Social Psychiatry and Psychiatric Epidemiology (2019) 54:1161–1175
https://doi.org/10.1007/s00127-019-01744-4

REVIEW

Case-finding for common mental disorders in primary care using routinely collected data: a systematic review

Harriet Larvin[1] · Emily Peckham[1] · Stephanie L. Prady[1]

Received: 31 January 2019 / Accepted: 24 June 2019 / Published online: 12 July 2019

Abstract

Purpose Case-finding for common mental disorders (CMD) in routine data unobtrusively identifies patients for mental health research. There is absence of a review of studies examining CMD-case-finding accuracy in routine primary care data. CMD-case definitions include diagnostic/prescription codes, signs/symptoms, and free text within electronic health records. This systematic review assesses evidence for case-finding accuracy of CMD-case definitions compared to reference standards.

Methods PRISMA-DTA checklist guided review. Eligibility criteria were outlined prior to study search; studies compared CMD-case definitions in routine primary care data to diagnostic interviews, screening instruments, or clinician judgement. Studies were quality assessed using QUADAS-2.

Results Fourteen studies were included, and most were at high risk of bias. Nine studies examined depressive disorders and seven utilised diagnostic interviews as reference standards. Receiver operating characteristic (ROC) planes illustrated overall variable case-finding accuracy across case definitions, quantified by Youden's index. Forest plots demonstrated most case definitions provide high specificity.

Conclusion Case definitions effectively identify cases in a population with good accuracy and few false positives. For 100 anxiety cases, identified using diagnostic codes, between 12 and 20 will be false positives; 0–47 cases will be missed. Sensitivity is more variable and specificity is higher in depressive cases; for 100 cases identified using diagnostic codes, between 0 and 87 will be false positives; 4–18 cases will be missed. Incorporating context to case definitions may improve overall case-finding accuracy. Further research is required for meta-analysis and robust conclusions.

Keywords Systematic review · Electronic health records · Anxiety · Depression · Adults

Introduction

Internationally, it is estimated that one in five people meet criteria for anxiety or depressive disorders (common mental disorders, CMD) [1]. Depression is the leading cause for global disability, and anxiety disorders are within the top 10 [1]. Research into the causes and consequences of CMD, and the effects of interventions, requires accurate case ascertainment for study recruitment [2, 3]. The high costs and participant burden associated with diagnostic interviewing and follow-up make unobtrusive identification using routinely collected data attractive [4]. There are also financial and resource benefits to recruitment using automated algorithms compared to manually identifying participants [5].

Most healthcare systems in the developed world make at least partial use of electronic data [1]. Data contained in electronic health records (EHR) typically comprise of a problem list detailing clinically important diagnoses and concerns, treatments including prescriptions, referrals, and other relevant encounter details. Structured coding systems allow for efficient searching and record retrieval and may be accessed by health researchers through a variety of ways. For example in the UK, some primary care research databases can be accessed by accredited researchers for a fee [6, 7] or care providers may be approached directly by researchers for data use [8].

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00127-019-01744-4) contains supplementary material, which is available to authorized users.

✉ Harriet Larvin
harrietlarvin@gmail.com

[1] Department of Health Sciences, The University of York, Seebohm Rowntree Building, Heslington, York YO10 5DD, UK

Mental health researchers may use EHR data to sample individuals for trial recruitment, identify and match cases and controls for observational studies or follow participants' progress and outcomes [9]. The effective use of these data in research is heavily reliant on accurately identifying patients using markers of a disorder, otherwise known as case-finding.

CMD-case definitions in EHR include current codes relating to specific depressive and anxiety diagnoses, signs or symptoms [10]. Researchers can also choose to interrogate treatment codes or mental health referrals, codes for antidepressant or anxiolytic prescriptions or codes indicating an historical depressive or anxiety observation [11, 12].

There are some pitfalls to re-purposing primary care EHR data which may limit its effectiveness as a data source for mental health research. Poor EHR uniformity and maintenance can reduce reliability and primary care practitioners (PCP) rates of depression diagnosis are usually lower than rates examined in epidemiological studies [11], as they do not usually record codes with research purposes in mind [13]. Diagnostic coding can also differ significantly between clinicians and practices over time, making the identification of patients using a specific case definition more difficult [10, 14]. Free text within EHR may be extensively used for clinical management, but is rarely available to researchers due to confidentiality concerns. Free-text extraction can also be difficult given variations in terminology and writing style—such data are unedited and often hastily written [15]. Stigma attached to mental disorder diagnoses can also prevent coding of free text [9]. This is problematic for research purposes when free text contains relevant information that is not otherwise coded in the record, such as signs, symptoms, or management plans [16].

Given concerns related to the re-purposing of data, it is important to understand the accuracy of case-finding CMD within EHR. Comparison to a reference standard ascertains case definition accuracy [17]; in mental health, such standards could be diagnostic interviews, screening instruments, or clinician judgement. Reviews of studies that compared routinely recorded case definitions against reference standards have indicated acceptable accuracy in secondary care settings [18], but there are no such reviews of accuracy within primary care.

The aim of this study was to systematically review studies that utilise case definitions for identifying CMD within routinely collected primary care data and independently verify the presence or absence of CMD against a reference standard. The findings of this review will inform the selection of CMD-case definitions for accurate case-finding in routine primary care data in mental health research.

Method

The review design and report follow the Preferred Reporting Items for Systematic Reviews and Meta-analyses extension for Diagnostic Test Accuracy (PRISMA-DTA) guidelines [19].

Eligibility

For inclusion, the study had to be set within an OECD state as of July 2018 and examine an adult population, or results of adults reported separately. CMD identified in papers comprised diagnostic sub-categories of depressive disorders, such as major depressive disorder and dysthymia or anxiety disorders, including generalised anxiety disorder and post-traumatic stress disorder, or both, as defined by WHO [1]. We excluded papers investigating severe mental disorders such as schizophrenia and bipolar disorder. Case-finding confirmation by reference standard (diagnostic interviews, screening instruments, or clinician judgement) was required to be within 1 year either side of the baseline, where this was not clear the study was considered at risk of bias. Studies had to examine registers managed by PCP and identify CMD in routinely recorded databases such as EHR and insurance claims data. Due to resource constraints, only studies published in English were reviewed. Exclusion criteria are outlined in Fig. 1.

Search strategy

Searches were carried out in MEDLINE, CINAHL, Cochrane and PsycINFO databases between 5th July 2018 and 6th July 2018. The search was not limited by publication date. The search string was a hybrid of the previous similar systematic review searches comprising methodological, case-finding index and condition terms, plus MeSH headings, subject keywords, synonyms, alternate phrasing and necessary adaptations depending on database to prevent overlooking relevant studies [20, 21].

Components of the search string were organised by (a) conditions of interest (CMD), (b) data source, (c) reference standard, and (d) methodological terms in the combination: (a and b) and (c or d). This algorithm considers CMD classification and medium of interest to be essential to appropriate publication search [20], see Supplement 1 for the search strategy.

Study identification

Search results were imported into Endnote [22] and duplicates removed. Titles and abstracts were screened against

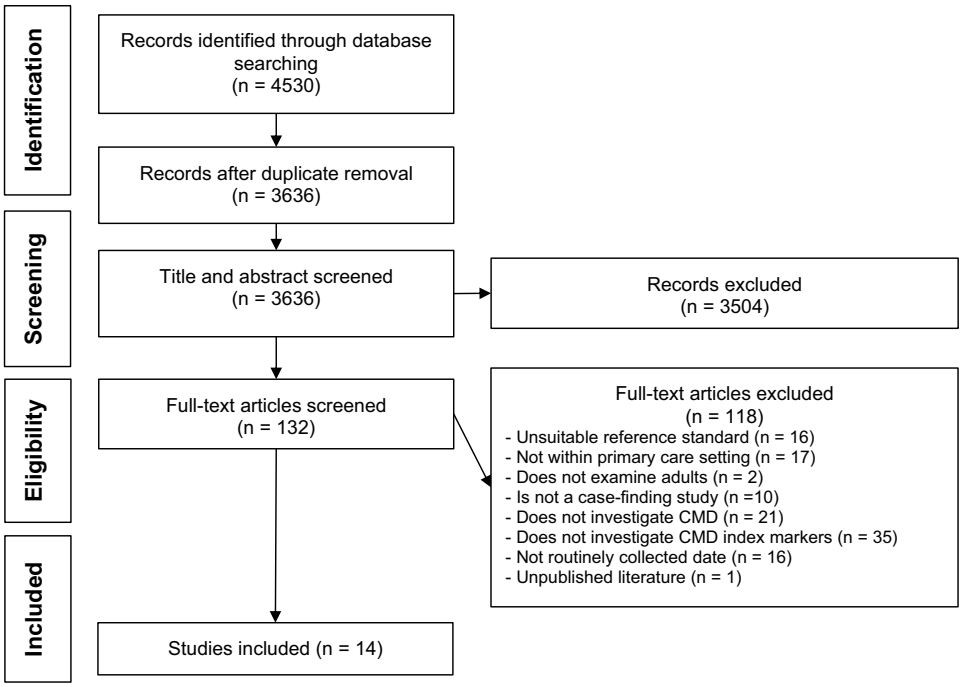

Fig. 1 Study selection flow chart

the inclusion and exclusion criteria and full texts of potentially eligible studies were screened by one author (HL). The reference lists from studies meeting all eligibility criteria were searched for additional potentially eligible studies.

Data extraction

One author (HL) undertook data extraction using a data extraction form developed for this study. Data were collected on: author, year, and country of study, number of patient entries, CMD sub-category, patient population demographics, details of case definition and reference standard and outcomes of study including: true positive, true negative, false positive and false negative values, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and Youden's index (YI). In cases where the required information could not be calculated from the presented data, primary authors were contacted once by email where possible for the missing information.

Quality assessment

The Quality Assessment of Diagnostic Accuracy Studies II (QUADAS-2) tool [23] measures risk of bias and was used to assess quality of included studies. Following quality assessment, an overall risk of bias rating was determined. Studies were classified with "high risk of bias", where one or more domain was categorised as high/unclear risk of bias.

Narrative synthesis

The review narrative first summarises quality assessment results. Following the overview of study characteristics, case-finding accuracy was presented by case definition investigated by the study: codes for diagnosis and symptoms, prescription codes, free text, and their combinations. ROC planes graphically display these findings to indicate overall case-finding accuracy [17]. Diagnostic codes refer to coded data describing diagnosis or defined problem, symptom, or sign.

Where there was more than one-case definition examining more than one disorder, studies are grouped by diagnostic sub-category and ordered by reference standard. Studies of case–control design [indicated by (*)] are not applicable for PPV/NPV reporting as representative incidence and prevalence cannot be determined when number of cases is contrived to the number of controls [24]. Where sensitivity, specificity, or YI was not reported or could not be calculated, studies are examined by available case-finding accuracy information.

Data synthesis

We examine the data initially by contrasting sensitivity and specificity. Sensitivity examines the proportion of true positive cases for CMD found by a case definition out of the number of positive cases identified by the reference standard; specificity determines the proportion of true negative CMD cases among those identified as negative by the reference standard. Sensitivity or specificity are considered

high when values exceed 0.67; values below 0.30 represent limited sensitivity or specificity [25]. YI provides a global measure of case definition performance and power in a single statistic, by combining true positive (sensitivity) and true negative (specificity) rates; an accurate marker is close to 1, while a poor marker has YI of closer to 0.

The most accurate case definitions are located in the top left quarter of a receiver operating curve (ROC) plane, where sensitivity and specificity are closest to 1. Forest plots demonstrate paired sensitivity and specificity of case definitions. Tabulations and calculations were prepared using Microsoft Excel and figures using MetaDiSc [26].

Study Selection

4530 papers were retrieved in the search. Following duplicate removal ($n=326$), title and abstract screening ($n=3636$) revealed 132 studies eligible for full text screening. Fourteen studies met eligibility criteria (Fig. 1) [28–41]. A meta-analysis was planned providing at least 20 studies were identified, with more than 10 of the studies being at low risk of bias according to QUADAS and with minimal clinical and methodological heterogeneity across extracted information [27]. Meta-analysis was not conducted due to insufficient quantity ($n\leq 20$) and large proportion of studies at high risk of bias ($n=10$; 71%).

Quality assessment

Four studies (29%) were at low risk of bias across all domains [30, 33, 39, 40] and the remaining studies (71%) were at risk of bias in one or more domains (S2.1). Of the studies at high risk of bias, the most common domain at risk was flow and timing; in all cases, this was due to unclear interval of time between case-finding index and reference standard baseline [31, 33–37, 43].

Overview of included studies

Table 1 summarises data extraction. All studies were published between 1994 and 2016 and were from USA ($n=6$), UK ($n=3$), The Netherlands ($n=2$), Canada ($n=1$), Spain ($n=1$), and Sweden ($n=1$). There was large variation in population demographics. For example, two studies examined older populations [37, 38], while another investigated patients aged 25–65 years [31]. One study investigated a population of mostly male veterans [32] and another a population with 63% women [30]. There were some absences of information for comprehensive population demographics in many studies, e.g., absent gender data [29, 31, 37].

Of 14 included studies, eight studies (57%) examined depressive disorders [28, 31, 35–38, 40, 41] and one additionally investigated both anxiety and depressive disorders [34]. The remaining five studies examined anxiety disorders only [29, 30, 32, 33, 39, 42]. All included papers examined accuracy of codes incorporated within algorithms; therefore, case definition accuracy can only be interpreted as codes within algorithms rather than accuracy of individual codes. Seven studies compared case definitions to a reference standard of diagnostic interview [29, 30, 33, 35, 37, 39, 41] and five compared to results from self-reported questionnaires [31, 32, 34, 36, 38]. Two studies examined case-finding accuracy compared to researcher and clinician reviewed EHR [28, 36] and one utilised a questionnaire completed by a physician [40]. Four studies were of case–control design and, therefore, increased risk of bias [28, 30, 32, 33].

Study design and reporting

We contacted four authors for raw data to populate contingency tables; none responded [28, 32, 33, 40]. Most studies investigated diagnostic codes only ($n=10$) [28, 29, 31–34, 37, 38, 40, 41], four investigated prescription codes only [34–36, 40], and one free text with codes [30]. Five studies examined combinations within case definitions [30, 34–36, 40]. Studies are grouped more than once if they examined more than one-case definition.

ROC planes are grouped by type of case definition. Forest plots illustrate possible variation in sensitivity and specificity by CMD and reference standard utilised. Study groupings are outlined in Table 2. One-case definition examined diagnostic and prescription codes with free text combined and is, therefore, not illustrated graphically [26].

Diagnostic codes only

Eight of the ten studies examining diagnostic codes were at high risk of bias [28, 29, 31, 32, 34, 37, 38, 41] and two were at low risk of bias [33, 40].

Figure 2a illustrates overall variable case-finding accuracy of diagnostic codes. Most points follow closely to the line of no effect; however, there are three case definitions located in the top left quartile indicate high sensitivity and specificity: two from Elhai et al. and one from van Weel-Baumgarten et al. [29, 41].

The King alteration of DSM-IV classification for PTSD had the highest YI of 0.88; the Simms alteration also resulted in high YI of 0.79 [29] (S2.2). These case definitions were investigated by Elhai et al. (2009) examining DSM-IV alterations to consider additional features of PTSD absent from current classification compared to clinical interview. Case definitions included ICD-9 and CPT codes for separation of avoidance and numbing symptoms of DSM-IV classification in King's alteration and combining hyperarousal and numbing signs in Simms. It should be noted that these case definitions compares the effect on cohorts to diagnostic criteria

Table 1 Summary of included studies

Author (year), country	Number of patients	Demographics and selection	Diagnostic sub-category	Patient database	Case definition	Case definition interpretation	Reference standard	Reference standard interpretation	Interval between case-finding and reference standard
Alaghehbandan et al. (2012)[a], Canada [28]	253	Patients aged 18+ with diagnosis of depressive disorder. Around 70% male. Recruited from 3 family practices in Newfoundland and Labrador. Data collected between January and July 2007. Bipolar disorders excluded. Controls were matched by age/gender	Depressive disorder	Administrative databases; Medical care plan data.	5 case definitions developed using ICD 9/10 edition codes. (> 1 hospital or > 2 PSY visits within 1 year; > 1 hospital or > 2 PSY visits within 2 years; (> 1 hospital or > 1 PSY visits within 1 year) or > 2 PCP within 1 year; (> 1 hospital or > 1 PSY) or > 2 PCP within 2 years; (> 1 hospital or > 1 PSY) or > 3 PCP within 3 years	Variables for development: diagnoses, date, service provider type	Diagnosis in EHR reviewed by researchers	Excluded cases were reviewed by psychiatrist	Unclear
Elhai et al. (2009), USA [29]	5692	Adults and adolescent samples separated. Of 5692 patients in sample, only those with traumatic event comprising initial fear/helplessness/horror criterion were queried for remaining DSM-IV PTSD symptoms	Anxiety disorder—PTSD	National Comorbidity Survey Replication	Current DSM-IV classification of PTSD plus King, Simms extensions of DSM-IV PTSD models compared; all to have symptoms A1 and A2, E and F. King: at least 1 re-experiencing symptom (B1–B5); at least 2 hyperarousal symptoms (D1–D5) + at least 1 avoidance symptom C1–C2. Simms: at least 1 re-experiencing symptom (B1–B5); at least 1 avoidance symptom C1–C2; at least 1 hyperarousal symptom (D4–D5) + at least 3 dysphoria symptoms (C3–C7, D1–D3)	King model separates criterion C. Simms model 3 hyperarousal symptoms combined with emotional numbing factor symptoms to form 'dysphoria'	CIDI: at least 1 re-experiencing symptom (B1–B5); at least 3 avoidance symptoms (C1–C7) + at least 2 hyperarousal symptoms (d1–D5)	Interviewer: unclear training and inter-rater independence	Unclear

Table 1 (continued)

Author (year), country	Number of patients	Demographics and selection	Diagnostic sub-category	Patient database	Case definition	Case definition interpretation	Reference standard	Reference standard interpretation	Interval between case-finding and reference standard
Fernández et al. (2012), Spain[a] [30]	3815	Mean age = 54.3 years. 63% females. Patients from 77 primary care centres in Catalonia, randomly selected and invited to join the study	Anxiety disorder	Diagnosis and treatment of mental disorders study	Codes ICPC-II: p01, p02, p74, p79. CIE-9: 300.0 and 300.00, 300.01, 300.02, 300.09, 300.2 and 300.20, 300.22, 300.23, 300.29, 300.3. CIE-10: F40, F40.0, F40.1, F40.2, F40.8, F41, F41.0, F41.1, F41.2, F41.3, F41.8, F41.9, F42. Also mention of anxiety disorder/ anxiety symptoms in EHR not coded	Previous 12 month EHR examined and extracted by blinded interviewers into dichotomous variable	Structured Clinical Interview for DSM, face-to-face	Trained clinical psychologists	1 year
Flyckt et al. (2014), Sweden [31]	90	Adults, sampled from a doctors waiting room in a wealthy catchment area with 50% pop. between 25 nd 65. Age and gender matched. All participants must have had primary care contact within the past year	Depressive disorder	EHR from primary practices	One or more of the following 'cues'. Signs of depression: note in medical record by PCP that described 1/9 depressive criteria. Also included number of physician-rated signs/symptoms, e.g., tearfulness and excessive worrying. Also decreased functioning (low GAF score). A case was exhibiting one or more of these cues	Experienced psychiatrist analysed the cues and signs retrieved in EHR	Montgomery-Asberg Rating Scale score > 12 and diagnostic interview (as determined by interview in waiting room)	Not reported	1 week
Gravely et al. (2011), USA[a] [32]	4777	Veterans, mostly male and middle aged (45-64). At least 1 new PTSD diagnosis, recruited over 30 weeks	Anxiety disorder—PTSD	Veterans Administration administrative data from National Patient Care Database	ICD-9 codes of one PTSD diagnosis (309.81), verses at least 2 PTSD diagnoses	Second PTSD diagnosis found within 4 months of first	PTSD checklist score > 50	PTSD checklist self-reported through national survey	1 year

Table 1 (continued)

Author (year), country	Number of patients	Demographics and selection	Diagnostic sub-category	Patient database	Case definition	Case definition interpretation	Reference standard	Reference standard interpretation	Interval between case-finding and reference standard
Holowka et al. (2014), USA[a] [33]	1649	Iraq/Afghanistan veterans, average 37.5 years of age and 50% men; 50% women. Random sample, participants with PTSD (presence of 2 PTSD diagnosis in ICD-9: code 309.81) and without diagnosis (3:1). Consecutive sampling to acquire necessary numbers	Anxiety disorder—PTSD	Project VALOR National patient care database	PTSD status by ICD-9 diagnoses in Encounter (services provided for condition) records and Patient Problem list (codes for diagnoses). Codes examined as current and within lifetime	Indicators of PTSD in problem list and encounter data abstracted by trained research assistants	Structured clinical interview for DSM over telephone	Trained doctoral-level diagnosticians. Blind to diagnostic status and interrater reliability examined by random subsample	1 year
John et al. (2016), UK [34]	2799	Welsh between 18 and 74. Baseline survey sent to population	Anxiety and depressive disorders collectively	General practice database at Swansea University	12 algorithms for Read codes version 2; current and historical (additional files) plus drug treatment	In waves (wave one = baseline survey); wave two = follow-up postal survey	5-item MHI	Self-reported questionnaire	Unclear
Joling et al. (2011), The Netherlands [35]	816	Aged between 18 and 65. Screening questionnaires sent to random sample who consecutively consulted Primary care practitioner for 4 months; screen positives followed up with CIDI	Depressive disorder	The Netherlands Study of Depression and Anxiety	ICPC codes, medication data (anatomical therapeutic chemical classification), referral data (working committee for information and automation) codes and fee text in EHR	2 scorers for 36 cases before total agreement across all cases	CIDI by telephone	Interviewers blind to diagnosis	Unclear
McGregor et al. (2010), UK [36]	168	Patients of 5 general practices in Swansea. 10% random sample of eligible patients	Depressive disorder—at least moderate to severe	Secure anonymised information linkage database	Algorithm of inclusion/exclusion criteria. Recent AD therapy, diagnosis of moderate to severe depression in medical history. Plus specific study exclusion criteria	Not reported	Clinical diagnosis in EHR	Psychiatrist judgement	Unclear

Table 1 (continued)

Author (year), country	Number of patients	Demographics and selection	Diagnostic sub-category	Patient database	Case definition	Case definition interpretation	Reference standard	Reference standard interpretation	Interval between case-finding and reference standard
Mullan et al. (1994), UK [37]	186	Aged 65 years plus. Consecutive attenders to the practice	Depressive disorder	Lower Clapton Health Centre	Primary care depression diagnosis in EHR	EHR traced for info on: PCP detected depression, PCP recorded depressive symptom, currently on AD, past history of depression and past history of AD	15-item GDS by brief interview with researcher	By psychiatrist blind to 15-item GDS data	Unclear
Noyes et al. (2011), USA [38]	1551	19 counties, in New York, West Virginia and Ohio states. A random sample of elderly primary care patients was taken. Average age 77 years; more than 66% female and < 4% non-white	Depressive disorder	Medicare claims database	Base was ICD-9—clinical modification codes 296.20-24 and 296.30-34 (depression). Extended to dysthymic, adjustment disorders with depressed mood, depressive order not elsewhere classified (300.4, 309.0 and 311)	Not reported	2 self-reported depression scales. MINI and GDS	No interviewer judgement required for these scales. Administration of reference standard was baseline	1 year before and after baseline
Shear et al. (2000), USA [39]	164	Aged between 18 and 65, urban and rural clinics. 68% female. Consecutive recruitment for non-psychotic patients who were seen at a rural or urban community treatment facility	Mood, anxiety or adjustment disorders	Two community health facilities, Pennsylvania	Primary diagnosis in EHR	Primary diagnosis, demographics and insurance coverage	SCID	One of 2 researchers lead interview (non-physician), reviewed by SCID trainer. Unresolved questions or disagreements discussed at weekly visits from supervisor and resolved	3 months
Trinh et al. (2011), USA [40]	82	Patients selected from representative group of primary care practices	Depressive disorder	Research Patient Data Registry	EHR field codes. Depression in billing diagnosis, depression in problem list and antidepressant in medication list. Compared with combinations	EHR fields tested against PCP assessment of survey	PCP survey	PCP assessment posted	Unclear

Table 1 (continued)

Author (year), country	Number of patients	Demographics and selection	Diagnostic sub-category	Patient database	Case definition	Case definition interpretation	Reference standard	Reference standard interpretation	Interval between case-finding and reference standard
van Weel-Baumgarten et al. (2000), The Netherlands [41]	65	Average age 46 years. Randomly selected from of university general practice. Must be 18 + and capable of communication	Depressive disorder	Continuous morbidity registry	Diagnostic codes for depression	ICPC criteria. Code could be first diagnosis or new episode of depression. Data processed anonymously	CIDI.	Index was blinded by interviewer	1 year

AD antidepressants, *CIDI* composite international diagnostic interview, *CIE* Clasificacion Internacional de Enfermedades, *DSM* diagnostic and statistical manual for mental disorders, *EHR* electronic health records, *GAF* global assessment for functioning, *ICD* international classification of diseases, *GDS* geriatric depression scale, *ICPC* international classification of primary care, *MHE* mental health inventory, *MINI* mini-international neuropsychiatric interview, *PCP* primary care physician, *PSY* physician, *PTSD* post-traumatic stress disorder, *SCID* structured clinical interview for diagnostic and statistical manual for mental disorders

[a]Case–control design

alterations and not case-finding accuracy using current diagnostic criteria. Current DSM-IV classification of PTSD was also compared to reference standard in this study; case-finding accuracy was significantly lower with YI of 0.38.

The case definition with the lowest case-finding accuracy in this group, illustrated by YI of 0.05 and 0.04 were examined by John et al. incorporating Read codes for current diagnosis plus a range of signs and symptoms of anxiety or depressive disorders—treated or untreated in two waves, respectively. These codes encompass case definitions for both anxiety and depressive disorders and utilised a self-reported questionnaire as comparison [34].

Sensitivity and specificity of diagnostic codes was variable when grouped by disorder sub-category (Supplement 3). Specificity was high across most diagnostic codes ranging from 0.82 (95% CI 0.68, 0.92) to 1.00 (1.00, 1.00). The most sensitive diagnostic codes were for anxiety disorders (PTSD), ranging from 0.80 (0.75, 0.83) to 0.88 (0.84, 0.91). Sensitivity was low in depressive disorders, not exceeding 0.38 (0.24, 53). Case definitions using ICPC codes for diagnosis or episode of depression compared to clinical interview as reference standard in the van Weel-Baumgarten's (2000) study exhibited considerably higher sensitivity than other studies of depressive disorders at 1.00 (0.87, 1.00). Diagnostic codes for anxiety and depressive disorders combined demonstrated low sensitivity, not exceeding 0.06 (0.05, 0.08).

Prescription codes only

Three of the four studies examining prescription codes only were at high risk of bias [35–37]; one was at low risk of bias [40].

Figure 2b illustrates overall variable case-finding accuracy of prescription codes. All points are above or on the line of no effect.

The most accurate case definition with YI of 0.44 was ICD-9 codes for antidepressant prescriptions in medication list for depressive disorders compared to a physician questionnaire (S2.3); however, absence in contingency table data hinders estimates of overall accuracy for this case definition [40]. Joling et al. investigated ICPC codes for antidepressants in current medication lists and in EHR history as case definition for depressive disorders giving a similar YI value of 0.41 [35].

The least accurate case definition in this group illustrated YI of 0.0 [37]. This study by Mullan et al. (1994) investigated current and historical prescriptions for antidepressants as case definitions for depressive disorders in an older population compared to brief diagnostic interview.

Sensitivity was variable and specificity moderately high for prescription codes when grouped by disorder sub-category (Supplement 3). Specificity did not fall below 0.72

Table 2 Summary of paper groupings

Author	Case definition type	Diagnostic sub-category	Reference standard
Holowka et al.[a] [33]	Diagnostic codes only	Anxiety disorder	Diagnostic interview
Elhai et al. [29]			
Gravely et al.[a] [32]			Self-reported questionnaire
van Weel-Baumgarten et al. [41]		Depressive disorder	Diagnostic interview
Mullan et al. [37]			
Alaghehbandan et al.[a] [28]			EHR review
Flyckt et al. [31]			Self-reported questionnaire
Noyes et al. [38]			
Trinh et al. [40]			Physician questionnaire
John et al. [34]		Anxiety and depressive disorders	Self-reported questionnaire
Joling et al. [35]	Prescription codes only	Depressive disorder	Diagnostic interview
Mullan et al. [37]			
Trinh et al. [40]			Physician questionnaire
John et al. [34]		Anxiety and depressive disorders	Self-reported questionnaire
Shear et al. [39]	Free text only	Depressive disorder anxiety disorder	Diagnostic interview
McGregor et al. [36]	Combined: diagnostic and prescription codes	Depressive disorder	Clinician judgement
Trinh et al. [40]			Physician questionnaire
John et al. [34]		Anxiety and depressive disorders	Self-reported questionnaire
Fernández et al.[a] [30]	Combined: diagnostic codes and free text	Anxiety Disorder	Diagnostic interview
Joling et al. [35]	Combined: diagnostic and prescription codes, plus free text	Depressive disorder	Diagnostic interview

CMD common mental disorder, *EHR* electronic health records, *PTSD* post-traumatic stress disorder

[a]Case–control design

(95% CI 0.68, 0.75). Sensitivity was variable across in prescription codes for depressive disorders and anxiety/depressive disorder combined. This ranged from 0.14 (0.07, 0.25) to 0.69 (0.63, 0.75) in depressive disorders and 0.33 (0.29, 0.36) to 0.48 (0.43, 0.53) in anxiety/depressive disorders combined. There were insufficient data to examine prescription codes as a case definition for anxiety disorders.

Free text only

The single study by Shear et al. examining free text only for anxiety and depressive disorders was at low risk of bias [39]. Accuracy of free-text primary diagnosis recorded by PCP as case definitions was determined for separate anxiety and depressive disorder diagnoses compared to clinical interview.

Figure 2c illustrates overall variable accuracy. One-case definition is located in the top left quarter, indicating high case-finding accuracy. This case definition identified depressive disorder as primary diagnosis by PCP.

Comparison of YI values in illustrates accuracy of free text was higher in depressive disorders with 0.35, compared to 0.13 in anxiety disorders (S2.4).

There was variable sensitivity and moderately high specificity of free text as case definition (Supplement 3). Sensitivity ranged from 0.20 (95% CI 0.12, 0.29) to 0.58 (0.47, 0.68) in. Specificity was highest in the case definition for identifying anxiety disorders at 0.94 (0.85, 0.98).

Combined: diagnostic and prescription codes

Two studies examining diagnostic and prescription codes combined were at high risk of bias [34, 36] and the third was at low risk [40].

Figure 2d illustrates overall variable case-finding accuracy in using diagnostic and prescription codes to identify CMD. Most points do not follow closely to the line of no effect indicating minimal threshold effect. Two points are located in the top left quartile suggesting high accuracy. These case definitions are reported by McGregor et al. and developed in a trial recruitment context, comprising Read codes for antidepressant prescription, lifetime depression diagnosis; exclusion Read codes: folate deficiency, pregnant, taking Lithium/anticonvulsants and life expectancy less than 1 year; clinician 1 judgement as reference standard; Read codes for antidepressant prescription, lifetime depression diagnosis; exclusion Read codes: folate deficiency,

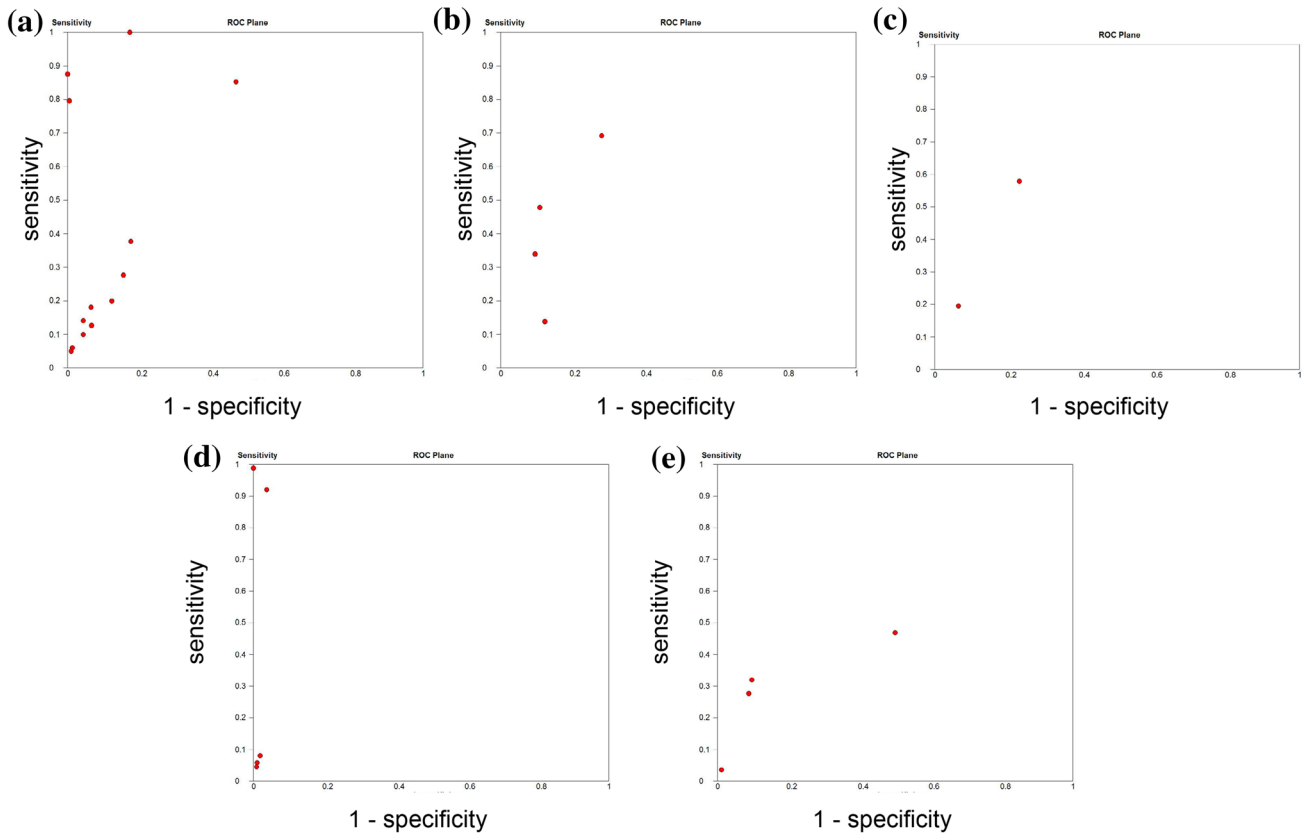

Fig. 2 ROC planes of case definitions. **a** Diagnostic codes, **b** prescription codes, **c** free text, **d** diagnostic and prescription codes combined, **e** diagnostic codes and free text combined. Grey trendline represents 45° line of no effect. *ROC* receiver operating characteristic

pregnancy and life expectancy less than 1 year; clinician 2 judgement as reference standard [36].

Comparison of YI values illustrate case definitions in McGregor et al. as the most accurate diagnostic codes with YI of 0.99 and 0.88, respectively [36]. Low-case-finding accuracy was illustrated in case definitions for anxiety and depressive disorders combined in John et al., as illustrated by YI not exceeding 0.04 [34] (S2.5).

Supplement 3 shows variation in sensitivity and specificity of diagnostic and prescription codes combined when grouped by disorder sub-category. Specificity was high across case definitions, YI at least 0.96 (95% CI 0.89, 0.99). Case definitions of at least moderate severity depression used by McGregor et al. (2010) were more sensitive ranging from 0.92 (0.84, 0.97) to 0.99 (0.87, 1.00) [36], than the algorithm used by John et al. for depression and/or anxiety disorders, where sensitivity did not exceed 0.38 (0.24, 53) [34].

Combined: diagnostic codes and free text

A single study by Fernández et al. examined diagnostic codes (ICPC and CIE codes) combined with free text for case-finding anxiety disorders and anxiety with or without symptoms of depression. This study compared case-finding accuracy to diagnostic interview and was at low risk of bias [30].

Figure 2e illustrates low-case-finding accuracy with points following closely to the line of no effect.

YI values across these case definitions were low, and the highest was 0.25 for identifying anxiety associated with depression. The case definition for any anxiety disorder has similar YI value of 0.22. Case-finding accuracy when the case definition was for generalised anxiety disorder, YI was 0.02 (S2.6).

There was highly variable sensitivity and specificity in this group of case definitions. Sensitivity was low, ranging from 0.04 (95% CI 0.01, 0.09) to 0.47 (0.38, 0.56). Specificity of all case definitions were above 0.90 (0.89, 0.91), except for anxiety associated with depression which was moderate (0.50; 0.39, 0.61) (Supplement 3).

Combined: diagnostic and prescription codes, plus free text

Only Joling et al. examined case definitions comprising diagnostic and prescription codes plus free text. One-case definition was examined: ICPC codes for depressive disorder/depressive feelings, with antidepressant prescription, free text and mental health referral, compared to diagnostic interview as reference standard. This study was at high risk of bias [35].

Sensitivity and PPV of this marker were moderately low at 0.41 (95% CI 0.36, 0.45) and 0.65 (0.60, 0.71), respectively; PPV in this study increased compared to prescription codes alone which was 0.54 (0.49, 0.59). Specificity was high at 0.90 (0.83–0.96). YI illustrated moderately low-case-finding accuracy of 0.30 (S2.7).

Discussion

Summary of findings

Most of the fourteen studies included in the review were at risk of bias. ROC planes illustrated variable overall accuracy across case definitions, while forest plots indicated generally high specificity but variable sensitivity. Meta-analysis was not conducted due to variability in demographics, study design, and overall high risk of bias.

The most accurate case definition assessed in this review comprised diagnostic and prescription Read codes along with contextual trial exclusion criteria [36]. The least accurate case definitions appeared to be current antidepressant prescription for PTSD in an older population and ICD-9 codes for identifying depression in older age groups.

Combining diagnostic codes and free text and diagnostic plus prescription codes and free text appeared to have low-case-finding accuracy in the one study that examined it; however, free text combined with diagnostic and prescription codes marginally increased PPV compared to prescription codes alone in one study (35). Combining case definitions for anxiety and depressive disorders did not demonstrate markedly high case-finding accuracy.

Limitations

Only fourteen studies met our inclusion criteria. Searching grey literature databases may have increased quantity of included studies, but not impact [42]. Meta-analysis was not conducted as the requirements were not met. Most studies in the present review were at high risk of bias which also impedes reliability of findings.

Threshold effect occurs when a significant change is observed following a quantitative limit. Many case definitions incorporated only one type of marker (e.g., diagnostic or prescription codes or free text) compared to a reference standard which varied greatly in reliability and conduct. This restricts outlook of potential threshold effect [27] and contributes to significant heterogeneity across studies making direct comparisons difficult to interpret.

Limitations to case definition types examined in this review include undefined location within EHR (e.g., prescription or problem lists) and potentially unreliable free-text extraction due to terminology and contextual variations [10]. Many studies included antidepressant prescription codes as markers for CMD; this does not consider CMD patients who refuse or are unsuitable for treatment [43], or patients taking antidepressants for other conditions such as chronic pain. Where marker types have been combined within case definitions, e.g., diagnostic and prescription codes, the effects of 'AND' and 'OR' within definitions has not been explored. It is possible that these combinations may greatly differ in their case-finding accuracy. The case-finding accuracy of encounter information and psychiatric referrals as case definitions in primary care EHR has also not been explored.

While utilising EHR routine primary care data is an unobtrusive method for identifying cases for mental health research, a large proportion of CMD in primary care is undiagnosed [1]. Patients with CMD identified by case-finding may, therefore, not be representative of community cases. This may bias generalisability of findings from mental health research that use EHR.

There is an argument that case–control designs can overestimate test accuracy and should not be compared with cohort studies in diagnostic accuracy systematic reviews [3]. While including these studies in the present review may limit reliability of conclusions, case–control design studies are identified in the narrative and potential bias outlined.

Studies were screened by a single author, introducing potential bias; around 8% of studies may be missed by single screening [44]. Due to time and resource constraints grey literature was not examined contributing to publication bias risk. Studies published in languages other than English were also not explored.

Interpretation with existing literature

An existing review examining effectiveness of case-finding for COPD in primary care found notable heterogeneity across studies [45]. This was shared as a significant barrier in the present review and could be causative of much of the variability in accuracy across case definitions.

Davis, Sudlow, and Hotopf reviewed studies using routine secondary care data for case-finding a variety of psychiatric

diagnoses finding that case-finding markers for depressive disorders were more accurate than those for anxiety disorders [18]. Their findings do not reflect the findings of the present review in that case-finding of depression and anxiety disorders appear equally variable using routine data within a primary care setting. Factors such as setting and care type may influence the case-finding accuracy between disorder types.

Fiest et al. examined accuracy of ICD codes for identifying depression in administrative data and conclude case-finding accuracy is dependent on amount context provided by case definition [45]. In the present review, the study utilising a detailed case definition: trial recruitment criteria also appeared to have higher case-finding accuracy.

The present review indicates combining free text with diagnostic codes and diagnostic plus prescription codes only marginally improves case-finding accuracy compared to prescription codes alone. The previous research in non-CMD conditions indicates free text significantly augments case-finding accuracy [15, 46]. Accuracy of free-text mining in EHR may be dependent on disease which could explain the differences in findings of the present review.

In this review, lower case-finding accuracy for depressive disorders was observed in studies within older populations. Older age group patients can have higher prevalence of comorbidities which can influence CMD diagnoses and prevent accurate case-finding [47]. Examining studies by age ranges may demonstrate the impact age can have on CMD-case-finding.

Implications for practice

The predominantly high specificity of CMD-case definitions suggests they may be more useful for identifying CMD patients as cases in mental health research with marginally low levels of false positives. To identify true positives, it may be necessary to utilise further screening or diagnostic assessments to confirm CMD cases as sensitivity was not consistently high.

Most of the evidence in our review came from studies examining diagnostic codes only. The findings suggest that CMD-case-finding accuracy using diagnostic code algorithms may be influenced by disorder. Further research is required to examine the differences between disorder sub-categories.

Case definitions incorporating the context of the research purpose may improve case-finding accuracy. Researchers may wish to prioritise contextual markers in case-finding. For example, using case definitions to encompass trial eligibility or classify a specific disorder within CMD classification.

Our findings also suggest case-finding for CMD using case definitions combining codes with free text resulted in variable accuracy; however, the previous evidence indicates free text significantly improves case-finding accuracy [48]. Researchers may choose to caveat free text as a marker of CMD in future practice.

Recommendations for future research

Accuracy of contextual case definitions should be investigated further, so thresholds and optimal markers for CMD may be determined. Lower case-finding accuracy in studies examining older populations and differences in CMD manifestation for this age group [47] indicates results by age ranges may produce more reliable results. Improved quality case-finding studies using reliable reference standards and structured case definitions is key to improve clarity of findings and enable meta-analysis. Further studies examining the addition of free-text data to case-finding algorithms are needed to understand whether and how the exclusion of these data in research extracts impacts on the accuracy of coded data.

Routine primary care databases used in the present review may have variable accuracy [49]. Developments to improve concordance of EHR coding for mental health research will enhance study reliability and synthesis precision.

Conclusion

The lack of high-quality studies included in this review prevents robust conclusions; however, high specificity and low sensitivity across case definitions indicates case-finding routine primary care data could effectively distinguish non-CMD cases but lacks sufficient sensitivity to accurately identify CMD cases. Presently, in mental health research, CMD-case-finding may need additional screening tools or diagnostic assessments for confirmation.

Acknowledgements The authors would like to thank the reviewers for their helpful comments which considerably strengthened the review.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. World Health Organisation (2017) Depression and other common mental disorders. global health estimates. http://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf;jsessionid=EEF8521C5188C4D578E6B3A7A66B8581?sequence=1. Accessed 2 Aug 2018
2. Howe LD et al (2013) Loss to follow-up in cohort studies: bias in estimates of socioeconomic inequalities. Epidemiology 24(1):1–9
3. Smeeth L, Donnan PT, Cook DG (2006) The use of primary care databases: case-control and case-only designs. Fam Pract 23(5):597–604
4. Crossan C et al (2017) Cost effectiveness of case-finding strategies for primary prevention of cardiovascular disease: a modelling study. Br J Gen Pract 67(654):e67–e77
5. Callard F et al (2014) Developing a new model for patient recruitment in mental health services: a cohort study using Electronic Health Records. BMJ Open 4:e005654
6. National Institute for Health Research (2018) Clinical Practice Research Datalink. https://www.cprd.com/home/. Accessed 2 Aug 2018
7. ResearchOne (2018) Transforming data into knowledge for evidence-based care. http://www.researchone.org/. Accessed 2 Aug 2018
8. Foster JM et al (2015) Barriers and facilitators to patient recruitment to a cluster randomized controlled trial in primary care: lessons for future trials. BMC Med Res Methodol 15:18
9. Perera et al (2016) Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) case register: current status and recent enhancement of an electronic mental health record-derived data resource. BMJ Open 6:e008721
10. Rait G et al (2009) Recent trends in the incidence of recorded depression in primary care. Br J Psychiatry 195(6):520–524
11. Cornish RP et al (2016) Defining adolescent common mental disorders using electronic primary care data: a comparison with outcomes measured using the CIS-R. BMJ open 6(12):e013167
12. Christensen KS, Sokolowski I, Olesen F (2011) Case-finding and risk-group screening for depression in primary care. Scand J Prim Health Care 29(2):80–84
13. Wright A, Maloney FL, Feblowitz JC (2011) Clinician attitudes toward and use of electronic problem lists: a thematic analysis. BMC Med Inform Decis Mak 11:36
14. Gulliford MC et al (2009) Selection of medical diagnostic codes for analysis of electronic patient records. Application to stroke in a primary care database. PloS one 4(9):e7168
15. Ford E et al (2016) Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 23(5):1007–1015
16. Graber M (2013) The incidence of diagnostic error in medicine. BMJ Qual Safety 22:ii21–ii27
17. Centre for Reviews and Dissemination (2009) Systematic reviews: CRD"s guidance for undertaking reviews in health care, systematic reviews. University of York, York
18. Davis KAS, Sudlow CLM, Hotopf M (2016) Can mental health diagnoses in administrative data be used for research? A systematic review of the accuracy of routinely collected diagnoses. BMC Psychiatry 16:263
19. McInnes MDF et al (2018) Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA 319(4):388–396
20. Glanville J et al (2018) Diagnostic accuracy. https://vortal.htai.org/?q=node/339. Accessed 4 July 2018
21. Chapman D (2009) Health-related databases. J Can Acad Child Adolesc Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent 18(2):148–149
22. Clarivate Analytics (2017) Endnote (TM). https://www.myendnoteweb.com/. Accessed 4 July 2018
23. Whiting PF et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155(8):529–536
24. Sauerbrei W, Blettner M (2009) Interpreting results in 2 × 2 tables: part 9 of a series on evaluation of scientific publications. Deutsches Arzteblatt Int 106(48):795–800
25. Power M, Fell G, Wright M (2013) Principles for high-quality, high-value testing. Evid-based Med 18(1):5–10
26. Zamora J et al (2006) Meta-DiSc: a software for meta-analysis of test accuracy data. BMC Med Res Methodol 6:31
27. Takwoingi Y et al (2017) Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. Stat Methods Med Res 26(4):1896–1911
28. Alaghehbandan R et al (2012) Using administrative databases in the surveillance of depressive disorders–case definitions. Popul Health Manag 15(6):372–380
29. Elhai JD et al (2009) Diagnostic alterations for post-traumatic stress disorder: examining data from the National Comorbidity Survey Replication and National Survey of Adolescents. Psychol Med 39(12):1957–1966
30. Fernández A et al (2012) Recognition of anxiety disorders by the general practitioner: results from the DASMAP Study. Gen Hosp Psychiatry 34(3):227–233
31. Flyckt L et al (2014) Clinical cues for detection of people with undiscovered depression in primary health care: a case–control study. Primary Health Care Res Dev 15(3):324–330
32. Gravely AA et al (2011) Validity of PTSD diagnoses in VA administrative data: comparison of VA administrative PTSD diagnoses to self-reported PTSD Checklist scores. J Rehabil Res Dev 48(1):21–30
33. Holowka DW et al (2014) PTSD diagnostic validity in Veterans Affairs electronic records of Iraq and Afghanistan veterans. J Consult Clin Psychol 82(4):569–579
34. John A et al (2016) Case-finding for common mental disorders of anxiety and depression in primary care: an external validation of routinely collected data. BMC Med Inform Decis Mak 16(1):35
35. Joling KJ et al (2011) Do GPs' medical records demonstrate a good recognition of depression? A new perspective on case extraction. J Affect Disord 133(3):522–527
36. McGregor J et al (2010) The Health Informatics Trial Enhancement Project (HITE): using routinely collected primary care data to identify potential participants for a depression trial. Trials 11:39
37. Mullan E et al (1994) Screening, detection and management of depression in elderly primary care attenders. II: Detection and fitness for treatment: a case record study. Fam Pract 11(3):267–270
38. Noyes K et al (2011) Medicare beneficiaries with depression: comparing diagnoses in claims data with the results of screening. Psychiatric Serv 62(10):1159–1166
39. Shear MK et al (2000) Diagnosis of nonpsychotic patients in community clinics. Am J Psychiatry 157(4):581–587
40. Trinh N-HT et al (2011) Using electronic medical records to determine the diagnosis of clinical depression. Int J Med Inform 80(7):533–540
41. van Weel-Baumgarten EM et al (2000) The validity of the diagnosis of depression in general practice: is using criteria for diagnosis as a routine the answer? Br J Gen Pract 50(453):284–287

42. Hopewell S et al (2007) Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.MR000010.pub3
43. Keller MB (2004) Remission versus response: the new gold standard of antidepressant care. J Clin Psychiatry 65(Suppl 4):53–59
44. Beahler CC, Sundheim JJ, Trapp NI (2000) Information retrieval in systematic reviews: challenges in the public health arena. Am J Preventive Med 18(4, Supplement 1):6–10
45. Fiest KM et al (2014) Systematic review and assessment of validated case definitions for depression in administrative data. BMC Psychiatry 14:289
46. Haroon SM et al (2015) Effectiveness of case finding strategies for COPD in primary care: a systematic review and meta-analysis. NPJ Prim Care Respir Med 25:15056
47. Killinger LZ (2012) Diagnostic challenges in the older patient. Chiropract Man Therapies 20(1):28
48. Price SJ et al (2016) Is omission of free text records a possible source of data loss and bias in Clinical Practice Research Datalink studies? A case-control study. BMJ Open 6(5):e011664
49. Stewart R, Davis K (2016) 'Big data' in mental health research: current status and emerging possibilities. Soc Psychiatry Psychiatr Epidemiol 51(8):1055–1072