

International Journal of Population Data Science

Journal Website: www.ijpds.org



Swansea University
Prifysgol Abertawe

Variation in Primary Care Diagnostic Codes Used to Define Health Conditions for Research: Comparison of 45 Health Data Research UK Phenotype Library Codelists for Eight Health Conditions

Brian Kelly¹, Clare MacRae¹, Regina Prigge¹, Harry Hemingway², Natalie Fitzpatrick², Ana Torralbo², and Bruce Guthrie^{1,*}

Submission History

Submitted:	05/02/2026
Accepted:	18/05/2026
Published:	xx/xx/20xx

¹ Advanced Care Research Centre, Usher Institute, The University of Edinburgh, Usher Building, 5-7 Little France Road, Edinburgh EH8 9AG

² Institute of Health Informatics, University College London, 222 Euston Road, London, NW1 2DA

Abstract

Introduction

The use of electronic healthcare records (EHRs) has become standard practice in most healthcare settings in many countries. These can be utilised for research purposes by developing codelists to define cohorts and health conditions of interest. In the UK the Health Data Research UK (HDR UK) Phenotype Library is a resource where researchers can access multiple codelists deposited by other researchers for a range of medical conditions.

Objectives

This paper examines variation in the published codelists in the HDR UK Phenotype Library for a number of conditions.

Methods

We extracted codelists held within the HDR UK Phenotype Library for eight common conditions (asthma, autism, coronary heart disease, chronic obstructive pulmonary disease, depression, hypertension, obesity, and type 2 diabetes). We categorised included codes as diagnostic, disease related, administrative, and unrelated to the condition. We explored the degree of agreement between codelists by estimating kappa statistics for codelists for each condition and category.

Results

There were 48 published codelists for the eight conditions, with wide variation in the number of codes used to define each condition (for example from 39 to 247 codes in eight asthma codelists). Only codelists for chronic obstructive pulmonary disease and depression consistently achieved fair (kappa 0.21-0.40) moderate agreement (kappa 0.41-0.60), with other conditions consistently achieving fair agreement at best. Agreement was higher for diagnostic codes, and lowest for administrative codes.

Conclusions

The HDR UK Phenotype Library publishes researchers' codelists to facilitate greater transparency and replicability of studies. This study finds that agreement between codelists in terms of the codes included was low. Further research is needed to determine how such variation alters identification of which individuals have conditions of interest.

Keywords

electronic health record, administrative data research, epidemiologic studies

*Corresponding Author:

Email Address: bruce.guthrie@ed.ac.uk (Bruce Guthrie)



Introduction

The use of electronic healthcare records (EHRs) has become standard practice in most healthcare settings in many countries [1, 2]. EHRs are a valuable resource for epidemiological research, as they are readily available to researchers and include rich longitudinal data on entire populations [3]. EHRs vary in how complete a record they are of an individual's healthcare history, but primary care EHRs are often the most complete such record, particularly in countries like the UK with strong primary care systems where primary care acts as a gatekeeper to specialist services [4, 5]. This is particularly true in UK primary care, where all residents are registered with a single general practice and records follow the patient if they change practices. Primary care EHR data is typically a mix of free-text notes and coded data of various kinds including diagnoses, symptoms, administrative data, test results, and medications. UK primary care records became nearly fully computerised in the early 2000s [6]. Most UK primary care practices historically coded events and other information about patients using Read codes, a comprehensive computerised semi-hierarchical clinical classification system [7]. Although these are in the process of being replaced by the Systematised Nomenclature of Medicine Clinical Terms [8], many datasets including SAIL Databank remain coded in Read.

To utilise EHRs for research purposes, it is necessary to define cohorts of interest, for example by identifying patients with particular health conditions [9]. This involves creating lists of codes that indicate that the patient has the condition of interest. Codelists may include diagnostic codes, but can also include codes that refer to the use of medication (for example, medications used for exclusively for the management of dementia) and administrative codes (for example, attendance at a diabetes clinic can be used to define the presence of diabetes). However, some medications can be used for multiple diseases and some administrative codes do not mean the presences of a disease, for example referral for screening or a test. In some circumstances, the ability of codelists to identify patients with a health condition can be improved by combining multiple data elements, such as diagnostic codes, symptom codes, medications, and test results [10, 11]. Some codes may indicate inclusion in the condition of interest, while others indicate exclusion which requires consideration of the clinical term attached to codes rather than assuming that all codes in a particular part of the hierarchy indicate the presence of the condition.

However, there are significant challenges that researchers face when using EHRs to define health conditions of interest. Firstly, routine healthcare data is recorded to support medical care for the individual and for administrative purposes like payment, not for research purposes, and clinicians themselves do not use codelists during clinical care [9]. Historically, researchers have therefore created their own codelists, ideally in a structured process that involves clinicians. However, many research studies do not report details of codes and algorithms used to create research cohorts [12–14], reducing reproducibility [13]. Secondly, even if adequately reported, researchers with differing clinical or epidemiological experience often have different views of the inclusion of certain codes; and levels of agreement between codelists or between individual researchers can be low [12]. Often it is not possible to

achieve consensus on many aspects of defining a condition or describing its sequelae [14]. This can often lead to large variations in the incidence and prevalence of conditions of interest, depending on the codelists and algorithms applied to the data [15]. Finally, the codes used by clinicians may change over time, either because new codes are created or because of changing clinical or administrative needs, so historical research codelists may also need updating [16, 17].

The choice of codelist is therefore very important for researchers, there is a risk of misclassification [18–20], which has implications when comparing results from different studies. These choices are not commonly reported, and there are calls to improve the reporting of choices made in developing codelists [9, 13, 21, 22]. In response, a number of resources have been created in recent years with the intention of documenting codelists used and phenotype definitions [23]. A UK example is the Health Data Research UK (HDR UK) Phenotype Library [24], a comprehensive, open access resource providing the research community with information, tools and phenotyping algorithms for UK electronic health records. By providing access to codelists and metadata deposited by researchers, the HDR UK Phenotype Library aims to support improvements in the quality of health data analysis through greater transparency [23]. Within the HDR UK Phenotype Library there are multiple codelists for any given condition. However, researchers still have to choose which to use, and it is unclear to what extent the available codelists vary.

This paper therefore examines how codelists for eight conditions published in the HDR UK Phenotype Library vary in terms of the Read codes included, including in how they vary by the types of code included, that relate to the same condition, for example, diagnostic, medication, procedure, and administrative codes?

Methods

Data

We extracted codelists held within the HDR UK Phenotype Library [24] for eight common conditions: asthma, autism, coronary heart disease (CHD), chronic obstructive pulmonary disease (COPD), depression, hypertension, obesity, and type 2 diabetes (accessed between 20th January 2026 and 10th March 2026, see supplementary material for exact dates). We focused on codelists that used Read codes, because UK primary care EHRs has more complete ascertainment of conditions than hospital data [4, 5].

Data Management

We report the contents of each codelist held within the HDR UK Phenotype Library for each of the eight conditions of interest, excluding codelists with multiple records where the same author had uploaded identical codelists for the same condition. Since Read codes are used for a variety of purposes beyond diagnosis (for example, to record clinic attendance for administrative or payment reasons), we classified the Read codes in each codelist into categories in order to better understand if variation between codelists at least

partly reflected broad choices about the type of codes to include as well as choices about each code individually. Supplementary material section 10 details the process used and provides examples, but in brief, our original intention was to distinguish diagnostic and administrative codes, but this was elaborated through discussion and consensus for codes which did not easily fit this dichotomy (for example, Read codes for the resolution of a diagnosis are evidence that the person had the condition at some point, but are not diagnostic on the date they are recorded). One author (BG) developed the initial set of categories which were then applied by BK, CM and RP and finalised in discussion. All four then classified all codes and the classification applied where there was universal agreement. If there was any disagreement, then BG and CM (both clinical academics who are also general practitioners and therefore users of the underlying electronic health records) reached a consensus on classification. The final classification included eight categories: diagnostic, administrative, diagnostic resolved, disease related event, disease related monitoring/ management, medication only for the condition, medication also for other conditions, and code unrelated to the condition. Diagnostic codes were those we considered to have no other function than to record a diagnosis. Disease related codes were those either for events which also named a diagnosis (for example, codes for emergency hospital admission for a named condition) or codes for condition monitoring or management which named a diagnosis (for example, medication review for a named condition). Administrative codes were those recording administrative events which named a condition (for example, recording attendance or non-attendance at a diabetes clinic). We considered some codes included in codelists as unrelated to the condition being defined; either completely unrelated (for example, 'cardiac asthma' is not asthma), or too non-specific to be certain (for example, 'dietary advice for diabetes' could be advice given to someone with pre-diabetes).

Data Analysis

Initial descriptive analysis documented for each condition the number of codes used in each codelist, and the overlap between codelists. In order to explore the degree of agreement between codelists we estimated Cohen's kappa statistics between all possible pairs of codelists within each condition with 95% confidence intervals calculated using bias-corrected bootstrapping using the `kapci` command in Stata BE v18.0. We then estimate the overall kappa by condition for all codelists related to that condition, and separately for each category of Read code (diagnostic, administrative etc.). Kappa can vary between -1 (complete disagreement) to 1 (complete agreement, with zero representing no agreement beyond chance. The kappa result be interpreted as follows: values ≤ 0 as indicating worse than chance agreement and 0.01 – 0.20 as none to slight, 0.21 – 0.40 as fair, 0.41 – 0.60 as moderate, 0.61 – 0.80 as substantial, and 0.81 – 1.00 as almost perfect agreement [25]. We use summary tables and figures to illustrate these results, and report the full findings in supplementary material. Since analysis only uses published codelists, there was no requirement for ethical review.

Results

Forty-two codelists were identified for the eight included conditions. Within each condition, there was wide variation in the number of codes used to define that condition (Table 1, supplementary tables S1–S8). Asthma had eight codelists, containing between 39 and 247 codes (334 distinct codes in total). Autism had three codelists, containing between 8 and 19 codes (21 distinct codes). CHD had eight codelists, containing between 30 and 129 codes (130 distinct codes). COPD had five codelists, containing between 41 and 136 codes (136 distinct codes). Depression had nine codelists, containing between 88 and 135 codes (187 distinct codes). Hypertension had eight codelists, containing between 19 and 130 codes (145 distinct codes). Obesity had four codelists, containing between 21 and 70 codes (90 distinct codes). Type 2 diabetes had three codelists, containing between 102 and 645 codes (646 distinct codes).

In general, codelists which included more codes had a higher percentage of codes which were unique to that codelist and had a lower percentage of codes which were present in all codelists for the same condition, but there is variation between the health conditions (Table 1). For example, for the eight hypertension codelists examined, no single code was present in all eight codelists, but with the exception of the hypertension 1 codelist (Mansfield et al 2021) almost all codes were included in more than one codelist. For the nine depression codelists, most codes included in individual codelists were included in all nine codelists (ranging from 45% to 58%). In contrast, for asthma there was much more variation in the percentage of codes included in individual codelists which were included in all eight asthma codelists (ranging from 11% to 69%).

Figure 1 illustrates the proportion of codes for each condition in each category of Read code. Overall, 67% of codes examined were classified as diagnostic, 15% were disease related, 8% were administrative, and 10% were unrelated to the condition. This varied by condition; for example, 87% of Read codes for CHD were classified as diagnostic compared to 26% for asthma, and 34% of Read codes for COPD were classified as unrelated to the condition compared to 12% for hypertension. Overall, diagnostic codes were most frequent in CHD and autism, and relatively high in depression, hypertension, and type 2 diabetes; and least frequent in asthma. Disease related codes were most frequent asthma and obesity and least frequent in depression and autism. Administrative codes were most frequent in type 2 diabetes, and relatively high in obesity and COPD; and least frequent (absent) in autism. Codes that were unrelated to the condition were most frequent in COPD, depression, and asthma; and least frequent in CHD.

Figure 2 shows pairwise kappa statistics for each codelist and all other codelists for the same condition (full data is shown in supplementary tables S11a–S11h). There is variation in the level of agreement between pairs of codelists within each condition. There is no to, at best, fair agreement for asthma, autism, hypertension, and obesity; fair agreement for CHD; moderate agreement for COPD and depression. Agreement for diabetes is zero for one codelist, and moderate for the other two.

Figure 3 shows the overall kappa statistic by condition of interest, broken down by classification of Read code (the

Table 1: Number of Read codes in Each Codelist Examined, and How Often Codes in One Codelist Were Also Included in All Codelists for Each Condition

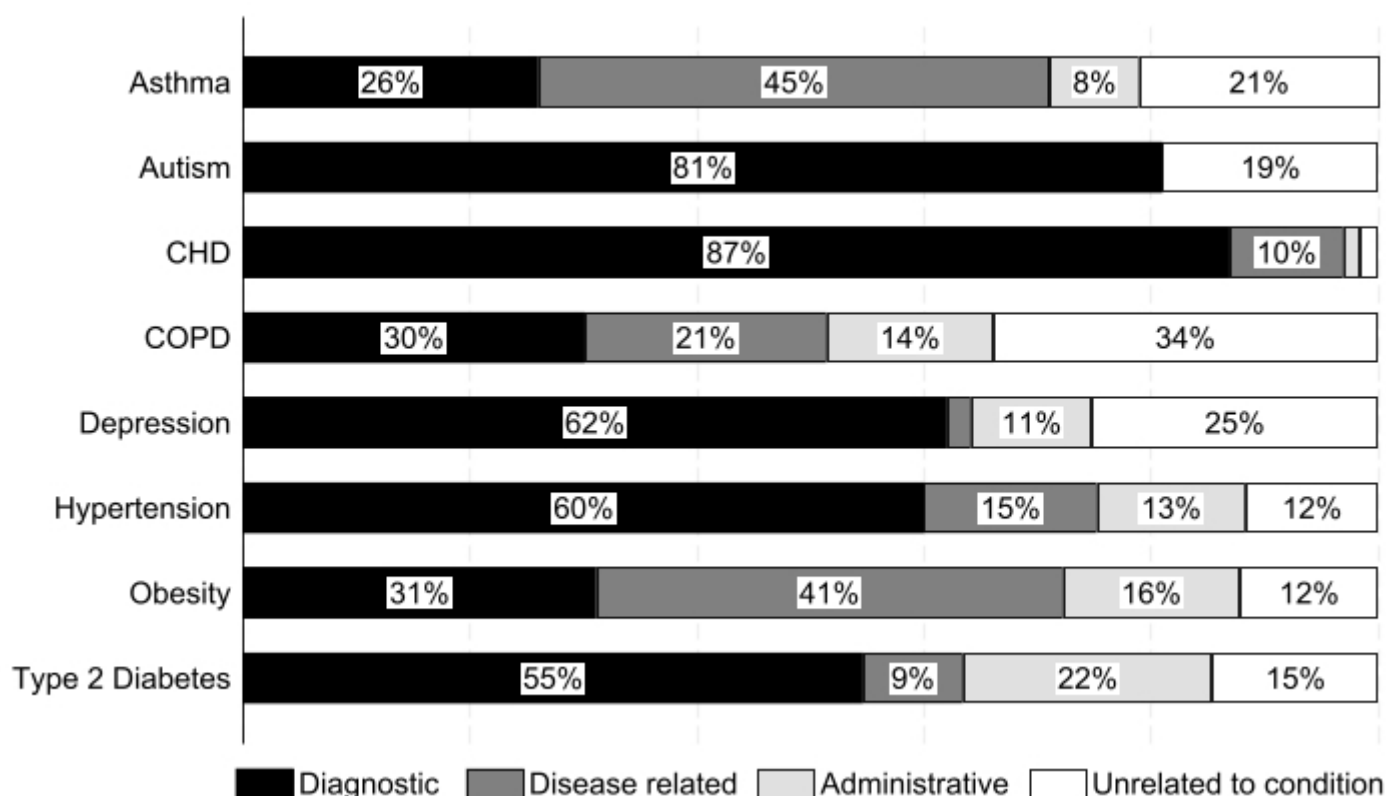
Codelist*	No. of codes included in codelist	% of codes only in this codelist	% in one or more other codelists	% in all codelists for the condition
Asthma 1	247	33.6	55.4	10.9
Asthma 2	165	9.1	74.5	16.4
Asthma 3	156	34.0	48.7	17.3
Asthma 4	146	0	81.5	18.5
Asthma 5	42	0	35.7	64.3
Asthma 6	40	0	32.5	67.5
Asthma 7	40	0	32.5	67.5
Asthma 8	39	0	30.8	69.2
Autism 1	19	21.1	42.1	36.8
Autism 2	16	6.3	50.0	43.8
Autism 3	8	12.5	0	87.5
CHD 1	129	0	81.5	18.6
CHD 2	129	0	81.5	18.6
CHD 3	62	0	61.3	38.7
CHD 4	62	0	61.3	38.7
CHD 5	30	3.3	16.7	80.0
COPD 1	136	35.3	46.3	18.4
COPD 2	67	0	62.7	37.3
COPD 3	58	0	56.9	43.1
COPD 4	55	0	54.5	45.5
COPD 5	41	0	39.0	61.0
Depression 1	135	11.9	42.9	45.2
Depression 2	131	22.9	30.5	46.6
Depression 3	121	4.1	45.6	50.4
Depression 4	107	0	43.0	57.0
Depression 5	105	0	42.0	58.1
Depression 6	105	0	42.0	58.1
Depression 7	100	0	39.0	61.0
Depression 8	100	0	39.0	61.0
Depression 9	88	0	30.6	69.3
Hypertension 1	130	15.4	84.6	0
Hypertension 2	85	0	100	0
Hypertension 3	83	0	100	0
Hypertension 4	76	1.3	98.7	0
Hypertension 5	69	0	100	0
Hypertension 6	65	0	100	0
Hypertension 7	61	3.2	96.7	0
Hypertension 8	19	0	100	0
Obesity 1	70	38.6	40.0	21.4
Obesity 2	52	23.1	48.1	28.8
Obesity 3	28	14.3	32.1	53.6
Obesity 4	21	19.0	9.6	71.4
Type 2 diabetes 1	645	82.9	2.2	14.9
Type 2 diabetes 2	105	1.0	7.6	91.4
Type 2 diabetes 3	102	0	5.9	94.1

*Supplementary tables S1a-c (asthma), S2a-c (autism), S3a-c (CHD), S4a-c (COPD), S5a-c (depression), S6a-c (hypertension), S7a-c (obesity) and S8a-c (type 2 diabetes) document all 42 codelists examined.

underlying kappa results are in supplementary material tables S12a and S12b.) Overall, there are low levels of agreement between all types of Read codes in codelists. Again, this

varies by condition, being lowest for type 2 diabetes and highest for depression. Agreement tends to be highest for diagnostic Read codes (with the exception of autism), and

Figure 1: Classification of Unique Read Codes by Condition*



lowest for administrative Read codes (which have mostly negative kappa).

Discussion

Summary of Findings

This paper aimed to explore variation between 45 HDR UK Phenotype Library codelists for eight health conditions, and the extent to which this variation differed by the classification of Read code (diagnostic, disease related, administrative, and codes unrelated to the condition). We found there was considerable variation in terms of the Read codes used to define the eight conditions we examined. For example, there are eight codelists for asthma containing between 39 and 247 Read codes. Codelists varied in the extent to which they included different categories of Read code. We found that there were generally low levels of agreement between codelists for the same condition; though this varied by condition. Codelists for depression had the greatest level of agreement ($\kappa=0.510$ consistent with moderate agreement), but all other conditions had a kappa lower than 0.250 (consistent with poor or minimal agreement), and both autism and type 2 diabetes had a $\kappa < 0$ (agreement less than would be expected at random). Not unexpectedly, all codelists included a high proportion of diagnostic codes, with more variation in whether they included other types of codes. For codelists using particular categories of codes, we found levels of agreement between codelists was greater for diagnostic and disease related codes, and lowest for administrative and unrelated codes.

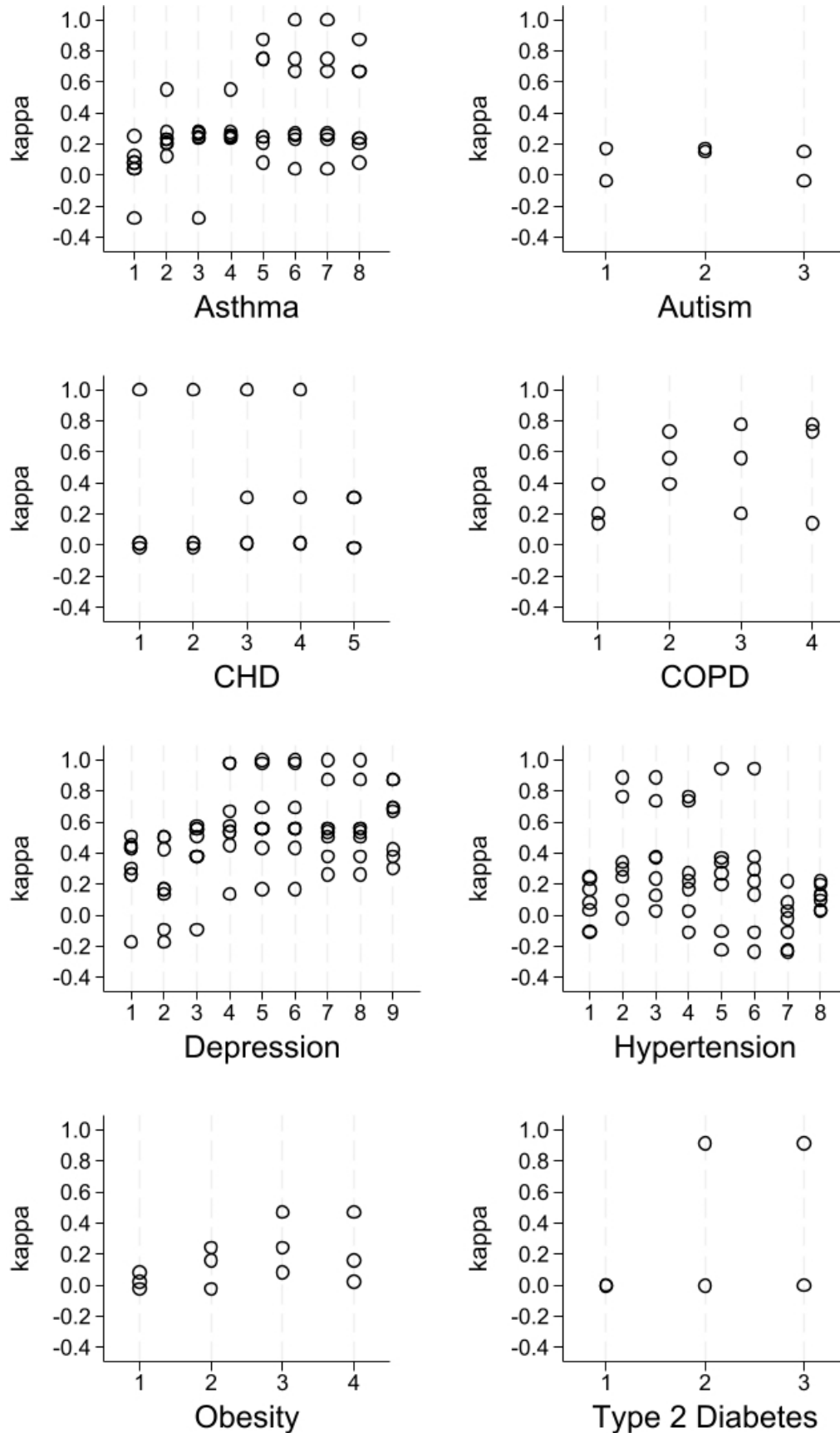
Strengths and Limitations

The main strength of this paper is that it systematically examines the extent to which published codelists agree with each other, which is under-researched. A further strength is that we used well documented primary care codelists from a single phenotype library in the UK, although this does mean that it is uncertain if the findings apply in other contexts: however, we think it very likely that this is a general problem. There are a number of limitations. First, codelists may legitimately differ because researchers have different aims or uses for the codelist, although this is not explicit in the phenotype library metadata. Related to this, our classification of Read codes into categories was arbitrary (although carried out in a systematic way), but this reflects the kinds of choices which researchers have to make when choosing codes. Second, since new Read codes were added over time and coding practices varied over time, observed differences in codelists could reflect the different set of codes available or used when the codelist was created.

Comparison with Other Literature

Our findings are consistent with previous research that identified variations in codelists used to identify health conditions [14, 15], and found that levels of agreement between different codelists can be low [12, 13]. This paper extends previous work by examining a much wider range of conditions and codelists, and by looking at agreement between different classifications of Read codes. Variation is likely to be less in other coding systems like ICD-10 because they have fewer codes and are less granular, but conversely may be worse

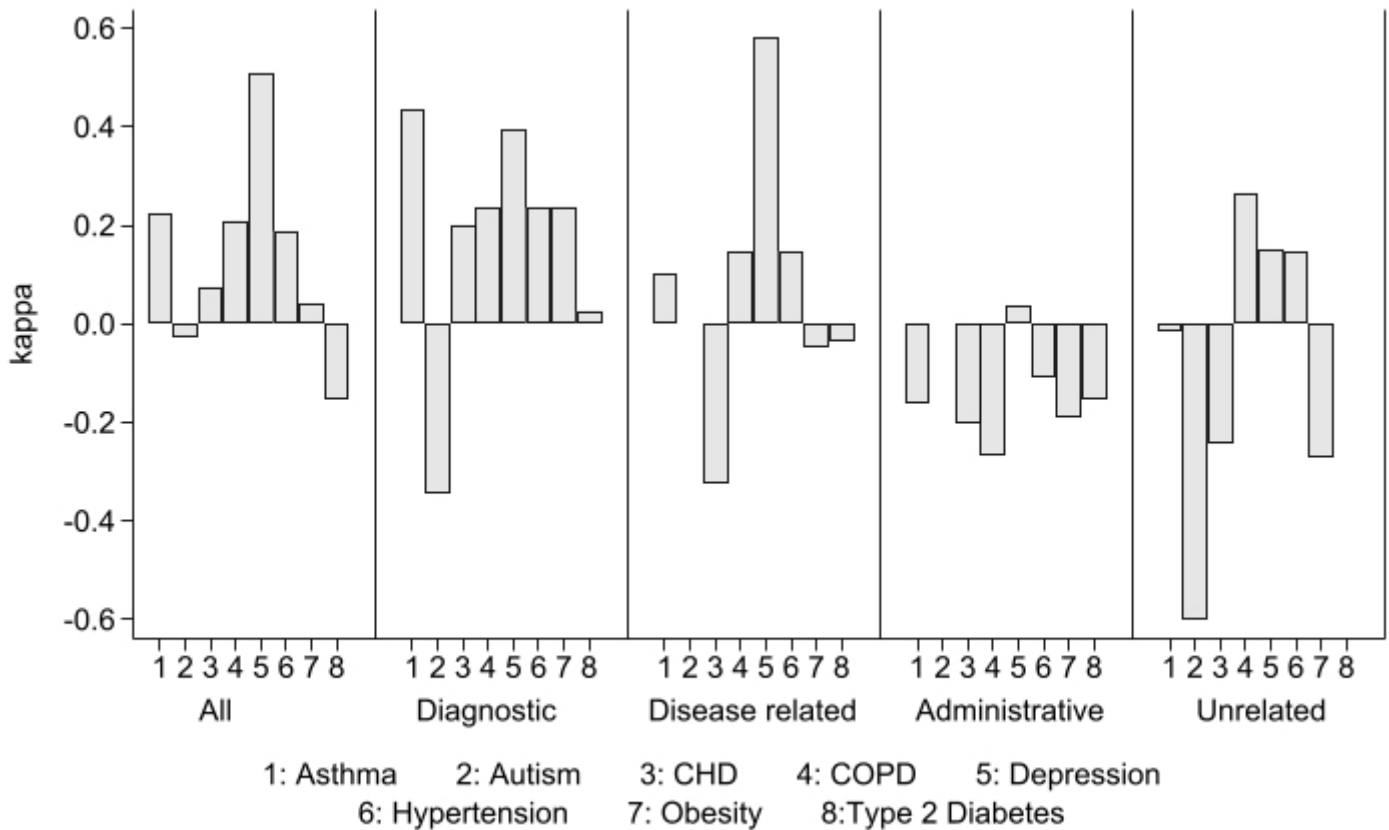
Figure 2: Pairwise Kappa Statistics Comparing Agreement Between All Pairs of Codelists for each condition



The numbers on the x-axis refer to individual codelists for that condition.
The markers represent the kappa statistic for two way comparisons
for that codelist against all other codelists for that condition

Kappa values can be interpreted as follows: values ≤ 0 as indicating worse than chance agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

Figure 3: Overall Kappa Statistic by Condition of Interest, Broken Down by Classification of Read Code*



*Because of small numbers of codes, diagnostic in figure combines: (a) diagnostic, (b) medication only for the condition; Disease related combines: (a) disease related event, (b) disease related monitoring/management; Unrelated to condition combines: (a) unrelated to condition, (b) medication also for other conditions.

Kappa values can be interpreted as follows: values ≤ 0 as indicating worse than chance agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

in the future in coding systems containing many more codes than Read such as SNOMED-CT.

Implications for Researchers

Our findings that codelists vary considerably in their choice of codes has a number of implications. Publication of codelists in phenotype libraries has improved the transparency and replicability of research, but the existence of multiple codelists with limited validation (or limited validation in the dataset being used) presents researchers with difficult choices in the face of multiple codelists on offer. Researchers could use different sets of codelists and conduct sensitivity analysis [12]. However, this is resource intensive and may not provide consistency required to deal with problems of different results arising from different codelists. There have also been calls to harmonise and standardise codelists for conditions of interest, which could lead to a single consensus codelist, or to a minimum set of codes for each condition of interest [26], but it is also important to recognise different purposes may require the use of different codes meaning that there will not always be a 'gold standard' set of codes and associated algorithms to define conditions [12]. Even if there were a 'gold standard,' it could be subject to drift over time because condition classifications and choice of codes by clinicians may

change over time [16], for example when codes are used for payment [27].

Guidance on which codelists to employ for research is outside the current aims of the HDR UK Phenotype Library. Rather, the stated aim of the HDR UK Phenotype Library is to provide a place for researchers to publish findable, consistently documented codelists, rather than to standardise or harmonise disease definitions [28]. Although the HDR UK Phenotype Library intends to better document how codelists have been validated [28], the onus remains on individual researchers to choose which phenotypes or codelists they wish to use. Similar issues apply to other phenotype libraries [29], and other coding systems, with CIPHER for example offering ICD-10 codelists which vary in terms of included codes (see for example, codelists for autism <https://phenomics.va.ornl.gov/web/cipher/search?search=autism>).

In practice therefore, researchers need to carefully consider the accuracy of published codelists (including evidence of their validity), and their relevance to the researcher's purpose and dataset (including evidence of their validity in a particular context [30, 31], since clinicians in different contexts vary in how they use clinical codes [17, 32–34]). The aim has to be to minimise measurement error which can lead to observing smaller effect sizes than exist if conditions were accurately

defined [35]. Ideally, researchers would either use codelists or algorithms validated in the data they are using, although this may not be feasible requiring judgements whose justification should be documented in publications.

There have been suggestions that the answer to all the above problems is that healthcare systems should be more standardised in how they use codes [27]. The UK Quality Outcomes Framework (QOF) is an example of such standardisation for conditions included in primary care pay-for-performance [36, 37]. The introduction of QOF in the UK led to improved recording of conditions [17]. However, such standardisation is always likely to be local, meaning that researchers still have to choose the best codelist or algorithm for their purpose and their context.

Future Research

There is a need for further research to determine how much difference variation between codelists makes to the individuals identified as having different conditions, and how the use of different categories of codes (such as administrative codes) alters the performance of codelists in identifying individuals with specific conditions.

Conclusion

Phenotype libraries help improve the transparency and reproducibility of research by publishing the codelists and algorithms used by researchers. However, published codelists vary considerably, meaning that researchers still need to think carefully about which codelist they apply in terms of its validity in their data for their particular purposes.

Acknowledgements

We would like to thank the HDR UK for the development of their phenotype library: <https://phenotypes.healthdatagateway.org/>.

Statement on Conflicts of Interest

There are no conflicts of interest.

Ethics Statement

No ethical approval was required. The study was conducted using codelists that are freely available and open access: <https://phenotypes.healthdatagateway.org/>.

Data Availability Statement

All codelists used in the analysis are freely available in the HDR UK Phenotype Library: <https://phenotypes.healthdatagateway.org/>.

AI Disclosure Statement

The authors declare that no generative AI tools were used in the preparation of this manuscript.

References

1. McMillan B, Eastham R, Brown B, Fitton R, Dickinson D. Primary care patient records in the United Kingdom: past, present, and future research priorities. *Journal of medical Internet research*. 2018 Dec 19;20(12):e11293. <https://doi.org/10.2196/11293>
2. Slawomirski L, Lindner L, de Bienassis K, Haywood P, Hashiguchi TCO, Steentjes M and Oderkirk J. 2023. Progress on implementing and using electronic health record systems: developments in OECD countries as of 2021. *OECD Health Working Papers*. 21 September 2023. Available from: https://www.oecd.org/en/publications/progress-on-implementing-and-using-electronic-health-record-systems_4f4ce846-en.html.
3. Meeks HD, Yu DZ, Fraser A, Maloney TN, Curtin K, Newman M, Smith KR. Utilize administrative data to assess impact of individual-level and neighborhood-level socioeconomic status on childhood health in a longitudinal study. *Public Health*. 2025 Mar 1;240:154-60. <https://doi.org/10.1016/j.puhe.2025.01.019>
4. MacRae C, Morales D, Mercer SW, Lone N, Lawson A, Jefferson E, McAllister D, van den Akker M, Marshall A, Seth S, Rawlings A. Impact of data source choice on multimorbidity measurement: a comparison study of 2.3 million individuals in the Welsh National Health Service. *BMC medicine*. 2023 Aug 15;21(1):309. <https://doi.org/10.1186/s12916-023-02970-z>
5. Prigge R, Fleetwood KJ, Jackson CA, Mercer SW, Kelly PA, Sudlow C, Norrie JD, Morales DR, Smith DJ, Guthrie B. Robustly measuring multimorbidity using disparate linked datasets. *Communications Medicine*. 2025 Jul 8;5(1):283. <https://doi.org/10.1038/s43856-025-00995-4>
6. Kontopantelis E, Stevens RJ, Helms PJ, Edwards D, Doran T, Ashcroft DM. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross-sectional population study. *BMJ open*. 2018 Feb 1;8(2):e020738. <https://bmjopen.bmj.com/content/8/2/e020738>
7. NHS Digital. Read coded clinical terms. secondary read coded clinical terms. Available from: <https://digital.nhs.uk/services/terminology-and-classifications/read-codes>, accessed 28th May 2025.
8. NHS Digital. SNOMED CT. Available from: <https://digital.nhs.uk/services/terminology-and-classifications/snomed-ct>, accessed 28th May 2025.

9. Liu M, Deng K, Wang M, He Q, Xu J, Li G, Zou K, Sun X, Wang W. Methods for identifying health status from routinely collected health data: An overview. *Integrative Medicine Research*. 2025 Mar 1;14(1):101100. <https://doi.org/10.1016/j.imr.2024.101100>
10. Nissen F, Morales DR, Mullerova H, Smeeth L, Douglas IJ, Quint JK. Validation of asthma recording in the Clinical Practice Research Datalink (CPRD). *BMJ open*. 2017 Aug 1;7(8):e017474. <https://doi.org/10.1136/bmjopen-2017-017474>
11. Gillmeyer KR, Lee MM, Link AP, Klings ES, Rinne ST, Wiener RS. Accuracy of algorithms to identify pulmonary arterial hypertension in administrative data: a systematic review. *Chest*. 2019 Apr 1;155(4):680-8. <https://doi.org/10.1016/j.chest.2018.11.004>
12. Gulliford MC, Charlton J, Ashworth M, Rudd AG, Toshke AM, eCRT Research Team. Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. *PloS one*. 2009 Sep 24;4(9):e7168. <https://doi.org/10.1371/journal.pone.0007168>
13. Al Sallakh MA, Vasileiou E, Rodgers SE, Lyons RA, Sheikh A, Davies GA. Defining asthma and assessing asthma outcomes using electronic health record data: a systematic scoping review. *European Respiratory Journal*. 2017 Jun 15;49(6). <https://doi.org/10.1183/13993003.00204-2017>
14. Guthrie G, Guthrie B, Walker H, James MT, Selby NM, Tonelli M, Bell S. Developing an AKI consensus definition for database research: findings from a scoping review and expert opinion using a Delphi process. *American Journal of Kidney Diseases*. 2022 Apr 1;79(4):488-96. <https://doi.org/10.1053/j.ajkd.2021.05.019>
15. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, Kadri SS, Angus DC, Danner RL, Fiore AE, Jernigan JA. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *Jama*. 2017 Oct 3;318(13):1241-9. <https://doi.org/10.1001/jama.2017.13836>
16. Zghebi SS, Reeves D, Grigoroglou C, McMillan B, Ashcroft DM, Parisi R, Kontopantelis E. Clinical code usage in UK general practice: a cohort study exploring 18 conditions over 14 years. *BMJ open*. 2022 Jul 1;12(7):e051456. <https://doi.org/10.1136/bmjopen-2021-051456>
17. Beaney T, Clarke J, Salman D, Woodcock T, Majeed A, Barahona M, Aylin P. Identifying potential biases in code sequences in primary care electronic healthcare records: a retrospective cohort study of the determinants of code frequency. *BMJ open*. 2023 Sep 1;13(9):e072884. <https://doi.org/10.1136/bmjopen-2023-072884>
18. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *Bmj*. 2010 Aug 19;341. <https://doi.org/10.1136/bmj.c4226>
19. Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC medical research methodology*. 2019 Mar 11;19(1):53. <https://doi.org/10.1186/s12874-019-0695-y>
20. Weinstein EJ, Ritchey ME, Lo Re III V. Core concepts in pharmacoepidemiology: validation of health outcomes of interest within real-world healthcare databases. *Pharmacoepidemiology and drug safety*. 2023 Jan;32(1):1-8. <https://doi.org/10.1002/pds.5537>
21. Almowil Z, Zhou SM, Brophy S, Croxall J. Concept Libraries for Repeatable and Reusable Research: Qualitative Study Exploring the Needs of Users. *JMIR human factors*. 2022 Mar 15;9(1):e31021. <https://doi.org/10.2196/31021>
22. Matthewman J, Andresen K, Suffel A, Lin LY, Schultze A, Tazare J, Bhaskaran K, Williamson E, Costello R, Quint J, Strongman H. Checklist and guidance on creating codelists for routinely collected health data research. *NIHR open research*. 2024 Sep 18;4:20. <https://doi.org/10.3310/nihropenres.13550.2>
23. Thayer DS, Mumtaz S, Elmessary MA, Scanlon I, Zinnurov A, Coldea AI, Scanlon J, Chapman M, Curcin V, John A, DelPozo-Banos M. Creating a next-generation phenotype library: the health data research UK Phenotype Library. *JAMIA open*. 2024 Jul 1;7(2):ooae049. <https://doi.org/10.1093/jamiaopen/ooae049>
24. Heath Data Research UK (HDRUK) Phenotype Library. Available from: <https://www.hdruk.ac.uk>, accessed 28th May 2025.
25. DeVellis RF. Inter-Rater Reliability. In Kempf-Leonard K, editor. *Encyclopedia of social measurement*. Elsevier; 2005, p. 317-322. <https://doi.org/10.1016/B0-12-369398-5/00095-5>
26. Richesson RL, Smerek MM, Cameron CB. A framework to support the sharing and reuse of computable phenotype definitions across health care delivery and clinical research applications. *eGEMs*. 2016 Jul 5;4(3):1232. <https://doi.org/10.13063/2327-9214.1232>
27. Calvert M, Shankar A, McManus RJ, Lester H, Freemantle N. Effect of the quality and outcomes framework on diabetes care in the United Kingdom: retrospective cohort study. *BMJ*. 2009 May 27;338. <https://doi.org/10.1136/bmj.b1870>
28. Thayer DS, Mumtaz S, Elmessary MA, Scanlon I, Zinnurov A, Coldea AI, Scanlon J, Chapman M, Curcin V, John A, DelPozo-Banos M, Davies H, Karwath A, Gkoutos GV, Fitzpatrick NK, Quint JK, Varma S, Milner C, Oliveira C, Parkinson H, Denaxas S, Hemingway H, Jefferson E. Creating a next-generation phenotype library: the health data research UK Phenotype Library. *JAMIA Open*. 2024 Jun 17;7(2):ooae049. <https://doi.org/10.1093/jamiaopen/ooae049>

29. Mohammadi, Campbell SC, Sturkenboom MCJM, Vaz A. A Systematic Review to Summarize and Critically Appraise Existing Phenotype Libraries Using Electronic Health Records. *Pharmacoepidemiology and Drug Safety* 2026;35(5);e70378 <https://doi.org/10.1002/pds.70378>
30. Feder SL. Data quality in electronic health records research: quality domains and assessment methods. *Western journal of nursing research*. 2018 May;40(5):753-66. <https://doi.org/10.1177/0193945916689084>
31. Chapman M, Mumtaz S, Rasmussen LV, Karwath A, Gkoutos GV, Gao C, Thayer D, Pacheco JA, Parkinson H, Richesson RL, Jefferson E. Desiderata for the development of next-generation electronic health record phenotype libraries. *Gigascience*. 2021 Sep;10(9):giab059. <https://doi.org/10.1093/gigascience/giab059>
32. Grimes DA. Epidemiologic research using administrative databases: garbage in, garbage out. *Obstetrics & Gynecology*. 2010 Nov 1;116(5):1018-9. <https://doi.org/10.1097/AOG.0b013e3181f98300>
33. Davies AH, Ahmed H, Thomas-Wood T, Wood F. Primary care health professionals' approach to clinical coding: a qualitative interview study. *British Journal of General Practice*. 2024 Jul 29. <https://doi.org/10.3399/BJGP.2024.0036>
34. Osterhage KP, Hser YI, Mooney LJ, Sherman S, Saxon AJ, Ledgerwood M, Holtzer CC, Gehring MA, Clingan SE, Curtis ME, Baldwin LM. Identifying patients with opioid use disorder using international classification of diseases (Icd) codes: challenges and opportunities. *Addiction*. 2024 Jan;119(1):160-8. <https://doi.org/10.1111/add.16338>
35. Innes GK, Bhondokhan F, Lau B, Gross AL, Ng DK, Abraham AG. The measurement error elephant in the room: challenges and solutions to measurement error in epidemiology. *Epidemiologic reviews*. 2021;43(1):94-105. <https://doi.org/10.1093/epirev/mxab011>
36. Gillam S, Steel N. The quality and outcomes framework—where next?. *BMJ*. 2013 Feb 7;346. <https://doi.org/10.1136/bmj.f659>
37. Langdown C, Peckham S. The use of financial incentives to help improve health outcomes: is the quality and outcomes framework fit for purpose? A systematic review. *Journal of Public Health*. 2014 Jun 1;36(2):251-8. <https://doi.org/10.1093/pubmed/ftd077>

Abbreviations

EHRs:	Electronic Healthcare Records
ICD:	International Classification of Diseases
HDR UK:	Health Data Research UK
NHS:	National Health Service (UK)
SNOMED	Systematised Nomenclature of Medicine
CT:	Clinical Terms

