	8	a. Internal checks	*Checked the number of observations in the CPRD Aurum (2023/03) code browser for the different categories: full codelist: 17.4 million diagnosis and symptom codes: 16.8 million definite atopic eczema: 6.4 million*
		b. External checks	*No checks were performed using external data*
Publish	9	a. Codelist published	*The codelist is published on LSHTM datacompass and the study GitHub repository*
		b. Resources published	*All resources are available at the study GitHub repository, including scripts and terms*

Guidance

Step 1: Define

To find or create a suitable codelist, it is necessary to clearly state the following: Firstly, (**1a - Concept**) state what the codelist intends to capture (e.g., a disease, drug, test results, etc..). Secondly, (**1b - Timeframe**) state if current (prevalent), new (incident) or previous events are of interest (e.g., a codelist for incident asthma may only aim at capturing codes indicating a first occurrence of asthma not including asthma-related administrative or treatment codes which are likely to indicate ongoing asthma). Thirdly, (**1c - Accuracy**) state if the codelist should prioritise sensitivity (i.e., includes codes "probably" indicating the clinical phenotype, e.g., "suspected asthma", "referred to asthma clinic") or specificity (e.g., includes codes that "definitely" match the concept)? Finally, (**1d - Setting**) state where the codes occur (e.g. the health care setting such as primary care or hospital care and what types of codes are included e.g. diagnostic codes, referrals, administrative codes, disease history codes). Together, this information makes up a clinical concept (e.g., "codes definitely describing current or previous asthma in primary care, including diagnostic, treatment, administrative and disease history codes").

Step 2: Search

(2a – Sources searched) Existing codelists that match your requirements can be identified (via an internet search (e.g., use a search-engine to search for "asthma codelist CPRD"), a search of publication databases, codelist repositories (e.g., the HDR UK phenotype library) or through existing collaboration and networks. Document which sources were searched. **(2b - Existing codelists found)** This search does not need to be systematic, but rather should identify codelists that may be directly reused or codelists that can help in creating a new codelist. To choose potentially suitable codelists, check the codelist metadata, including which clinical concept the codelist aims to capture, when the codelist was created, which database it was used in, which terminology, and which version of the terminology was used (as different versions of the same data source and terminology can contain different codes), and if there are any copyright restrictions. Codelists in other terminologies may also be useful, especially if these can be reliably mapped to the terminology of interest; however, this is not always possible. Document which suitable codelists you found.

Step 3: Verify

In addition to matching your requirements (in terms of concept, terminology, etc.) the quality of existing codelists needs to be verified. **(3a - Verified by others)** Identify which information is available, besides the metadata, to allow you to judge if the codelist was created using good practice. Projects or published studies dedicated to, or including codelist validation, may be of particular interest[23]. (**3b - Verified by yourself**) If available information isn't sufficient to judge the quality of an existing codelist, various checks can be conducted depending on the specific use-case. The codelist may be cross-checked with other existing codelists to verify if different authors consistently include the same codes. A review of the existing codelist may be performed, similar as would be done for a newly created codelist (see Step 7). If you have access to your study data or the number of observations for each code, you may also check the number of records the codelist retrieves, which may be compared to expectations based on clinical knowledge or previous studies.

Step 4: Reference

(**4a - Existing codelists used**) Any existing codelists that are used should be referenced, giving credit to the author(s), and making it easy for others to evaluate your study, or find and adapt the codelist for their own purposes. You should reference whether you have identified a codelist that suits your purposes without modification, whether it required changes to be suitable for your study, or whether it was used to check or inform the creation of a new codelist. You should also state what the existing codelist was originally used for. We suggest wording such as "codelist(s) for [clinical concept] are from/were adapted from/were cross checked with …". References to existing codelist should include the author(s), year, and permanent identifier (such as a DOI, URL or manuscript reference). You may include these references directly as part of this checklist, in your study or codelist repository (see Step 8), or the section of your manuscript or manuscript appendix that describes study variables.

Step 5: Prepare

(5a - Synonyms) Identify synonyms and related words to the clinical concept (e.g., "asthma" for an asthma codelist; "stomach/gastric", "cancer/neoplasm/malignant tumour", etc., for a stomach cancer codelist; "beta-blocker", "beta-adrenoceptor-antagonist", and substance and trade names for a beta-blocker codelist). Consulting and referencing sources of clinical information can be useful. For example Medical Subject headings on PubMed[24], clinical knowledge summaries and guidelines (such as those provided by the National Institute for Health and Care Excellence (NICE) in the UK[25]), and websites of patient organisations may all contain useful information. **(5b - Exceptions)** At this stage, identifying exceptions to the concept that shouldn't be included in the codelist is also important (e.g., if only "allergic" forms of asthma should be included, identify the words "non-allergic", "exercise-induced", etc.).

Step 6: Create

In this step, you create and iteratively refine a draft codelist. (**6a - Method used**) This can be done in a variety of ways. Guidance on the use of specific methods for creating codelists is available elsewhere, including on using Stata scripts[8], online tools[7], and for specific use-cases, such as drug codelists[10]. **(6b - Search terms)** Most approaches will involve searching a dictionary (also referred to as browser) firstly using search terms that correspond to the clinical concept or synonyms thereof, and secondly using exclusion terms to exclude codes that should not be in the codelist. For example, you create a script that searches for a list of predefined search terms (e.g., "asthma", "inhaler", etc..) and then exclude terms based on predefined exclusion terms (e.g., "referral", "review", etc..). Once finalised, report this list of search terms, and if applicable, exclusion terms. (**6c - Hierarchy used to extend search**) Make use of dictionary hierarchies, e.g., through

checking codes that are in the same or a descendant chapter as already included codes, to identify further codes that are related but may have different names or labels (e.g., check which other names for a disease or brand names for drugs may be included in the same Read code or ICD chapter or SNOMED-CT concept). **(6d - Decisions made while iterating)** When developing the draft codelist, the search should be iteratively refined by repeatedly checking the retrieved and excluded codes and adding terms to the list of search terms and exclusion terms. It may be better to also include codes where you are unsure if they should be in the codelist, as it is easier to exclude codes in the review stage than it is to add codes. Record important decisions made while refining the search, e.g., document the reasons for in- or exclusions. If necessary, revisit the definition of the clinical concept, and record additional decisions in descriptions or comments. **(6e - *Categories*)** You may want to specify categories within the codelist, e.g., incident and prevalent codes, more sensitive or specific, only diagnosis codes or diagnosis and administrative codes, (e.g., allowing for the conduct of secondary or sensitivity analyses).

Step 7: Review

Your codelist, and how it was created, needs to be reviewed to check for omissions and mistakenly included codes. **(7a - Reviewers)** A suitable reviewer with relevant knowledge about your clinical concept of interest and experience of the health care setting of your study should be identified. Reviewers may be within your research group, or you may need to reach out to other researchers in the field (e.g., an asthma codelist may be reviewed by a general practitioner, asthma researcher or internal medicine physician). The actual review process can be handled in real time or asynchronously (e.g., via email or a GitHub issue thread). Having multiple reviewers that need to agree on the final codelist can further increase trust in the review process. **(7b - Scope of review)** The reviewer(s) should first read the description of the clinical concept, then, for each of the codes in the draft codelist, decide if the code is appropriate to include. Reviewing only the codelist, without reviewing the process of how it was generated risks missing codes that should be included; therefore, the method of how the codelist was created should also be reviewed. It is particularly important to give the full list of search terms and exclusion terms (e.g., are all terms included that could possibly refer to asthma?). Make sure to implement all the required changes and re-review if necessary. Whether or not to re-review is up to your judgment, but in general it will be more important when new search terms need to be added as compared to when only a few codes need to be dropped. **(7c - Evidence of review)** During the review process, interactions between the reviewer(s) and codelist creator(s) should be documented, e.g., via a GitHub Issue thread, or a spreadsheet where reviewers mark each code with yes/no or possible/probable/unlikely (e.g., "referral to asthma clinic", may be marked as codes to be excluded, or codes to be included in a category of "possible asthma").

Step 8: Check

Where possible, code lists should be checked against the database they were created for (internal) or other data sources (external). (**8a – Internal checks**) Internal methods within the intended database could include the reporting of the numbers of individuals who were identified with the clinical concept of interest and potential sensitivity analyses comparing versions of the code list with different inclusion/exclusion criteria applied. (**8b – External checks**) External checks could include the comparison of prevalence and incidence measured within the dataset to external literature or a validation study using GP questionnaires to investigating differences between clinical diagnoses and electronic recording. More detail on validation methods can be found in a previous publication[23].

Step 9: Publish

Finally, you should publish your codelist and metadata required by reporting guidelines such as RECORD. You should also publish resources used to create the codelist and related documentation to help readers to review, evaluate or reproduce your study, and reuse or adapt your codelist for future work. **(8a - Codelist published)** Codelists can be uploaded to general purpose repositories, ideally adhering to FAIR (Findable, Accessible, Interoperable, Reusable) principles[26]. Examples of such repositories include zenodo.org or the Open Science Framework. You may also be able to adhere to FAIR principles when using your organisation's research output repository, a GitHub or Gitlab repository, or uploading your codelist(s) as supplemental materials to your study. Codelists should be shared in a suitable format that is both human- and machine-readable (.txt, or .csv). **(8b - Resources published)** Share all resources used to create the codelist, such as search terms, scripts, and references, alongside the codelist. Depending on where the codelist is hosted, there may be predefined fields for metadata, or metadata can be included as part of the checklist.

Discussion

We have developed a checklist to support the creation, adaptation, and re-use of high-quality codelists for research using routinely collected health data, accompanied by step-by-step guidance. These were developed by researchers with relevant expertise and experience including members of the EHR research group at LSHTM, which has employed codelist based data extraction for hundreds of studies for a large range of health-related topics. In Table 2 we include an example of a filled in checklist.

We expect these guidelines to be implemented by a wide range of institutions and research groups, including the EHR group at LSHTM. The guidelines can be used to train new EHR researchers, and develop or strengthen internal guidelines for publishing codelists. Developers of code list sharing platforms will also benefit from these guidelines to identify metadata that is required to allow codelists to be updated and reused. In comparison to previously published recommendations, the checklist and guidance here aim to be as universally applicable as possible within a research context, assuming as little as possible about the way of working, type of codelists to be created, type of terminology used, or tools used to create the codelist. As a consequence, it is not

possible to cover every specific case in detail, therefore more narrow guidance may be useful. Examples of more specific guidance include guidance on creating drug codelists[10], SNOMED-CT codelists using concept hierarchies[16–18], codelists using Stata scripts[8], codelists using the "termset" method[7].

The guidance was developed with more challenging coding systems in mind, such as SNOMED-CT and Read codes, which have a complex or overlapping hierarchical structures. The checklist is designed to cope with this complexity, however some steps of the codelist creation process in other settings (e.g. using only ICD coding) may be simplified.

The guidance was developed with research as the use case; however codelists developed for research may end up being used in clinical practice. Further guidance, developed with public, patient, and healthcare worker input, is needed for a clinical care setting to maximise clinical benefit and prevent avoidable harm.

This guidance underwent different validation steps[27], including a literature search, pilot testing and survey of peers. We have published the guidance in NIHR Open Research to support collaboration with the wider EHR community through open peer review, and to enable others to build upon the ideas presented here. Subsequent iterations, subject to funding, should involve pilot testing and input from larger groups of stakeholders, to ensure recommendations are useful for EHR researchers working in a range of different settings and on different topics.

While codelists are shared alongside the majority of studies that use them (a recent review found about 70% reported at least one diagnostic or treatment code), the resources used to create these codelists are rarely shared[28]. Besides journals necessitating (as with analytical code)[29,30] that codelists be published alongside manuscripts, data providers and research organisations should be encouraged to establish and maintain repositories that facilitate sharing of more complete codelist information. Future research may review current codelist banks with a view to improving the completeness of information captured.

Conclusion

Codelists form the foundation of research using routinely collected health data, however they may often be of suboptimal standard, not capturing what they are supposed to capture, and the way in which they are created and shared often precludes reuse and reproducibility. With this work, we provide a checklist, and step-by-step guidance, to help researchers adhere to best practice.

Data availability

Zenodo: Data for "Checklist and guidance on creating codelists for electronic health records research"; https://zenodo.org/doi/10.5281/zenodo.10852954[31]

This project contains the following data:

- Example codelist
- Questionnaire

Acknowledgements

We thank all members of the Electronic Health Records research group at the London School of Hygiene for testing the developed checklist and guidance and giving feedback and recommendations, including those named as co-authors, and Mia Harley, Marleen Bokern, Astrid Coste, Harriet Forbes and Alasdair Henderson, who have given their permission for their names and affiliations to be included in this publication.

References

1. Farmer R, Mathur R, Bhaskaran K, *et al.*: **Promises and pitfalls of Electronic Health Record analysis.** *Diabetologia.* 2018; **61**(6): 1241–8.
PubMed Abstract | Publisher Full Text | Free Full Text

2. **International Classification of Diseases (ICD).** [cited 2023 Dec 6].
Reference Source

3. Vuokko R, Vakkuri A, Palojoki S: **Systematized Nomenclature of Medicine-Clinical Terminology (SNOMED CT) clinical use cases in the context of Electronic Health Record systems: systematic literature review.** *JMIR Med Inform.* 2023; **11**: e43750.
PubMed Abstract | Publisher Full Text | Free Full Text

4. 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.
PubMed Abstract | Publisher Full Text

5. Altman DG, Simera I: **Using reporting guidelines effectively to ensure good reporting of health research**. In: Moher D, Altman DG, Schulz KF, Simera I, Wager E, editors. *Guidelines for Reporting Health Research: A User's Manual*. 1st ed. Wiley, 2014; 32–40.
Publisher Full Text

6. Benchimol EI, Smeeth L, Guttmann A, *et al.*: **The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement.** *PLoS Med.* 2015; **12**(10): e1001885.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Williams R, Brown B, Kontopantelis E, *et al.*: **Term sets: a transparent and reproducible representation of clinical code sets.** Olier editor. *PLoS One.* 2019; **14**(2): e0212291.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Watson J, Nicholson BD, Hamilton W, *et al.*: **Identifying clinical features in primary care Electronic Health Record studies: methods for codelist development.** *BMJ Open.* 2017; **7**(11): e019637.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Davé S, Petersen I: **Creating medical and drug code lists to identify cases in primary care databases.** *Pharmacoepidemiol Drug Saf.* 2009; **18**(8): 704–7.
PubMed Abstract | Publisher Full Text

10. Graul EL, Stone PW, Massen GM, *et al.*: **Determining prescriptions in electronic healthcare record data: methods for development of standardized, reproducible drug codelists.** *JAMIA Open.* 2023; **6**(3): ooad078.
PubMed Abstract | Publisher Full Text | Free Full Text

11. **What are codelists and how are they constructed?** Bennett Institute for

Applied Data Science, 2023; [cited 2023 Oct 18].
Reference Source

12. **Phenotype Library.** [cited 2023 Dec 6].
Reference Source

13. **OpenCodelists.** [cited 2023 Dec 6].
Reference Source

14. **LSHTM data compass.** [cited 2023 Dec 6].
Reference Source

15. **Primary care unit.** [cited 2023 Dec 6]. CPRD @ Cambridge - Codes Lists (GOLD).
Reference Source

16. Willett DL, Kannan V, Chu L, *et al.*: **SNOMED CT concept hierarchies for sharing definitions of clinical conditions using Electronic Health Record data.** *Appl Clin Inform.* 2018; **9**(3): 667–82.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

17. Chu L, Kannan V, Basit MA, *et al.*: **SNOMED CT concept hierarchies for computable clinical phenotypes from Electronic Health Record data: comparison of intensional versus extensional value sets.** *JMIR Med Inform.* 2019; **7**(1): e11487.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

18. Elkheder M, Gonzalez-Izquierdo A, Qummer Ul Arfeen M, *et al.*: **Translating and evaluating historic phenotyping algorithms using SNOMED CT.** *J Am Med Inform Assoc.* 2023; **30**(2): 222–232.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

19. Bahr NJ, Nelson SD, Winnenburg R, *et al.*: **Eliciting the intension of drug value sets - principles and quality assurance applications**. *Stud Health Technol Inform*. 2017; **245**: 843–847.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

20. Fung KW, Xu J, Gold S: **The use of inter-terminology maps for the creation and maintenance of value sets.** *AMIA Annu Symp Proc.* 2020; **2019**: 438–47.
PubMed Abstract | **Free Full Text**

21. Li L, Grando A, Sarker A: **A data-driven iterative approach for semi-automatically assessing the correctness of medication value sets: a proof of concept based on opioids.** *Methods Inf Med.* 2021; **60**(S 02): e111–9.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

22. LSHTM: **Ethics.** Research governance and integrity, [cited 2024 Mar 22].
Reference Source

23. Herrett E, Thomas SL, Schoonen WM, *et al.*: **Validation and validity of diagnoses in the General Practice Research Database: a systematic review.** *Br J Clin Pharmacol.* 2010; **69**(1): 4–14.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

24. **Home - MeSH - NCBI.** [cited 2023 Dec 6].
Reference Source

25. **CKS**. NICE. [cited 2023 Dec 6].
Reference Source

26. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR guiding principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

27. Arundel C, James S, Northgraves M, *et al.*: **Study reporting guidelines: how valid are they?** *Contemp Clin Trials Commun.* 2019; **14**: 100343.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. Tazare J, Wang SV, Gini R, *et al.*: **Sharing is caring? International society for Pharmacoepidemiology review and recommendations for sharing programming code.** *Pharmacoepidemiol Drug Saf.* Wiley Online Library, 2024; **33**(9): e5856. [cited 2024 Sep 16].
PubMed Abstract | **Publisher Full Text**

29. Loder E, Macdonald H, Bloom T, *et al.*: **Mandatory data and code sharing for research published by *The BMJ***. *BMJ.* 2024; **384**: q324.
PubMed Abstract | **Publisher Full Text**

30. Schultze A, Tazare J: **The role of programming code sharing in improving the transparency of medical research.** *BMJ.* 2023; **383**: 2402.
PubMed Abstract | **Publisher Full Text**

31. Matthewman J, Strongman H, Suffel A, *et al.*: **Data for "checklist and guidance on creating codelists for Electronic Health Records research".** [Data set]. Zenodo. 2024.
http://www.doi.org/10.5281/zenodo.10852954

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 04 October 2024

https://doi.org/10.3310/nihropenres.14932.r32963

✔ **Elizabeth Ford**
University of Sussex, Brighton, England, UK

I am happy to approve this version without reservations, the changes are sufficient for me.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 27 September 2024

https://doi.org/10.3310/nihropenres.14932.r32961

✔ **Duncan Edwards**
University of Cambridge, Cambridge, England, UK

I am content with the amendments made in response to my review.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: General Practitioner using and generated electronic health records to manage the health of individuals and populations. Clinical researcher using electronic health records.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 27 September 2024

https://doi.org/10.3310/nihropenres.14932.r32962

Shirley Wang
Howard Hughes Medical Institute - Harvard Medical School, Boston, Massachusetts, USA

The authors adequately addressed my comments

Competing Interests: I have been an ad hoc consultant for Exponent Inc, Cytel Inc, and MITRE a federally funded research and development center for the Centers for Medicare and Medicaid Services

Reviewer Expertise: Phamacoepidemiology, meta-research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 24 June 2024

https://doi.org/10.3310/nihropenres.14709.r31892

Duncan Edwards
University of Cambridge, Cambridge, England, UK

- Is the rationale for developing the new method (or application) clearly explained?

Yes. The need for improving the quality of codelists is well recognised by those involved in this area and this guidance is more robust and detailed guidance than I have seen before, which has tended to be in the form of expert opinion from smaller groups. I suspect many researchers will immediately seek to embed this guidance in their groups' operating procedures.

- Is the description of the method technically sound?

Yes. Iterative development and trialling from an expert group, was an appropriate method. It's a weakness that the group contained only folk from LSHTM, though at least this fact is stated.

- Are sufficient details provided to allow replication of the method development and its use

by others?

Mostly. I am surprised that for 7b (or perhaps earlier) that some sort of incidence/prevalence check when a draft codelist is applied to a population is not explicitly included (i.e. comparing vs previous estimates in other studies, or versus clinical judgement by the reviewer of what a reasonable number and the gender/age breakdown would be) - in my experience this data exploration stage is where a lot of major bugs with codelists are found. The description of 7b seems insufficiently detailed without explicit mention of this issue (though perhaps it is best covered elsewhere, I'm not sure 7b is the right place) - please consider a minor amendment.

- If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No relevant source data.

- Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes. However, there is no mention of the subsequent use of codelists in clinical practice. This is a huge issue in itself, and codelists developed based upon research papers can extremely swiftly be used in clinical practice to great population benefit (but also avoidable harm).
e.g. https://www.judiciary.uk/prevention-of-future-death-reports/alexander-reid-prevention-of-future-deaths-report/
I'm not arguing that this paper should cover this issue (perhaps akin to the phase IV monitoring after drug trials) in great depth, because it is such a complex and important issue in its own right. However, zero mention of the issue seems inappropriate and fails to highlight to researchers their role and responsibility in the subsequent use by others of their codelists. So I would strongly suggest an amendment to the paper, albeit small, to mention this issue. Perhaps a paragraph in the discussion recognising the importance of phase IV monitoring of codelists subsequently used on the public, and the need for public, patient and frontline health worker engagement in developing best practice in this area too (perhaps the authors will tackle this in a subsequent paper!?).

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

No source data required

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: General Practitioner using and generated electronic health records to manage the health of individuals and populations. Clinical researcher using electronic health records.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Sep 2024

Julian Matthewman

- COMMENT 3.1: Is the rationale for developing the new method (or application) clearly explained?

Yes. The need for improving the quality of codelists is well recognised by those involved in this area and this guidance is more robust and detailed guidance than I have seen before, which has tended to be in the form of expert opinion from smaller groups. I suspect many researchers will immediately seek to embed this guidance in their groups' operating procedures.

AUTHOR RESPONSE 3.1: Many thanks for your helpful and positive comments. We will be very happy to see our guidance embedded in other groups' operating procedures.

- COMMENT 3.2: Is the description of the method technically sound?

Yes. Iterative development and trialing from an expert group, was an appropriate method. It's a weakness that the group contained only folk from LSHTM, though at least this fact is stated.

AUTHOR RESPONSE 3.2: We agree that our manuscript and future work would be improved by wider collaboration. Future endeavors are, of course, subject to funding. We have made related changes to the manuscript under reviewer 1's comments (4 and 5).

- COMMENT 3.3: Are sufficient details provided to allow replication of the method development and its use by others?

Mostly. I am surprised that for 7b (or perhaps earlier) that some sort of incidence/prevalence check when a draft codelist is applied to a population is not explicitly included (i.e. comparing vs previous estimates in other studies, or versus clinical judgement by the reviewer of what a reasonable number and the gender/age breakdown would be) - in my experience this data exploration stage is where a lot of major bugs with codelists are found. The description of 7b seems insufficiently detailed without explicit mention of this issue (though perhaps it is best covered elsewhere, I'm not sure 7b is the right place) - please consider a minor amendment.

AUTHOR RESPONSE 3.3: Your comment aligns with reviewer 1 comment 7; we have considered it in our response.

AUTHOR CHANGE 3.3: see reviewer 1 comment 7.

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No relevant source data.

- COMMENT 3.4: Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes. However, there is no mention of the subsequent use of codelists in clinical practice. This is a huge issue in itself, and codelists developed based upon research papers can extremely swiftly be used in clinical practice to great population benefit (but also avoidable harm). e.g. https://www.judiciary.uk/prevention-of-future-death-reports/alexander-reid-prevention-of-future-deaths-report/
I'm not arguing that this paper should cover this issue (perhaps akin to the phase IV monitoring after drug trials) in great depth, because it is such a complex and important issue in its own right. However, zero mention of the issue seems inappropriate and fails to highlight to researchers their role and responsibility in the subsequent use by others of their codelists. So I would strongly suggest an amendment to the paper, albeit small, to mention this issue. Perhaps a paragraph in the discussion recognising the importance of phase IV monitoring of codelists subsequently used on the public, and the need for public, patient and frontline health worker engagement in developing best practice in this area too (perhaps the authors will tackle this in a subsequent paper!?).

AUTHOR RESPONSE 3.4: Many thanks for alerting us to the potential use of research codelists in clinical care. We have now highlighted this issue in the discussion.

AUTHOR CHANGE 3.4: added to discussion p16: The guidance was developed with research as the use case; however codelists developed for research may end up being used in clinical practice. Further guidance, developed with public, patient, and healthcare worker input, is needed for a clinical care setting to maximize clinical benefit and prevent avoidable harm.

Is the rationale for developing the new method (or application) clearly explained?
Yes
Is the description of the method technically sound?
Yes

- COMMENT 3.5: Are sufficient details provided to allow replication of the method development and its use by others?

Partly

AUTHOR RESPONSE 3.5: We believe that our methods are clearer following our review of the manuscript. For example, we added a summary of questionnaire responses to the manuscript.

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No source data required
Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Report 12 June 2024

https://doi.org/10.3310/nihropenres.14709.r31895

?

Elizabeth Ford
University of Sussex, Brighton, England, UK

This article describes the development of, and then presents, a researcher checklist to improve practice around code list development for EHR research. The aim is to improve practice of code list development, and reporting of code list meta data and origins. The article is very well written with good justification and explanation. The authors used reasonable checklist development methods (review, expert consensus, feedback workshop) and there are high levels of expertise in the team, with a strong track record within EHR research. Minor corrections to manuscript:
Error in table 1: items under verify are labelled 3a and 3a – second item should be 3b.
In the example checklist, the name of the code list author is also given as the name of the reviewer. This seems to endorse the practice of self-review – however, the written guidance suggests getting outside help for review. Suggest change this to another name to encourage checklist users to seek an outside pair of eyes to review their checklist draft.
Top lines of page 9, second column "...whether it was used to check or inform the creation of a new codelist, the existing codelist." Sentence does not make sense - please correct.
Suggestions for possible revision of the checklist:
All the items in the list are well thought out and well described in the guidance section.
Overall I endorse the checklist as it is, except could an item be included in the checklist on "has this codelist been trialled in the target data and numbers ascertained and sense checked against outside sources?".
The reason I suggest this is that the NHS is putting a lot of funding into regional Secure Data Environments across England which will likely house novel datasets just being created from NHS data and providers for the first time – quality and coding culture and practices will largely be unknown. Taking and adapting off the shelf codelists from phenotype libraries seems like good practice, however, in our local experience, once patient lists ascertained via available code lists are compared against QOF returns, they are significantly under-ascertaining cases. Code lists must therefore be locally adapted.
So an additional checklist item which does not mandate but allows the option of reporting on what external sources the code list returns have been compared to, and how well it performed against these, would be useful for the next user of the code list and would encourage local teams to be cautious about applying code lists without checking them.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Health data science, Electronic Health Records research, Data Governance, Public Engagement

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Sep 2024

Julian Matthewman

COMMENT 2.1: This article describes the development of, and then presents, a researcher checklist to improve practice around code list development for EHR research. The aim is to improve practice of code list development, and reporting of code list meta data and origins. The article is very well written with good justification and explanation. The authors used reasonable checklist development methods (review, expert consensus, feedback workshop) and there are high levels of expertise in the team, with a strong track record within EHR research.

AUTHOR RESPONSE 2.1: Many thanks for reviewing and supporting our article, and for your considered advice.

Minor corrections to manuscript:

COMMENT 2.2: Error in table 1: items under verify are labelled 3a and 3a – second item should be 3b.

AUTHOR CHANGE 2.2: Done.

COMMENT 2.3: In the example checklist, the name of the code list author is also given as the name of the reviewer. This seems to endorse the practice of self-review – however, the written guidance suggests getting outside help for review. Suggest change this to another name to encourage checklist users to seek an outside pair of eyes to review their checklist draft.

AUTHOR RESPONSE 2.3: We have now changed the order of the two authors stated as reviewers so that the reviewer who is not the author of the codelist is stated first. In addition to getting outside help for the review, it may sometimes also be appropriate for the codelist author themselves to review the codelist if they have sufficient knowledge (which was the case in this example). Therefore we have kept the author of the codelist in the reviewer field, and additionally specified that they are the codelist author.

AUTHOR CHANGE 2.3: *Table 2: Example of filled in Checklist, under Row 7 a (Reviewers) changed to*: "Sinéad Langan (dermatologist and expert on atopic eczema research using electronic health records), Julian Matthewman (codelist author; clinician; conducted multiple studies on atopic eczema using UK primary care data)"

COMMENT 2.4: Top lines of page 9, second column "...whether it was used to check or inform the creation of a new codelist, the existing codelist." Sentence does not make sense - please correct.

AUTHOR CHANGE 2.4: Deleted "the existing codelist" to correct the sentence.

Suggestions for possible revision of the checklist:

COMMENT 2.5: All the items in the list are well thought out and well described in the guidance section. Overall I endorse the checklist as it is, except could an item be included in the checklist on "has this codelist been trialled in the target data and numbers ascertained and sense checked against outside sources?". The reason I suggest this is that the NHS is putting a lot of funding into regional Secure Data Environments across England which will likely house novel datasets just being created from NHS data and providers for the first time – quality and coding culture and practices will largely be unknown. Taking and adapting off the shelf codelists from phenotype libraries seems like good practice, however, in our local experience, once patient lists ascertained via available code lists are compared against QOF returns, they are significantly under-ascertaining cases. Code lists must therefore be locally adapted. So an additional checklist item which does not mandate but allows the option of reporting on what external sources the code list returns have been compared to, and how

well it performed against these, would be useful for the next user of the code list and would encourage local teams to be cautious about applying code lists without checking them.

AUTHOR RESPONSE 2.5: Thank you for this comment. We agree with you and encourage the reader to trial their codelist against internal and external data. Your comment aligns with reviewer 1 comment 1.7; we have considered it in our response.

AUTHOR CHANGE 2.5: see reviewer 1 comment 7.

Competing Interests: No competing interests were disclosed.

Reviewer Report 05 June 2024

https://doi.org/10.3310/nihropenres.14709.r31896

?

Shirley Wang
Howard Hughes Medical Institute - Harvard Medical School, Boston, Massachusetts, USA

This paper describes a checklist for code list development and sharing of the code lists after creation. This is very important work, and the authors should be commended for the thoughtfully developed and comprehensive checklist for creating code lists. While the questionnaire and example code lists are provided in supplemental material, the results of the questionnaire that were used to support discussion and adaptation of the checklist are not.

Minor comments:

I would contend that this checklist applies not only to electronic health records research, but also to claims based research.

All members of the task group and workshop were from LSHTM. However, the potential impact of this paper could have been far larger if the authors engaged members/perspectives outside of their own organization, from groups that have been thinking deeply about and already actively implementing processes for sharing code lists. For example, representation from groups like OHDSI, Sentinel, CPRD, N3C, etc. Such groups have experiences, preferences, and pipelines that may differ from experiences at LSHTM that would be helpful to inform what criteria would represent “good practice” for code list development. A task force gathering of experts could have been facilitated through a research society like the International Society for Pharmacoepidemiology or the International Society for Pharmacoeconomics Research. Just a thought for future work in this space.

The recommendations are predicated on investigators intending to share code lists (which I very

much support!) It would be helpful to have some discussion of the current prevalence of code list sharing (probably quite low). Do the authors have thoughts on how to encourage more routine citation of code lists? What other encouragement or culture shifts would be necessary for researchers outside LSHTM to routinely share their code lists?

Code list creation based on mapping from other ontologies is mentioned. A little more discussion or cautionary words about the hazards of doing so given imperfect mappings and unmappable concepts could be helpful.

Step 3 Verify. It would be helpful to emphasize more the importance of documenting performance characteristics of the code list (whether existing or new) or documenting the absence of such performance characteristics. If performance characteristics are available, then some details about the population in which it was validated would also be useful, for the user to understand how applicable those measurement characteristics would be to the population in which the code list is used.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Phamacoepidemiology, meta-research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 19 Sep 2024

Julian Matthewman

COMMENT 1.1: This paper describes a checklist for code list development and sharing of the

code lists after creation. This is very important work, and the authors should be commended for the thoughtfully developed and comprehensive checklist for creating code lists.

AUTHOR RESPONSE 1.1: Many thanks for reviewing our manuscript and for your kind comments and expert advice.

COMMENT 1.2: While the questionnaire and example code lists are provided in supplemental material, the results of the questionnaire that were used to support discussion and adaptation of the checklist are not.

AUTHOR RESPONSE 1.2: We have added the meeting minutes and transcribed questionnaire responses to the study's online repository.

AUTHOR CHANGE 1.2: No changes to the manuscript necessary. The repository DOI cited in the manuscript represents all versions, and now resolves to the latest version of the repository.

Minor comments

COMMENT 1.3: I would contend that this checklist applies not only to electronic health records research, but also to claims based research.

AUTHOR RESPONSE 1.3: We agree that the checklist applies to all routinely collected database studies that require the generation of codelists to answer the study question, especially those that use complex dictionaries or ontologies such as SNOMED or ATC codes that are used in many databases worldwide. We changed our terminology throughout the manuscript, and we are now using the term "routinely collected health data" instead of "electronic health records" to emphasise that this checklist applies to all types of routinely collected health data including claims data as well.

AUTHOR CHANGE 1.3: We changed the term "electronic health records" into "routinely collected health data" throughout the manuscript.

COMMENT 1.4: All members of the task group and workshop were from LSHTM. However, the potential impact of this paper could have been far larger if the authors engaged members/perspectives outside of their own organization, from groups that have been thinking deeply about and already actively implementing processes for sharing code lists. For example, representation from groups like OHDSI, Sentinel, CPRD, N3C, etc. Such groups have experiences, preferences, and pipelines that may differ from experiences at LSHTM that would be helpful to inform what criteria would represent "good practice" for code list

development. A task force gathering of experts could have been facilitated through a research society like the International Society for Pharmacoepidemiology or the International Society for Pharmacoeconomics Research. Just a thought for future work in this space.

AUTHOR RESPONSE 1.4: Our working group initially aimed to recommend an existing codelist bank for everyone to use or develop a codelist bank for LSHTM's EHR group. We quickly realised how complex and ambitious this task was, especially given the lack of allocated funding or time. We therefore prioritised developing broad internal recommendations for codelist creation and sharing that are applicable to multiple use cases. We have shared this work in our manuscript.
A future mission is to improve the quality and availability of codelist banks. We are open to working with a wider group of experts to secure funding for this work, and agree that this would increase the relevance and impact of our work

AUTHOR CHANGE 1.4: see comment 5 below as the changes are linked.

COMMENT 1.5: The recommendations are predicated on investigators intending to share code lists (which I very much support!) It would be helpful to have some discussion of the current prevalence of code list sharing (probably quite low). Do the authors have thoughts on how to encourage more routine citation of code lists? What other encouragement or culture shifts would be necessary for researchers outside LSHTM to routinely share their code lists?

AUTHOR RESPONSE 1.5 : Thank you for your comment. There has been a welcome recent push towards open science and reproducibility by publishing both programming code and codelists. A requirement for authors to publish all programming code and study code lists with any manuscript might help to create a culture of routinely sharing code lists. Additionally, encouraging data providers and research organisations to establish and maintain accessible repositories for these resources would further enhance this. An example of such an initiative is OpenSAFELY and OpenCodelists, where the sharing of codelists and programming code is an integral part of the study workflow. These platforms facilitate reproducibility and improve the overall transparency of medical research.

AUTHOR CHANGE 1.5: Added discussion (p16): While codelists are shared alongside the majority of studies that use them (a recent review found about 70% reported at least one diagnostic or treatment code), the resources used to create these codelists are rarely shared [28]. Besides journals necessitating (as with analytical code)[29,30] that codelists be published alongside manuscripts, data providers and research organisations should be encouraged to establish and maintain repositories that facilitate sharing of more complete codelist information. Future research may review current codelist banks with a view to improving the completeness of information captured.

COMMENT 1.6: Code list creation based on mapping from other ontologies is mentioned. A little more discussion or cautionary words about the hazards of doing so given imperfect mappings and unmappable concepts could be helpful.

AUTHOR RESPONSE 1.6: We have added some words of caution describing the challenges of mapping between dictionaries and ontologies.

AUTHOR CHANGE 1.6: (underlined text added to Methods p6):
The use of inter-terminology maps is recommended to check for codelists completeness when codelists exist in multiple terminologies (e.g. when creating a codelist in SNOMED CT, map an existing ICD-10 codelist to SNOMED and check for overlap and differences).(20) However, caution is needed when mapping terms from different ontologies as they may have been created for different purposes (e.g., documentation, billing, registries, referrals or information sharing) and are often used in different care settings (e.g., SNOMED CT in primary care in the UK and ICD-10 codes in secondary care).

COMMENT 1.7: Step 3 Verify. It would be helpful to emphasize more the importance of documenting performance characteristics of the code list (whether existing or new) or documenting the absence of such performance characteristics. If performance characteristics are available, then some details about the population in which it was validated would also be useful, for the user to understand how applicable those measurement characteristics would be to the population in which the code list is used.

AUTHOR RESPONSE 1.7: Thank you for this comment which links to comments from other reviewers. In alignment with Herrett el al.'s (2010) publication on validation methods of diagnoses in CPRD, we added an additional step to the checklists (Step 8 – Checks) which offers the opportunity to comment any internal or external validation steps. Further, we expanded in the main text on what these internal and external validation steps might look like.

AUTHOR CHANGE 1.7A (Added step 8 to Table 1p9):
Check (8)
a. Internal checks
What method(s) were used for internal checks, if any, and what are the findings?
b. External checks
What method(s) were used for external checks, if any, and what are the findings?

AUTHOR CHANGE 1.7B (Explanation added to Results p14):
Where possible, code lists should be checked against the database they were created for (internal) or other data sources (external). Internal methods within the intended database could include the reporting of the numbers of individuals who were identified with the clinical concept of interest and potential sensitivity analyses comparing different versions of the code list with different inclusion/exclusion criteria applied. External checks could consist of a comparison of prevalence and incidence measured within the dataset to that published in external literature, or a validation study involving primary data collection to estimate the

sensitivity, specificity or positive predictive value of a certain code list. More detail on potential options for conducting a code list validation can be found in a previous publication (Herrett et al.).

AUTHOR CHANGE 1.7C: (Added step 8 to Table 2, example of filled in checklist, p18)
Check (8)
a. Internal checks
Checked the number of observations in the CPRD Aurum (2023/03) code browser for the different categories:
full codelist: 17.4 million
diagnosis and symptom codes: 16.8 million
definite atopic eczema: 6.4 million
b. External checks
No checks were performed using external data

Competing Interests: No competing interests were disclosed.