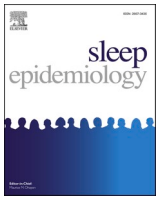
Sleep Epidemiology 5 (2025) 100110

Contents lists available at ScienceDirect

Sleep Epidemiology

journal homepage: www.elsevier.com/locate/sleepe



Validation of algorithms identifying diagnosed obstructive sleep apnoea and narcolepsy in coded primary care and linked hospital activity data in England

Helen Strongman [a,*], Sofia H. Eriksson [b,c], Kwabena Asare [a], Michelle A. Miller [d], Martina Sýkorová [a], Hema Mistry [d], Kristin Veighey [e], Charlotte Warren-Gash [a], Krishnan Bhaskaran [a]

[a] Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK
[b] National Hospital for Neurology and Neurosurgery/UCLH, London, UK
[c] Department of Clinical and Experimental Epilepsy, Institute of Neurology, University College London, London, UK
[d] Warwick Medical School, University of Warwick, Coventry, UK
[e] University Hospital Southampton, NHS Foundation Trust, Southampton, UK

ARTICLE INFO

Keywords:
Narcolepsy
Obstructive sleep apnoea
Electronic health records
Routinely collected data
Validation
Primary care

ABSTRACT

Purpose: To assist sleep epidemiology research, we created and tested the accuracy of five algorithms identifying diagnosed Obstructive Sleep Apnoea (OSA) and narcolepsy in routinely collected data from England (01/01/1998–29/03/2021).

Methods: The primary algorithm identified the first coded record in Clinical Practice Research Datalink (CPRD) primary care or linked hospital admissions data as an incident diagnosis of OSA (n = 92,222) or narcolepsy (n = 1072). Alternative algorithms required codes in CPRD, both datasets, or an additional proximate possible-sleep-related outpatient visit or excessive daytime sleepiness drug prescription (narcolepsy only). Staff in 73/1574 CPRD practices completed online questionnaires for a convenience sample of 144 OSA and 101 narcolepsy cases. We estimated Positive Predictive Values (PPVs) describing the proportion of cases confirmed by a gold standard hospital specialist diagnosis, the percentage of gold standard cases from the primary algorithm retained with alternative algorithms, and time between specialist and recorded diagnosis dates.

Results: Using the primary algorithm, the PPV (95 % CI) was 75.3 % (69.2–81.3) and 65.2 % (57.0–73.4) for OSA and narcolepsy, respectively: 80.6 % and 62.7 % of confirmed cases were recorded within 6 months of the specialist diagnosis. The CPRD-only algorithm increased the PPV to 85.3 (77.3–91.4, OSA) and 71.0 (58.8–81.3, narcolepsy) and retained high proportions of gold standard cases. Requiring additional outpatient or prescribing data increased PPVs, and for OSA improved diagnostic date accuracy, but omitted a high proportion of gold standard cases.

Conclusion: Highly accurate OSA diagnoses can be identified in routinely collected data. Recorded cases of narcolepsy are moderately accurate, but diagnosis dates are not.

Background

Sleep is a vital function to human health and daily living. Sleep can be disrupted by multiple environmental, lifestyle and medical factors including the primary sleep disorders Obstructive Sleep Apnoea (OSA) and narcolepsy [1], both of which are predominantly diagnosed in hospital-based specialist sleep centres using laboratory-based and ambulatory sleep studies [2].

Epidemiological research into sleep and other clinical fields can use routinely collected clinical and administrative data from healthcare systems. The early adoption of Electronic Health Records (EHR) to inform clinical care in National Health Service (NHS) general practices in the UK has supported the creation of large, longitudinal research databases linked to administrative hospital activity data[3,4]. In England, research data are restricted to structured data fields including coded diagnoses from primary care and hospital admissions activity

* Corresponding author.
E-mail address: helen.strongman@lshtm.ac.uk (H. Strongman).

https://doi.org/10.1016/j.sleepe.2025.100110

Available online 20 July 2025

data. Coded diagnostic data are not routinely collected in hospital outpatient activity data[3]. Studies investigating conditions treated in the hospital outpatient setting therefore rely on accurate coding of medical diagnoses by general practice staff based on receipt of clear information from hospital specialists, and the creation and use of high-quality code lists and algorithms by researchers to identify diagnoses[5]. There is limited information assessing the quality of recording of OSA or narcolepsy diagnoses in routinely collected clinical data or the validity of codelists and algorithms that researchers use to identify these conditions.

To assess the utility of routinely collected NHS England data for sleep disorder research, we therefore investigated the validity of algorithms identifying OSA and narcolepsy diagnoses in coded primary care and linked hospital activity data.

Methods

Study design

This validation study generated and tested the accuracy of five algorithms to identify diagnosed OSA and narcolepsy in coded primary care (CPRD Aurum[6]) or linked hospital activity data (HES Admitted Patient Care – APC[7] and outpatient) against a gold standard definition of diagnosis by a hospital specialist measured through a GP questionnaire study. GP practice staff members completed the questionnaire using detailed information available in the full medical record including, where available, information about diagnostic testing in letters received from hospital specialists.

This study was approved by the London School of Hygiene & Tropical Medicine Ethics Committee (Ref 101,296) and CPRD's Research Data Governance process (protocol 22_001887). The study protocol is available online[8]. CPRD supplies anonymised data for public health research; therefore, individual patient consent was not required for this study.

Setting

UK General Practices provide a wide range of primary care services including the diagnosis, treatment and prevention of common conditions, and act as a gate keeper to specialist services. Practices collect data to support and audit these services in electronic health record software including EMIS Web[9]. These data include Systematized Nomenclature of Medicine Clinical Terms (Snomed-CT) coded records of key clinical events and Dictionary of Medicines and Devices (DM+*D*) coded records of prescriptions. More detailed unstructured information describing clinical events are recorded in text boxes and by uploading documents such as letters received from hospitals and other care settings.

The Clinical Practice Research Datalink (CPRD) imports de-identified structured data from practices using EMIS software that have opted-in to providing data. Coded data is imported for all patients registered in these practices over time, except for people who have opted-out[10]. Imported data include key demographic information such as sex and year of birth and a pseudonymised identifier that can only be decoded by GP practice staff. Personal identifiers, and unstructured data recorded in text boxes or uploaded documents are not imported. CPRD process imported data to form the CPRD Aurum database; this includes anonymised patient and practice identifiers[6].

Periodically, CPRD Aurum primary care data are linked to multiple datasets, including separate HES databases for each hospital setting (APC, Outpatient, Accident & Emergency), Office for National Statistics (ONS) death data, and area-based deprivation datasets through a trusted third party[11]. Resultant linkage datasets include people with correct information necessary for data linkage at the time of processing.

We used the September 2023 CPRD Aurum build linked to the linkage dataset released in January 2022. The end of the data coverage period was 29/03/2021. The online questionnaire was administered through CPRD's Providing Online Verification of Electronic Health Records service (CPRD PROVE Plus). Direct Object Identifiers (DOIs) for each dataset and links to information about CPRD PROVE are provided in Supplementary Appendix Table 1.

Participants

Study population

The study population included people registered in 1574 practices actively contributing to CPRD Aurum in May 2023 and included in the linkage dataset. To avoid duplication, practices that contributed to CPRD both before and after a practice merger were excluded. We further excluded people whose records failed CPRD's data quality checks on recording and consistency of key variables[6], and for whom there was insufficient follow-up for an incident diagnosis to be identified (i.e. at least 90 days). The 90 day period was selected by visualising incidence rates of OSA and narcolepsy in the year following registration; records of OSA and narcolepsy during this time are likely to reflect earlier diagnoses[12] (Supplementary Appendix Fig. 1).

Incident sleep disorder cohorts

We used lookups provided by CPRD and NHS England to develop codelists including all codes for sleep apnoea and narcolepsy/cataplexy and recorded our decisions in a checklist[5] (Table 1). We developed and applied algorithms to identify incident OSA and narcolepsy cohorts and recorded diagnosis dates using these codelists. We first included people with a coded record of OSA or unspecified sleep apnoea or narcolepsy in CPRD Aurum or HES APC data. The date of the first inclusion record was assigned as the recorded diagnosis date. We then excluded people who were under 18 at diagnosis or had a prior record of central or primary sleep apnoea prior to or at diagnosis from the sleep apnoea cohort. To exclude prevalent diagnoses, we excluded people with a recorded diagnosis date before or within the first 90 days of practice registration from both cohorts.

Sample size considerations

We used convenience sampling to collect data for a minimum of 100 cases in each cohort with each recruited practice completing three to four questionnaires. This approach balanced the scientific requirement to produce estimates with reasonable precision with practical considerations related to practice recruitment. The sample size calculation is included in Supplementary Appendix Text 1.

Recruitment sample

We restricted the incident cohort to a recruitment sample consisting of one or two narcolepsy cases and two OSA cases per practice. People registered in practices with no narcolepsy cases were excluded. Random sampling was used to select two cases of each sleep disorder in practices with additional cases.

We sent a list of CPRD Aurum patient identifiers, practice identifiers, and diagnosis dates to CPRD for each sleep disorder cohort. CPRD further eliminated practices that have recently stopped contributing data, those that have stated that they do not want to participate in research studies, and practices from a small number of Clinical Research Networks (CRNs) with no or more complex information governance processes in place for CPRD PROVE studies.

Validation sample

All remaining practices were invited to sign up to a CPRD-web based agreement for this study on 18/03/2024 and for a member of practice staff to complete questionnaires for all cases in the recruitment sample for their practice on a first come, first served basis. Each practice received £110 per completed questionnaire, with an additional £110 incentive for completing all assigned questionnaires within one month.

The questionnaire was closed on 10/06/2024 when the minimum sample size of 100 completed questionnaires for the narcolepsy sample

Table 1
Checklists describing codelist creation methods for sleep apnoea and narcolepsy.

Code List Checklist*: Sleep Apnoea	
Metadata	
Title	Sleep Apnoea
Name	sleep_apnoea
Authors	Helen Strongman, Tim Quinnell, Sofia Eriksson
Target data source	Clinical Practice Research Datalink Aurum (March 2023) and linked Hospital Episode Statistics (HES) data
Terminology	SNOMED-CT (Aurum), EMIS (Aurum) and NHS 5th edition ICD-10 (HES).
Definition of clinical concept	
Concept	To identify sleep apnoea diagnoses and categorise them as follows: 1 "Obstructive Sleep Apnoea (OSA)" 2 "Obstructive sleep apnoea syndrome (OSAS)" 3 "Sleep apnoea NOS" 4 "Sleep apnoea syndrome NOS" 5 "Central sleep apnoea only" 6 "Primary sleep apnoea only" Notes: Non-specific codes are commonly used - Central and primary sleep apnoea are distinct from OSA and have different co-morbidities and health implications. - Mixed sleep apnoea codes are included in codes 1 to 4 because OSA is likely to be the dominant or at least co-dominant condition. In the validation study, we defined incident OSA using the following approach: [1] First ever record of OSA (category 1 to 4). [2] Exclude if prior record of central sleep apnoea or primary sleep apnoea (codes 5 & 6).
Timeframe	No restrictions
Accuracy	Algorithms accurately represent hospital specialist diagnoses. More stringent definitions improve accuracy but identify fewer cases. See study findings for more detail.
Setting	Diagnoses recorded in primary care and hospital activity inpatient data
Identify and evaluating existing code lists	
Source searched	Google search using the terms “cprd apnoea” “cprd apnea” and “cprd sleep apnoea”
Existing code lists found	None found
Verified by others	N/A
Verified by yourself	N/A
Existing code lists used	No
Create a new code list	
Synonyms	Sleep apnoea, sleep apnoea, OSA, sleep apnoea hypopnoea syndrome,
Exceptions	Primary sleep apnoea of newborn
Methods used	A script of lower-case search terms was used to identify relevant Read, medcodeid, snomedctconceptid in CPRD medical dictionaries and ICD-10 fields in NHS medical dictionaries. Terms anywhere in the string were identified. An additional search of Read Chapters was undertaken. Text searches were used to categorise codes.
Search terms	Inclusion: apnoea, apnea, OSA Exclusion: mosaic, mask, questionnaire, clinic, appliance, treatment, monitoring, recording, assessment, operation Terms without both “sleep” and either “apnoea” or “apnea” were also excluded
Hierarchy used to extend search	Inclusion: Read Chapters H5B, R005 and Fy03
Decisions made while iterating	Separate categories were created for central and primary sleep apnoea and care pathway codes were excluded from the code list because they were rare and mostly indicate sleep apnoea testing.
Categories	1. Obstructive Sleep Apnoea (OSA) This category includes terms including the strings "obstructive" or "mixed" but not "syndrome" 2. Obstructive Sleep Apnoea Syndrome (OSAS) This category comprised terms that included "obstructive" or "mixed" and "syndrome" 3. Sleep apnoea NOS This category includes terms not otherwise classified 4. Sleep apnoea syndrome NOS This category includes terms not otherwise classified that include the string "syndrome" 5. Central sleep apnoea only This category comprised terms that included the string "central" 6. Primary sleep apnoea only This category comprised terms that included the string "primary"
Review, finalise and publish	
Reviewers	Dr Sofia Eriksson (Neurologist & sleep specialist, University College London Hospitals NHS Foundation Trust and UCL): code list validity. Dr Tim Quinnell (Respiratory & sleep specialist, Royal Papworth Hospital Trust)
Scope of review	The draft code list, search terms and exclusion terms were reviewed.
Evidence of review	The process is documented in HTML files: (See “Resources published” section)
Internal checks	Undertaken by Helen Strongman
External checks	Validated as part of this study using a questionnaire asking GPs to confirm secondary care (hospital) diagnosis/ treatment for sleep apnoea. This was used as the “Gold Standard” comparison to assess the CPRD/HES codes identified (see resources published).
Code list published	The code list is published on the project’s LSHTM Data Compass page https://doi.org/10.17037/DATA.00004742. And in the study’s Github repositories at: https://github.com/hstrongman/OSA-narc_CPRD_validation Files include: codelist_sleep_apnoea_aurum.txt codelist_sleep_apnoea_aurum.dta (Github only)
Resources published	Strongman, H., S. H. Eriksson, K. Asare, M. A. Miller, M. Sykorova, H. Mistry, K. Veighey, C. Warren-Gash and K. Bhaskaran. "Validation of algorithms identifying diagnosed Obstructive Sleep Apnoea and narcolepsy in coded primary care and linked hospital activity data in England." Sleep Epidemiology 2025 https://doi.org/10.1016/j.sleepe.2025.100110 . The Do file and HMTL documents describing the code list derivation and search strategy are in the study’s GitHub repositories. Files include: codelist_sleep_apnoea.do (refers to associated text files) codelist_sleep_apnoea_description.html (early version of this checklist) codelist_sleep_apnoea_derivation_aurum.html codelist_sleep_apnoea_derivation_gold.html codelist_sleep_apnoea_derivation_hesicd.html
Code List Check List*: Narcolepsy	
Metadata	
Title	Narcolepsy
Name	narcolepsy
Authors	Helen Strongman, Tim Quinnell, Sofia Eriksson
Target data source	Clinical Practice Research Datalink Aurum (March 2023) and linked Hospital Episode Statistics data
Terminology	SNOMED-CT (Aurum), EMIS (Aurum) and NHS 5th edition ICD-10 (HES).
Definition of clinical concept	
Concept	To identify all narcolepsy diagnoses
Timeframe	No restrictions
Accuracy	Records of diagnosed narcolepsy are mostly accurate but diagnosis dates are not. More stringent definitions improve accuracy but identify fewer cases. See study findings for more detail.
Setting	Diagnoses recorded in primary care and hospital activity inpatient data
Identify and evaluating existing code lists	
Source searched	Google search using the term “cprd narcolepsy” and “cprd cataplexy”
Existing code lists found	None found

(continued on next page)

Table 1 (*continued*)

Code List Checklist*: Sleep Apnoea	
Verified by others	N/A
Verified by yourself	N/A
Existing code lists used	No
Create a new code list	
Synonyms	Narcolepsy, Cataplexy
Exceptions	None
Methods used	A script of lower-case search terms was used to identify relevant Read, medcodeid, snomedctconceptid in CPRD medical dictionaries and ICD-10 fields in NHS medical dictionaries. Terms anywhere in the string were identified. An additional search of Read Chapters was undertaken.
Search terms	Inclusion: narcolep, cataplexy Exclusion: none
Hierarchy used to extend search	Inclusion Read Chapter: F27
Decisions made while iterating	There were Read codes in chapter F27 with no associated observations/terms. These are not included in the codelist.
Categories	1: Narcolepsy This category included terms with the string "narcolep" in the CPRD data dictionaries. Therefore, this category included terms containing both "narcolep" and "cataplexy". 2: Cataplexy only This category included terms which included the string "cataplexy" but not "narcolep". Cataplexy only codes are uncommon and were not included in definitions of narcolepsy for this study.
Review, finalise and publish	
Reviewers	Dr Sofia Eriksson (Neurologist & sleep specialist, University College London Hospitals NHS Foundation Trust and UCL): code list validity. Dr Tim Quinnell (Respiratory & sleep specialist, Royal Papworth Hospital Trust)
Scope of review	The draft code list, search terms and exclusion terms were reviewed.
Evidence of review	The process is documented in HTML files (See "Resources published" section).
Internal checks	Undertaken by Helen Strongman
External checks	Validated in this study using a questionnaire asking GPs to confirm secondary care (hospital) diagnosis/treatment for narcolepsy. This was used as the "Gold Standard" comparison to assess the CPRD codes identified (see resources published)
Code list published	The code list is published on the project's LSHTM Data Compass page: https://doi.org/10.17037/DATA.00004742. And in the study's Github repositories at: https://github.com/hstrongman/OSA-narc_CPRD_validation Files include: codelist_narcolepsy_aurum.txt codelist_narcolepsy_aurum.dta (Github only) codelist_narcolepsy_hesapc.txt codelist_narcolepsy_hesapc.dta (Github only)
Resources published	Strongman, H., S. H. Eriksson, K. Asare, M. A. Miller, M. Sykorova, H. Mistry, K. Veighey, C. Warren-Gash and K. Bhaskaran. "Validation of algorithms identifying diagnosed Obstructive Sleep Apnoea and narcolepsy in coded primary care and linked hospital activity data in England." Sleep Epidemiology 2025 https://doi.org/10.1016/j.sleepe.2025.100110. The Do files and HMTL documents describing the code list derivation and search strategy are in the study's GitHub repositories Files include: codelist_narcolepsy.do (refers to associated text files) codelist_narcolepsy_description.html (early version of this checklist) codelist_narcolepsy_derivation_aurum.html codelist_narcolepsy_derivation_hesicd.html

*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 [version 2; peer review: 3 approved]. NIHR Open Res. 2024 Sep 18;4:20.

was reached. The resulting narcolepsy and OSA samples are referred to as the validation samples.

Validation questionnaire

The questionnaire was designed and reviewed in a text format by the full study team including epidemiologists, statisticians, sleep disorder clinicians and GPs. Following scientific approval of the published protocol, changes were made based on testing and narrative feedback by a General Practitioner (KV) and recommended adaptations for online questionnaire design by the CPRD PROVE team. The questionnaire was hosted on RedCap and tested by the study team and CPRD PROVE team. The system allowed respondents to indicate uncertainty in comments boxes instead of answering otherwise compulsory questions; missing values were returned for these responses and reset to "uncertain" for questions that included this response.

Fig. 1 summarises the questionnaire structure. The full questionnaire is provided in Supplementary Appendix Text 2.

Variables

Full variable definitions are provided in Supplementary Appendix Table 2. All codelists are published online[13]. Data management and statistical analyses were completed in Stata MP, version 17. Full programming code is available[14]

Validation questionnaire variables

We designated gold standard diagnosed cases as those confirmed to be diagnosed or treated by a hospital specialist (Q1a "Has this patient been diagnosed or treated for the specified sleep disorder (see above) by a hospital specialist at any time?"=yes). The specialist diagnosis date was captured in Q1b "When was the patient diagnosed by a hospital specialist?" for gold standard cases. Further questions were designed to capture information about methods used to diagnose gold standard cases and the origin of false positive diagnoses. Data transformations for further descriptive variables are described in Supplementary Appendix Table 2.

Sleep disorder data and demographic variables

We used routinely collected data to define the following variables describing data used to identify sleep disorder cases: category of diagnosis code (OSA only), origin of diagnosis code (CPRD and/or HES APC), calendar year at the recorded diagnosis and estimated person-years of follow-up prior to the recorded diagnosis. To support potential alternative sleep disorder algorithms, we additionally measured Excessive Daytime Sleepiness (EDS) drug prescriptions recorded at any time (narcolepsy only) and proximate HES outpatient visits (overall and sleep-related) in the 6 months before or after the recorded diagnosis. Sleep-related visits included those to outpatient clinics lead by neurology, respiratory, paediatrics, ENT and anaesthetics consultants. All analyses using HES outpatient data were restricted to people included in linkage processing for the HES outpatient data whose recorded diagnosis date was within the HES outpatient data coverage period (01/04/2003–30/10/2020). EDS drug prescriptions were identified in CPRD Aurum data and included prescriptions for modafinil, methylphenidate or dexamfetamine (i.e. drugs that are typically prescribed as 1st or 2nd line treatment).

Further demographic data defined using CPRD data included age at recorded diagnosis, sex, Body Mass Index (OSA only), ethnicity, practice area-based deprivation and urban-rural status, and practice size.

Alternative sleep disorder algorithms

We developed and applied alternative incident sleep disorder algorithms to the validation sample. These identified subsets of the cohort that met more stringent criteria and were hypothesised to reduce false positives, increasing the PPV, while reducing the number of gold

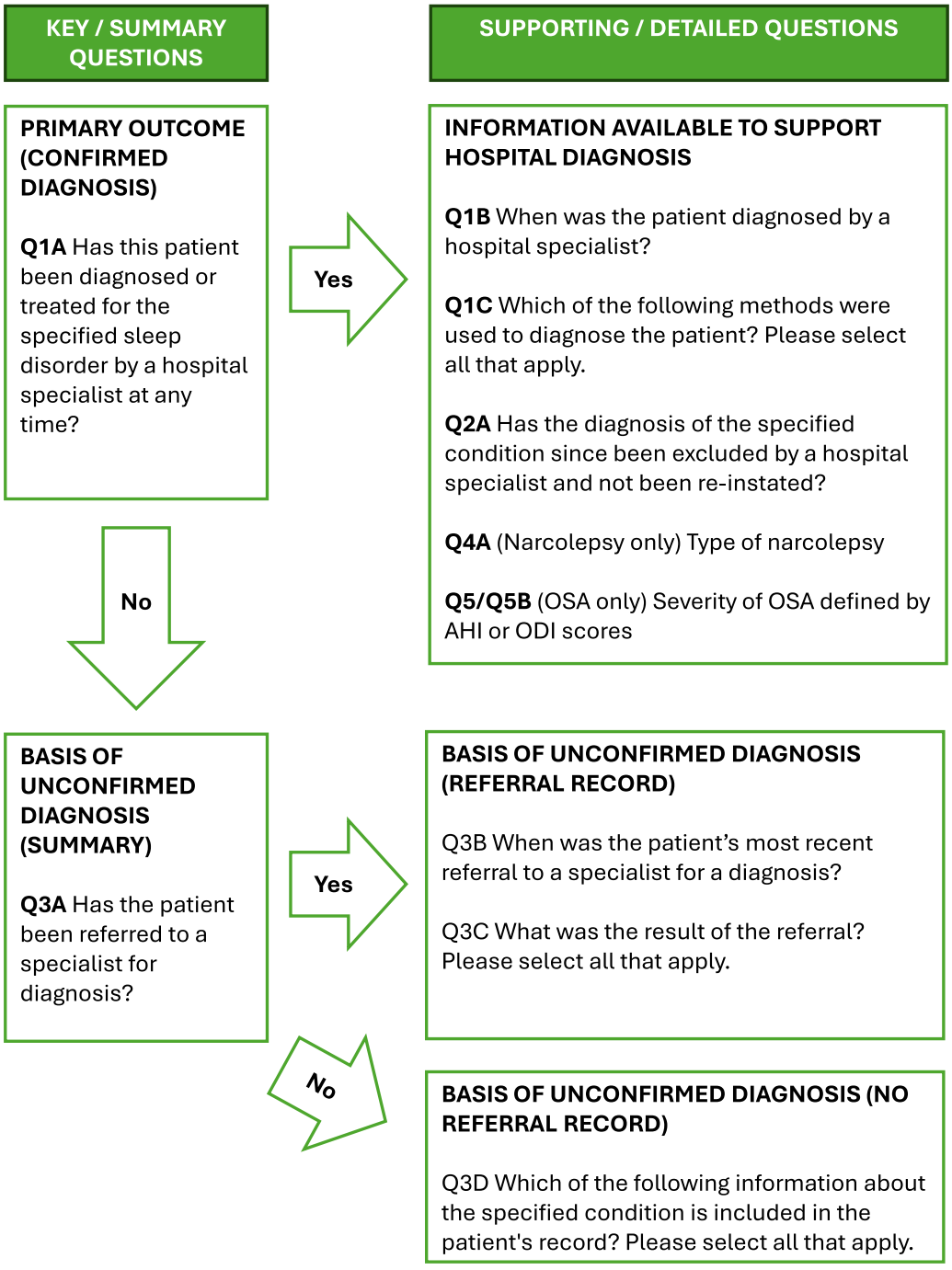


Fig. 1. Validation questionnaire summary.

standard cases identified. Pre-planned alternative algorithms included: coded records in CPRD data (no requirement for a code in HES APC data) and coded records in both CPRD and HES APC data. We explored possible alternative algorithms requiring both coded diagnoses and either EDS drug prescriptions (narcolepsy only) or proximate outpatient visits as defined above. We did not consider data describing OSA severity or use of Positive Airway Pressure devices as this is rarely recorded in primary care data. Recorded diagnosis dates for all algorithms were set to the first date when all requirements were met.

Statistical methods

Descriptive and exploratory analyses

We compared sleep disorder data and demographic characteristics recorded in the incident cohorts and validation samples, and in gold standard and false positive cases within the validation study. Sleep disorder data and demographic variables whose distribution differed substantially in both comparisons were identified as standardisation variables. Possible alternative sleep disorder algorithms were confirmed for variables where distributions differed between gold standard and false positive cases.

Validation estimates

We estimated the PPV for each sleep disorder algorithm by dividing the number of cases confirmed in the gold standard definition (i.e. those confirmed by GP staff to have been diagnosed/treated by a hospital specialist) by the total number of cases identified using the algorithm. PPVs for standardisation variables were directly standardised to match the distribution in the incident cohorts. We additionally estimated the percentage of gold standard cases identified using the primary algorithm that were retained when using the stricter alternative algorithms (gold standard cases meeting criteria for alternative algorithm/gold standard cases for primary algorithm). Exact binomial methods were used to estimate 95 % confidence intervals.

To compare PPVs between demographic groups, we fitted generalized linear models with a binomial distribution and robust standard errors. This approach estimates relative differences in PPVs between groups. This analysis was based on binary variables to maximise power; ethnicity was not included due to small numbers in the non-white groups.

Numbers representing 1 to 4 people, are reported as <5, in line with CPRD policy.

Results

Participants

From a study population of over 10 million people still registered in active CPRD Aurum practices at the last data collection date, 92,222 and 1072 individuals were identified as being diagnosed with incident OSA or narcolepsy, respectively, while registered in the practice (Supplementary Appendix Fig. 2). The recruitment sample included 1288 people with OSA and 901 with narcolepsy, for whom 144 and 101 questionnaires were completed by general practice staff in 73 practices.

Descriptive and exploratory analyses

The median (IQR) age of the OSA and narcolepsy validation samples was 53.8 (44.7, 61.4) and 40.1 (IQR 23.0, 48.8), respectively. 27.8 % (49) of the OSA and 52.5 % (53) of the narcolepsy validation samples were female. The majority (>85 %) of both validation samples were white and 69.4 % of the OSA validation sample was obese (Table 2).

Sleep disorder diagnoses were recorded between 1998 and 2021 with most codes derived from the CPRD data; non-specific sleep apnoea codes were used for one-third-of the OSA cohort/sample. The majority of people in the validation sample were eligible for linkage to HES outpatient data. Of these, 85.3 % (110, OSA) and 85.3 % (64, narcolepsy) had a proximate outpatient visit and 53.5 % (OSA) and 66.7 % (narcolepsy) had proximate possible sleep-related outpatient visit. 48.5 % ($n = 49$) of the validation sample had at least one prescription of an EDS drug in their primary care record (Table 2).

Diagnoses recorded in HES APC were more prevalent in the validation samples (34.0 % OSA, 38.6 % narcolepsy) than the incident cohorts (28.1 % OSA, 33.1 % narcolepsy) (Table 2) and less prevalent in gold standard cases (23.1 % OSA, 28.8 % narcolepsy) than false positive cases (62.5 % OSA, 52.4 % narcolepsy) (Supplementary Appendix Table 3). The source of the diagnosis code was therefore identified as a stratification variable.

Substantial differences were observed in the distribution of outpatient visits to a possible sleep-related specialist (for OSA) and EDS prescription (for narcolepsy) in gold standard vs false positive cases (Supplementary Appendix Table 3). We therefore selected alternative algorithms including these variables.

Outcomes and estimation

For each algorithm, Fig. 2 describes the PPV, percentage of gold standard cases from the primary algorithm that were retained using the

Table 2
Sleep disorder data and demographic characteristics recorded in the full incident cohorts and validation samples for OSA and narcolepsy.

	incident OSA	validation OSA	incident narcolepsy	validation narcolepsy
N+	92,222	144	1072	101
RECORDING OF SLEEP DISORDER DIAGNOSIS				
Source of diagnostic code*				
Primary care	66,306 (71.9)	95 (66.0)	717 (66.9)	62 (61.4)
Inpatient hospital activity data	25,916 (28.1)	49 (34.0)	355 (33.1)	39 (38.6)
Most specific sleep apnoea code type recorded at index				
OSA code	55,557 (60.2)	95 (66.0)		
sleep apnoea	36,665 (39.8)	49 (34.0)		
Person-years before diagnosis/index				
Mean (SD)	16.6 (13.4)	16.5 (13.6)	13.0 (11.4)	13.8 (11.5)
Median (IQR)	13.7 (5.7, 24.1)	14.5 (5.0, 23.5)	9.9 (4.3, 18.3)	12.2 (4.9, 19.3)
Calendar year at index				
Mean (SD)	2014.4 (5.1)	2014.1 (5.3)	2012.7 (6.1)	2012.9 (6.6)
Median (IQR)	2016 (2011, 2018)	2016.0 (2011, 2018)	2014.0 (2008, 2018)	2014.0 (2010, 2019)
ADDITIONAL INFORMATION IN ROUTINELY COLLECTED DATA				
Outpatient visit within 6 months of diagnosis (HES OP data)				
Linked OP data available	79,511 (86.2)	129 (89.6)	875 (81.6)	75 (74.3)
All	72,285 (90.9)	110 (85.3)	771 (88.1)	64 (85.3)
Neurology	4,805 (6.0)	<5 (<3.9)	363 (41.5)	35 (46.7)
Respiratory	39,967 (50.3)	56 (43.4)	264 (30.2)	23 (30.7)
Paediatric (NOS)	63 (0.1)	0 (0.0)	111 (12.7)	8 (10.7)
Ear Nose & Throat	15,251 (19.2)	23 (17.8)	75 (8.6)	<5 (<6.7)
Anaesthetics	63 (0.1)	0 (0.0)	111 (12.7)	8 (10.7)
Sleep-related consultants combined	50,226 (63.2)	69 (53.5)	612 (69.9)	50 (66.7)
Excessive Daytime Sleepiness drug prescriptions				
EDS drug prescription ever			594 (55.4)	49 (48.5)
Days between recorded diagnosis date & 1st EDS drug prescription (positive = recorded first)				
Mean (SD)			-88.5 (1207.7)	-222.7 (819.2)
Median (IQR)			36.0 (-2.0, 200.0)	16.0 (-165.0, 158.0)
Recorded diagnosis compared to date of 1st EDS drug prescription				
> 6 months before			101 (9.4)	11 (10.9)
within 6 months			335 (31.2)	27 (26.7)
> 6 months after			158 (14.7)	11 (10.9)
missing			478 (44.6)	52 (51.5)
CHARACTERISTICS				
Age at diagnosis (years)				
Mean (SD)	52.4 (12.5)	52.5 (12.4)	37.5 (18.3)	37.6 (17.2)

(*continued on next page*)

Table 2 (*continued*)

	incident OSA	validation OSA	incident narcolepsy	validation narcolepsy
Median (IQR)	52.4 (43.8, 61.0)	53.8 (44.7, 61.4)	37.3 (23.5, 50.4)	40.1 (23.0, 48.8)
Sex				
male	62,722 (68.0)	104 (72.2)	512 (47.8)	48 (47.5)
female	29,500 (32.0)	40 (27.8)	560 (52.2)	53 (52.5)
Body Mass Index				
Under/normal weight	8,062 (8.7)	13 (9.0)		
Overweight	20,967 (22.7)	27 (18.8)		
Obesity class I	22,889 (24.8)	42 (29.2)		
Obesity class II	16,699 (18.1)	29 (20.1)		
Obesity class III+	19,269 (20.9)	29 (20.1)		
missing	4,336 (4.7)	<5 (<3.5)		
Ethnicity				
0. White	80,345 (87.1)	126 (87.5)	902 (84.1)	88 (87.1)
1. South Asian	5,691 (6.2)	8 (5.6)	39 (3.6)	<5 (<5.0)
2. Black	3,467 (3.8)	9 (6.2)	83 (7.7)	6 (5.9)
3. Other	1,386 (1.5)	0 (0.0)	16 (1.5)	<5 (<5.0)
4. Mixed	968 (1.0)	<5 (<3.5)	21 (2.0)	0 (0.0)
missing	365 (0.4)	0 (0.0)	11 (1.0)	<5 (<5.0)
Carstairs quintile				
1 (least deprived)	11,167 (12.1)	32 (22.2)	119 (11.1)	24 (23.8)
2	18,150 (19.7)	27 (18.8)	189 (17.6)	18 (17.8)
3	20,763 (22.5)	34 (23.6)	260 (24.3)	26 (25.7)
4	20,480 (22.2)	25 (17.4)	233 (21.7)	18 (17.8)
5 (most deprived)	21,662 (23.5)	26 (18.1)	271 (25.3)	15 (14.9)
Urban Rural				
urban	77,639 (84.2)	113 (78.5)	924 (86.2)	78 (77.2)
rural	14,319 (15.5)	31 (21.5)	143 (13.3)	23 (22.8)
missing	264 (0.3)	0 (0.0)	5 (0.5)	0 (0.0)
Practice size				
Mean (SD)	42003.6 (34154.6)	41277.8 (33621.7)	41990.5 (34241.2)	43438.7 (37083.2)
Median (IQR)	32319 (22644, 47488)	31002 (20960, 46988)	32201 (22470.5, 46898.5)	31759 (20749, 48163)

* A small number of people had a code in both sources on the index date (<2%). These are coded as primary care to avoid small cell counts.

stricter alternative algorithms, and time difference between the recorded and specialist diagnosis date. Using the primary algorithm, the standardised percentage (95 % CI) of gold standard cases was 75.3 (69.2–81.3) for OSA and 65.2 (57.0–73.4) for narcolepsy. The median number of days between the recorded and specialist diagnosis date, and percentage of gold standard cases recorded within 6 months of the specialist diagnosis, was −1.0 (IQR −72.0,0.0, 80.6 %) for OSA and 0.0 (IQR −71.0, 273.0, 62.7 %) for narcolepsy. Using the CPRD only algorithm increased the percentage (95 % CI) of gold standard cases to 85.3 (77.3–91.4) for OSA and 71.0 (58.8–81.3) for narcolepsy, retained 89.4 % (95 % CI 81.9–94.6) of gold standard OSA and 83.1 % (95 % CI 71.0–91.6) of gold standard narcolepsy cases from the primary algorithm, and made little difference to the accuracy of the diagnosis date.

For OSA, requiring CPRD data with a proximate possible sleep-related outpatient visit to identify diagnosed cases provided the best balance between maximising the PPV (98.2 %, 95 % CI 90.3–100) while retaining the highest proportion of gold standard cases from the primary algorithm 55.7 % (95 % CI 45.2–65.8). Using this algorithm also increased the percentage of recorded cases within 6 months of the specialist diagnosis to 86.9 %. For narcolepsy, requiring CPRD data with an EDS drug prescription provided the best balance; PPV 88.9 % (95 % CI 75.9–96.3), gold standard cases retained 67.8 % (95 % CI 54.4–79.4) with a further reduction in the accuracy of the diagnosis date.

There is little evidence that the PPV differed between demographic and practice characteristics, except for Body Mass Index for OSA: PPV (95 % CI) 67.1 % (55.8–77.1) for the obesity class I and lower category and 82.8 % (70.6–91.4) for the obesity class II and above category. There is weak evidence that the crude PPV for narcolepsy was lower in later calendar-years. (Supplementary Appendix Table 4).

Descriptive analyses of questionnaire responses

Table 3 describes responses to the full validation questionnaire and sample and Supplementary Table 5 summarises key information across all algorithms.

At least one objective diagnostic method was identified for 77.9 % and 59.3 % of gold standard OSA and narcolepsy cases, respectively. This proportion was similar across all algorithms for OSA but increased with more stringent narcolepsy algorithms (max 70.4 % CPRD with outpatient visit to sleep-related specialty). The type of narcolepsy was identified in half of cases (30.5 % type 1, 20.3 % type 2). OSA severity was recorded for nearly 80 % of cases with a fairly even split between mild, moderate and severe cases. The original hospital specialist diagnosis was excluded at a later date for 9.6 % ($n = 10$) of OSA and 8.5 % ($n = 5$) of narcolepsy gold standard cases reflecting the complexity in diagnosing these conditions and potential alleviation of OSA symptoms.

9.7 %[14] and 21.8 %[22] of people in the OSA and narcolepsy validation samples, respectively, were referred to hospital specialists but not diagnosed. Amongst these cases, the diagnosis had been excluded or another sleep disorder had been diagnosed in 50 % of OSA and nearly all narcolepsy cases.

There was no evidence of referral to a hospital specialist for 18.1 % (26) and 19.8 %[20] of people in the OSA and narcolepsy validation samples, respectively. The reason for this is unclear in 69.2 % (OSA) and 50 % (narcolepsy) of cases. Others were either suspected cases, diagnosed by a GP, or not referred for diagnosis.

Discussion

Summary

We tested the accuracy of 5 algorithms to identify diagnosed OSA and narcolepsy in coded primary care (CPRD Aurum) and hospital activity data (HES APC and HES outpatient) against a gold standard definition of hospital specialist diagnosis recorded in the full GP record. The primary algorithm of a coded record in either CPRD Aurum or HES APC data had the lowest PPV for both OSA (75.3 95 % CI 69.2–81.3) and narcolepsy (65.2 95 % CI 57.0–73.4). The accuracy of the diagnostic date was moderate for OSA and poor for narcolepsy. Using CPRD data only increased the PPV to 85.3 (95 % CI 77.3–91.4) for OSA and 71.0 (95 % CI 58.8–81.3) for narcolepsy, while losing 10 to 20 % of gold standard cases from the primary algorithm. More stringent algorithms that achieved the best balance between increasing the PPV while retaining gold standard cases required CPRD data with a proximate possible sleep-related outpatient visit for OSA (PPV 98.2 %, 95 % CI 90.3–100.0) and CPRD data with a prescription record for an EDS drug for narcolepsy (PPV 88.9 % 95 % CI 75.9–96.3).

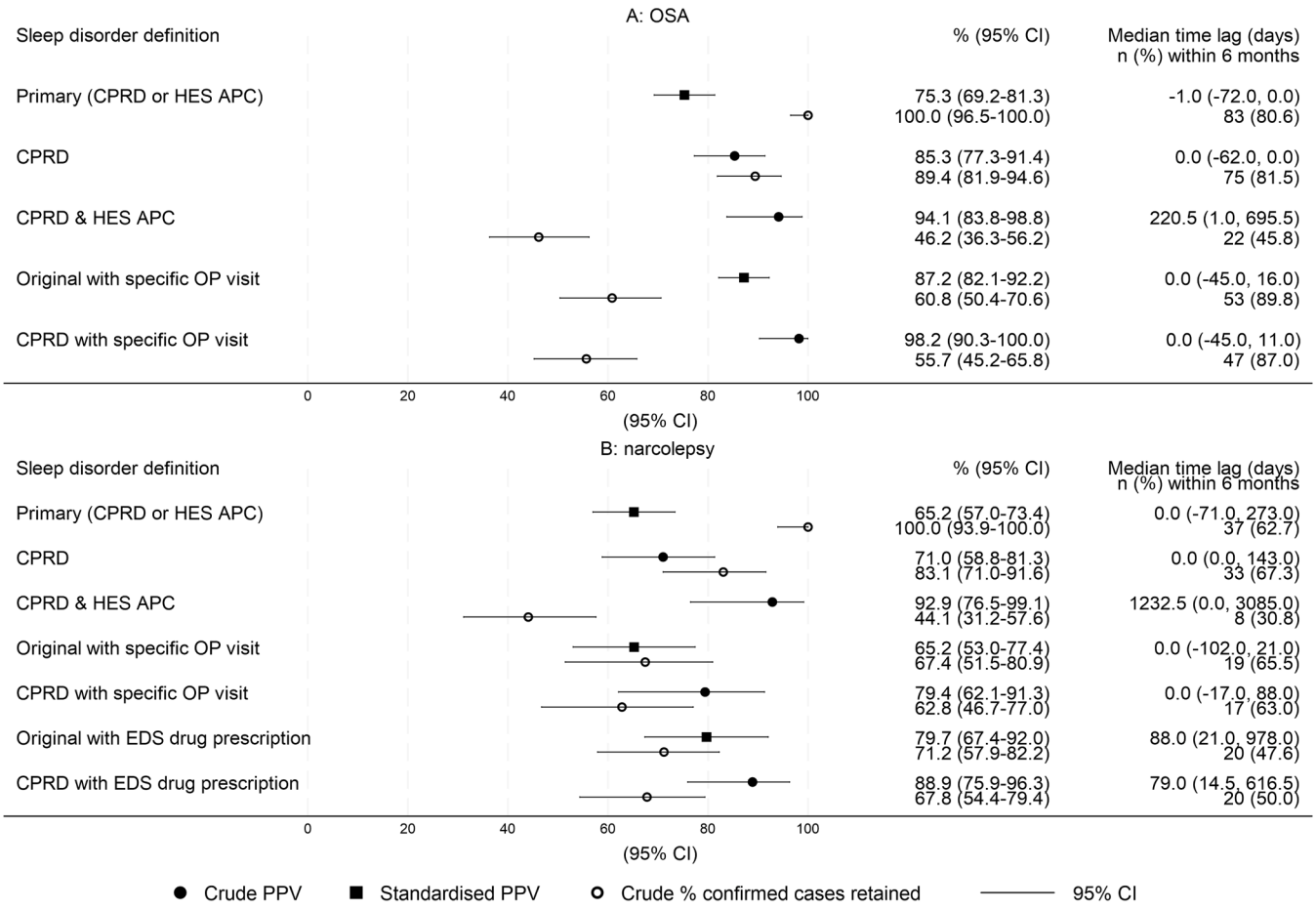


Fig. 2. PPV, percentage of gold standard cases from the primary algorithm retained with stricter alternative algorithms, and time between recorded and specialist diagnosis dates, by algorithm

Caption [1]: gold standard cases were identified as being diagnosed or treated by a hospital specialist in a GP questionnaire [2]. Substantial differences were observed in the proportion of cases identified in clinical practice research datalink (CPRD) or hospital episode statistics admitted patient care data (HES APC) in the incident cohort versus the validation sample and gold standard cases versus false positive cases. Standardised PPVs for the primary algorithm are therefore directly standardised to match the distribution of the incident cohort[3]. Gold standard cases retained are included in both the primary and alternative algorithm [4]. Median time lag, positive = recorded diagnosis date late.

Strengths and limitations

The key strength of our study is that we were able to assess whether coded incident diagnoses of OSA and narcolepsy in de-identified data represented specialist diagnoses using additional information that is available in the full primary care record, through a structured questionnaire. Setting hospital specialist diagnoses as a gold standard matches diagnosis and care pathways for both OSA and narcolepsy, both of which require specialist care. To enable our algorithms to be used in future studies, we have recorded methods used to construct code lists for each source[5], and our analysis code.

Our gold standard definition relies on specialists transferring accurate information to GPs and GP staff being able to find and interpret this information. Moderate proportions of objective testing for gold standard cases (77.9 % OSA, 59.3 % narcolepsy) suggests a lack of clear and detailed information being consistently transferred to GPs from hospital specialists. In the absence of this information, we have assumed that the specialist diagnosis was based on objective testing. GP staff members were able to extract information about OSA severity and narcolepsy type for a proportion of true positives (80 % OSA, 50 % narcolepsy), but this information is not commonly recorded in coded primary care data. Furthermore, we assessed whether coded records accurately identified specialist diagnoses at the time of recording. Questionnaire responses indicated that 9.6 % of OSA and 8.5 % of narcolepsy diagnoses were excluded by a hospital specialist at a later date. These may be borderline cases for which symptoms vary over time or cases where OSA symptoms have been alleviated by lifestyle changes; diagnoses may also be excluded as knowledge increases or people are referred to tertiary sleep centres for more stringent objective testing.

Whilst the origin of the majority of false positive cases identified using the primary algorithm was confirmed in the questionnaire (e.g. diagnosis excluded after a referral), 45 % (40) and 22.7 %[18] of false positive OSA and narcolepsy cases, respectively, had no record of referral to a specialist and no clear reason for the condition being recorded. Algorithms using CPRD data only resulted in a higher PPV and lower proportion of false positive OSA cases with no referral to hospital specialists. This may reflect information not being transferred from hospitals to GPs, or GP staff completing this questionnaire not being able to find this information in scanned letters, resulting in false negative cases. This pattern persists for algorithms requiring outpatient visits and EDS drug prescribing by the general practice (narcolepsy only).

Our analysis plan considered uncertainty and potential selection bias introduced through our convenience sampling strategy. This strategy allowed us to reach our target sample within 3 months, include a large number of practices, and provide reasonably certain estimates and detailed questionnaire responses for the primary algorithm. Wider confidence intervals for narrower alternative algorithms mean that differences between algorithms may be due to chance alone, and small

Table 3
Validation questionnaire responses.

	OSA	Narcolepsy
***N*+**	144	101
ALL CASES		
Recorded role of hospital specialist		
Diagnosis or treatment of sleep disorder	104 (72.2)	59 (58.4)
Referral but no diagnosis	14 (9.7)	22 (21.8)
No referral	26 (18.1)	20 (19.8)
CASES DIAGNOSED OR TREATED BY A HOSPITAL SPECIALIST (OSA $n = 104$, narcolepsy $n = 59$)		
Days between the index and diagnosis date (positive = recorded diagnosis later)		
Mean (SD)	−137.7 (654.4)	503.6 (2526.5)
Median (IQR)	−1.0 (−72.0, 0.0)	0.0 (−71.0, 273.0)
Months between the Recorded and specialist diagnosis date*		
Recorded diagnosis > 6 months before	13 (12.5)	6 (10.2)
Recorded diagnosis 1 to 6 months before	21 (20.2)	11 (18.6)
Recorded diagnosis within 1 month	57 (54.8)	20 (33.9)
Recorded diagnosis 1 to 6 months after	5 (4.8)	6 (10.2)
Recorded diagnosis > 6 months after	7 (6.7)	16 (27.1)
Methods used to diagnose the patient (all that apply)		
Electroencephalogram (EEG) telemetry	0 (0.0)	5 (8.5)
Polysomnography	36 (34.6)	24 (40.7)
Hospital respiratory polygraphy	14 (13.5)	<5 (<8.5)
Multiple Sleep Latency Test (MSLT)	<5 (<4.8)	24 (40.7)
Home respiratory polygraphy	6 (5.8)	<5 (<8.5)
Home oximetry	17 (16.3)	<5 (<8.5)
Successful Continuous positive airway pressure therapy (CPAP) trial	30 (28.8)	<5 (<8.5)
Lumbar puncture	0 (0.0)	<5 (<8.5)
Patient history	27 (26.0)	22 (37.3)
Unclear or no information available	21 (20.2)	12 (20.3)
None of the above	<5 (<4.8)	<5 (<8.5)
At least one objective diagnostic method identified	81 (77.9)	35 (59.3)
Type of narcolepsy(1)		
Type 1 narcolepsy or cataplexy		18 (30.5)
Type 2 narcolepsy or no cataplexy		12 (20.3)
No information available		16 (27.1)
Other/unclear		13 (22.0)
Severity of OSA (AHI or ODI(2))		
Mild (AHI 5 to 14)	22 (21.2)	
Moderate (AHI 15 to 30)	29 (27.9)	
Severe (AHI >30)	32 (30.8)	
Unclear / no information available	21 (20.2)	
Diagnosis excluded by specialist at a later date	10 (9.6)	5 (8.5)
CASES REFERRED TO A HOSPITAL SPECIALIST BUT NOT DIAGNOSED (OSA $n = 14$, narcolepsy $n = 22$)		
Months between index and referral date (positive = index later)		
Mean (SD)	6.9 (56.2)	−8.3 (62.4)
Median (IQR)	1.7 (−23.8, 13.4)	1.8 (−2.3, 7.4)
Result of the referral (all that apply)*		
Patient on waiting list for initial appointment	<5 (<35.7)	0 (0.0)
Patient undergoing investigation at hospital	<5 (<35.7)	0 (0.0)
Diagnosis of specified condition excluded by hospital specialist	7 (50.0)	8 (36.4)
Diagnosed with sleep apnoea or other sleep disorder	0 (0.0)	12 (54.5)
Referral rejected	5 (35.7)	<5 (<22.7)
CASES NOT REFERRED TO A HOSPITAL SPECIALIST (OSA $n = 26$, narcolepsy $n = 20$)		
Information included in patient's record (all that apply)		
Suspected narcolepsy/OSA/sleep apnoea	<5 (<19.2)	6 (30.0)
OSA/sleep apnoea/narcolepsy diagnosed by GP	6 (23.1)	6 (30.0)
Patient request not to be referred for diagnosis	<5 (<19.2)	0 (0.0)
GP decision not to refer for diagnosis	<5 (<19.2)	<5 (<25.0)
None of the above	18 (69.2)	10 (50.0)

(1)There was one missing value in the narcolepsy cohort for each of these variables. Missing narcolepsy type recoded as other/unclear.
[2]<5 values were provided as ODI scores and coverted to AHI.

cell count requirements prevent us from describing responses in full. We identified higher proportions of cases recorded in HES APC data only in the validation sample, compared to the full eligible incident sample, as a potential selection bias. We mitigated this bias by stratifying PPVs by source for algorithms affected by this bias. We were unable to identify mechanisms of selection bias caused by unmeasured differences between the full incident sample and validation sample or by restricting our study to active practices and cases. Additionally, recording practices may have changed during the Covid-19 pandemic[15]; this may explain our observation of weak evidence of a lower PPV for narcolepsy later in the study period.

As we did not validate records with no recorded diagnosis, we were unable to estimate negative predictive values (i.e. the proportion of people without a recorded diagnosis who have not been diagnosed) or the sensitivity of our algorithms (i.e. the proportion of gold standard (hospital specialist) diagnosed OSA and narcolepsy cases in the study population that we identified with our algorithms. The negative predictive value of our algorithms is likely to be high as the ratio of diagnosed to undiagnosed/sleep disorder free people in the population is low, even in high risk OSA groups.

Strengths and limitations in comparison to existing studies

There is minimal published evidence assessing the validity of OSA and narcolepsy records in routinely collected data. A validation study of recording of sleep disorders diagnosed in a single Canadian sleep centre found that non-specific coding of sleep disorders was common in all data sources, and particularly poor in inpatient data[16]. It is possible that GPs and hospitals use similarly generic terms to record OSA and narcolepsy in England; these cases would be not be identified using our algorithms. In contrast, recording of sleep apnoea was found to be highly accurate in electronic health record data from US hospitals that participated in a sleep apnoea genetics study[17]. The latter study included coded data from outpatient appointments. We found that requiring a proximate visit to one of four possible sleep related specialties improved the PPV for both sleep disorders, while substantially reducing the number of cases identified. Whereas US hospitals may be motivated to accurately record these data for re-imbursement purposes, recording of diagnostic data in English HES outpatient data is not mandated by NHS England and therefore highly incomplete[3]. We were therefore unable to use this source to identify cases or validate cases identified in other sources.

A systematic review estimated a median PPV of 89 % (range 24–100 %) for 183 different diagnoses in CPRD primary care data (linked data were not available at the time)[18]. There was no clear pattern by ICD-10 Chapter and data for individual diagnoses or papers was not presented. Our estimated PPV for OSA using CPRD data only of 85.3 % (95 % CI 77.3 %−91.4 %) is close to the median whereas our estimated PPV for narcolepsy of 71.0 % (95 % CI 58.8–81.3 %) is lower than median but within the range.

A recent concordance study including linked cancer registration data in the gold standard algorithm estimated PPVs greater than 80 % for CPRD records for each of the 20 most common cancers[19]. PPVs were highest for common, clearly defined and well-understood cancers with higher survival rates; lack of familiarity with narcolepsy among GPs and the complexity of diagnosing the condition may therefore explain lower than average PPVs. In contrast to our findings, PPVs for cancer algorithms that included records in CPRD, HES or ONS mortality data were similar to CPRD only algorithms. Unlike our study, this concordance study did not rely on data transfer from hospitals to GP practices to measure gold standard cases. Our hypothesis that this biased the PPV downwards is further supported by a study by Winstone et al. [20]. This study reported high accuracy albeit low completeness of childhood narcolepsy cases recorded in English HES data based on a review of clinical notes and investigation records from specialist centres by three narcolepsy experts. Alternatively, incorrect HES records of narcolepsy

diagnoses may be less common in children.

Meaning of the study and future research

Incident diagnoses recorded in primary care data in England are highly accurate for OSA and reasonably accurate for narcolepsy. However, diagnosis dates are moderately accurate for OSA and poorly recorded for narcolepsy. Our findings suggest that including coded records from hospital admissions data reduces the accuracy of the recorded diagnosis; however this may be due to limitations associated with information transfer between hospitals and primary care. Requiring a proximate possible sleep related outpatient visit or an EDS drug prescription for narcolepsy increases the accuracy of recorded diagnoses substantially but removes a high proportion of gold standard cases. As UK primary care EHR systems are designed to be inter-operable these findings are likely to generalise to other EHR systems (e.g. SystmOne) and nations of the UK. Future research is needed to validate algorithms identifying a wider range of sleep disorders, including those commonly diagnosed in primary care such as insomnia[21] and restless legs syndrome[22].

Use of routinely collected data is often either the only way to study population based epidemiological questions, or a useful supplement to studies using more accurate prospectively collected clinical data from smaller less representative samples of the population. When using routinely collected data from England's NHS to study OSA or narcolepsy, epidemiologists should select the most appropriate sleep disorder algorithm for their study, perform sensitivity analyses using alternative algorithms, and transparently report limitations associated with measurement bias and data availability (e.g. incomplete information on OSA severity, type of narcolepsy and sleep disorder treatments).

The recently released Sudlow report provides recommendations for improving the UK health data landscape[3]. These recommendations include mandating the inclusion of diagnosis and procedural codes in national hospital outpatient episodes data, and enabling automated coding of unstructured information from electronic health records. We believe that both of these recommendations would substantially improve case completeness and our ability to accurately identify OSA and narcolepsy cases. We further recommend that specialist sleep centres agree on a common use of ICD-10 coding in HES outpatient data to record suspected and diagnosed cases; this would support future epidemiological research. Audits of information transferred from hospital specialists to primary care[23], and improving medical education about sleep disorders[24], may be used to improve recording in primary care and future automated coding of unstructured data. More accurate recording of sleep disorders in general practice would directly improve the quality and safety of care in addition to supporting impactful research[25].

Conclusion

Recorded diagnoses of two well-defined sleep disorders, OSA and narcolepsy, in routinely collected GP data are highly and reasonably accurate, respectively. Diagnosis dates are moderately accurate for OSA but poorly recorded for narcolepsy. Stricter algorithms using HES outpatient data and CPRD prescribing data (narcolepsy only) improve accuracy substantially, while missing high numbers of gold standard diagnosed cases. Epidemiologists should use the algorithm most suited to their study and include sensitivity analyses using alternative algorithms. We support recommendations to mandate recording of diagnosis codes in hospital outpatient episodes data and encourage automated coding of free text records; this is likely to improve case ascertainment, accuracy and characterisation.

Funding sources

Helen Strongman, NIHR Advanced Fellowship NIHR301730, is funded by the National Institute for Health and Care Research (NIHR) for this research project. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care. Sofia H. Eriksson is supported in part by the National Institute for Health and Care Research University College London Hospitals Biomedical Research Centre funding scheme. CWG is supported by a Wellcome Career Development Award (225868/Z/22/Z). KB is funded by a Wellcome Senior Research Fellowship (220283/Z/20/Z).

CRediT authorship contribution statement

Helen Strongman: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Sofia H. Eriksson:** Writing – review & editing, Methodology. **Kwabena Asare:** Writing – review & editing. **Michelle A. Miller:** Writing – review & editing, Methodology. **Martina Sýkorová:** Writing – review & editing, Visualization. **Hema Mistry:** Writing – review & editing, Methodology. **Kristin Veighey:** Writing – review & editing, Methodology. **Charlotte Warren-Gash:** Writing – review & editing, Methodology. **Krishnan Bhaskaran:** Writing – review & editing, Methodology, Formal analysis.

Declaration of competing interest

Helen Strongman volunteers as a Director & Trustee of Narcolepsy UK, a patient-lead support charity. Michelle A Miller is a voluntary elected member of the British Sleep Society Executive Committee.

Sofia H Eriksson has received honoraria for educational activities from Eisai, Fidia, Lincoln and UCB Pharma.

Hema Mistry is a member of the NIHR HTA General Funding Commissioning Committee.

KA, MS, KV, CW-G and KB have no conflicts of interest to declare.

Acknowledgements

This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. Hospital Episodes Statistics Data and Office of National Statistics Mortality data were reused with the permission of The Health & Social Care Information Centre (All rights reserved, Copyright (2021)). The data are provided by patients and collected by the NHS as part of their care and support. We would like to thank Harley Kennedy and Iryna Brooks for managing data collection at CPRD. The interpretation and conclusions based in this study are those of the authors alone.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.sleepe.2025.100110.

References

[1] The Lancet. Waking up to the importance of sleep. Lancet 2022 Sep 24;400 (10357):973.
[2] Hartley S, Goncalves M, Penzel T, Verbraecken J, Young P. Revised European guidelines for the accreditation of sleep medicine centres. J Sleep Res 2024;33(6): e14200.
[3] Sudlow C. Uniting the UK's health data: a huge opportunity for society [Internet]. Zenodo, https://zenodo.org/records/13353747; 2024.
[4] Edwards L, Pickett J, Ashcroft DM, Dambha-Miller H, Majeed A, Mallen C, et al. UK research data resources based on primary care electronic health records: review and summary for potential users. BJGP Open 2023;7(3):BJGPO.2023.0057.
[5] Matthewman J, Andresen K, Suffel A, Lin LY, Schultze A, Tazare J, et al. Checklist and guidance on creating codelists for routinely collected health data research [version 2; peer review: 3 approved]. NIHR Open Res 2024;4:20. Sep 18.

[6] Wolf A., Dedman D., Campbell J., Booth H., Lunn D., Chapman J., et al. Data resource profile: clinical practice research datalink (CPRD) aurum. Int J Epidemiol. 2019 Mar 11;48(6):1740–1740g.
[7] Herbert A, Wijlaars L, Zylbersztejn A, Cromwell D, Hardelid P. Data resource profile: hospital episode statistics admitted patient care (HES APC). Int J Epidemiol 2017;46(4):1093. Aug 1–1093i.
[8] Strongman H, Belot A, Bhaskaran K, Eriksson S, Leschziner G, Miller M, et al. Protocol: chronology of healthcare resource use and comorbidities in people with obstructive sleep apnoea and narcolepsy before and after diagnosis: a descriptive study. Real World Evid Regist 2022 [Internet]Aug 26 [cited 2023 Aug 16]; Available from: https://osf.io/f5ukw.
[9] Kumarapeli P, de Lusignan S. Using the computer in the clinical consultation; setting the stage, reviewing, recording, and taking actions: multi-channel video study. J Am Med Inf Assoc 2013;20(e1):e67–75. Jun 1.
[10] Tazare J, Henderson AD, Morley J, Blake HA, McDonald HI, Williamson EJ, et al. NHS national data opt-outs: trends and potential consequences for health data research. BJGP Open 2024;8(3). OctBJGPO.2024.0020.
[11] Padmanabhan S, Carty L, Cameron E, Ghosh RE, Williams R, Strongman H. Approach to record linkage of primary care data from clinical practice research datalink to other health-related patient data: overview and implications. Eur J Epidemiol 2019;34(1):91–9. Jan 15.
[12] Leahy TP, Sammon C, Ramagopalan S. Identification of an appropriate time window for the exclusion of prevalent cases recorded following registration with the clinical practice research datalink. In: Abstracts of the 36th International Conference on Pharmacoepidemiology & Therapeutic Risk Management, Virtual. Pharmacoepidemiology & Drug Safety; 2020. p. 2465. September 16–17 [Internet].
[13] Strongman H, Eriksson SH, Quinnell T, Mathur R, Lyons A, Wong A. Codelists for "Chronology of healthcare resource use and comorbidities in people with obstructive sleep apnoea and narcolepsy before and after diagnosis: a descriptive study" [internet]. London, United Kingdom: London School of Hygiene & Tropical Medicine; 2025 [cited 2025 Jul 17]. Available from: https://datacompass.lshtm.ac.uk/id/eprint/4742/.
[14] Strongman H. Validation of algorithms identifying diagnosed obstructive sleep apnoea and narcolepsy in coded primary care and linked hospital activity data in England: Github repository for Sleep Epidemiology Publication [internet]. Zenodo 2025 [cited 2025 Jul 21]. Available from: https://zenodo.org/records/16276052.
[15] Mansfield KE, Mathur R, Tazare J, Henderson AD, Mulick AR, Carreira H, et al. Indirect acute effects of the COVID-19 pandemic on physical and mental health in the UK: a population-based study. Lancet Digit Health 2021;3(4):e217–30. Apr 1.
[16] Jolley RJ, Liang Z, Peng M, Pendharkar SR, Tsai W, Chen G, et al. Identifying cases of sleep disorders through international classification of diseases (ICD) codes in administrative data. Int J Popul Data Sci 2018;3(1).
[17] Keenan BT, Kirchner HL, Veatch OJ, Borthwick KM, Davenport VA, Feemster JC, et al. Multisite validation of a simple electronic health record algorithm for identifying diagnosed obstructive sleep apnea. J Clin Sleep Med 2020;16(2):175–83.
[18] Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the general practice research database: a systematic review. Br J Clin Pharmacol 2010;69(1):4–14. Jan.
[19] Strongman H, Williams R, Bhaskaran K. What are the implications of using individual and combined sources of routinely collected data to identify and characterise incident site-specific cancers? A concordance and validation study using linked english electronic health records data. BMJ Open 2020;10(8):e037719. Aug 20.
[20] Winstone AM, Stellitano L, Verity C, Andrews N, Miller E, Stowe J, et al. Clinical features of narcolepsy in children vaccinated with AS03 adjuvanted pandemic A/H1N1 2009 influenza vaccine in England. Dev Med Child Neurol 2014;56(11):1117–23. Nov.
[21] De Lange MA, Richmond RC, Eastwood SV, Davies NM. Insomnia symptom prevalence in England: a comparison of cross-sectional self-reported data and primary care records in the UK Biobank. BMJ Open 2024;14(5):e080479. May 1.
[22] van de Vijver DAMC, Walley T, Petri H. Epidemiology of restless legs syndrome as diagnosed in UK primary care. Sleep Med 2004;5(5):435–40. Sep 1.
[23] Varyani F, Card T, Kaye P, Aithal GP, West J. The communication of a secondary care diagnosis of autoimmune hepatitis to primary care practitioners: a population-based study. BMC Health Serv Res 2013;13(1):161. May 1.
[24] Romiszewski S, May FEK, Homan EJ, Norris B, Miller MA, Zeman A. Medical student education in sleep and its disorders is still meagre 20 years on: a cross-sectional survey of UK undergraduate medical education. J Sleep Res 2020;29(6):e12980.
[25] NHS England. NHS England » purpose of the GP electronic health record [internet] [cited 2025 July 18], https://www.england.nhs.uk/long-read/purpose-of-the-gp-electronic-health-record/; 2025.