Performance of French medico-administrative databases in epidemiology of infectious diseases: a scoping review

Marc-Florent Tassi, Nolwenn Le Meur, Karl Stéfic, Leslie Grammatico-Guillon

▶ **To cite this version:**

Marc-Florent Tassi, Nolwenn Le Meur, Karl Stéfic, Leslie Grammatico-Guillon. Performance of French medico-administrative databases in epidemiology of infectious diseases: a scoping review. Frontiers in Public Health, 2023, 11, 10.3389/fpubh.2023.1161550 . hal-04114348

HAL Id: hal-04114348
https://hal.ehesp.fr/hal-04114348

Submitted on 1 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

frontiers | Frontiers in Public Health

TYPE Systematic Review
PUBLISHED 12 May 2023
DOI 10.3389/fpubh.2023.1161550

Check for updates

OPEN ACCESS

EDITED BY
Jian Wu,
Suzhou Municipal Hospital, China

REVIEWED BY
Marco Dettori,
University of Sassari, Italy
Pasquale Stefanizzi,
University of Bari Aldo Moro, Italy

*CORRESPONDENCE
Marc-Florent Tassi
✉ marc.tassi@etu.univ-tours.fr

RECEIVED 08 February 2023
ACCEPTED 17 April 2023
PUBLISHED 12 May 2023

CITATION
Tassi M-F, le Meur N, Stéfic K and Grammatico-Guillon L (2023) Performance of French medico-administrative databases in epidemiology of infectious diseases: a scoping review. *Front. Public Health* 11:1161550. doi: 10.3389/fpubh.2023.1161550

COPYRIGHT
© 2023 Tassi, le Meur, Stéfic and Grammatico-Guillon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Performance of French medico-administrative databases in epidemiology of infectious diseases: a scoping review

Marc-Florent Tassi[1]*, Nolwenn le Meur[2], Karl Stéfic[1,3] and Leslie Grammatico-Guillon[1,4]

[1]INSERM U1259, Université de Tours, Tours, France, [2]Univ Rennes, EHESP, CNRS, Inserm, Arènes-UMR 6051, RSMS-U 1309, Rennes, France, [3]Laboratoire de virologie et CNR VIH-Laboratoire associé, CHRU de Tours, Tours, France, [4]Service d'Information Médicale d'Epidémiologie et d'Economie de la Santé, CHRU de Tours, Tours, France

The development of medico-administrative databases over the last few decades has led to an evolution and to a significant production of epidemiological studies on infectious diseases based on retrospective medical data and consumption of care. This new form of epidemiological research faces numerous methodological challenges, among which the assessment of the validity of targeting algorithm. We conducted a scoping review of studies that undertook an estimation of the completeness and validity of French medico-administrative databases for infectious disease epidemiological research. Nineteen validation studies and nine capture-recapture studies were identified. These studies covered 20 infectious diseases and were mostly based on the evaluation of hospital claimed data. The evaluation of their methodological qualities highlighted the difficulties associated with these types of research, particularly those linked to the assessment of their underlying hypotheses. We recall several recommendations relating to the problems addressed, which should contribute to the quality of future evaluation studies based on medico-administrative data and consequently to the quality of the epidemiological indicators produced from these information systems.

KEYWORDS

infectious diseases, medico-administrative data, SNDS, validation studies, capture recapture studies, scoping review

1. Introduction

Major epidemics of the last thirty years, such as HIV, Ebola, or SARS, have brought epidemiology back to its historical link with infectious diseases (1–3). The recent pandemic of COVID-19 has dramatically increased this phenomenon (4–6). Indeed, a simple search on PubMed for "epidemiology" and "infection" returned an annual average of 37,600 references for the years 2015 to 2019 vs. 72,800 for 2020 and 2021.

From a methodological perspective, the last decades have also strongly affected epidemiological research. The rise of Big Data during the information age has notably materialized in the health sector through the progressive implementation of medico-administrative information systems. These data warehouses may contain various medical and demographic information but share the common feature of being fed in a passive, regular and sustainable way for administrative and financial management purposes (7, 8). In France, public health care agencies collect and monitor health care expenses using two principal data warehouses. The nature and scope of the information they contain has

already been extensively detailed (9, 10). Briefly, the *Datamart de Consommation Inter Régime* (DCIR), gathers information on primary care expenditures whereas the *Programme de Médicalisation des Systèmes d'Information* (PMSI), relates to hospital care including information extracted from anonymous discharge summaries. The *Système National des Données de Santé* (SNDS) was conceived to host and link these two data warehouses so that they can be used jointly for research purposes. Although these information systems were initially developed for financial management purposes, their content in medical data covering 99% of the French population associated with significant historical depth has made them increasingly valuable as data sources for public health research and in particular epidemiology (11).

The growing success of administrative medical databases for research purposes should not blind researchers to the fact that these sources of information have many limitations that are likely to cause significant bias (7, 12–14). A major concern is the lack of clinical and biological information available which sometimes limits the accuracy and questions the reliability of the information used to identify population of interest. To overcome this limitation, researchers develop algorithms of varying complexity that aim to minimize patient misclassification, whether in terms of inclusion criteria, exposure, comorbidities, or outcome.

In France, users of French medico-administrative databases have formed the REDSIAM network to mutualise expertise concerning the development and evaluation of algorithms for epidemiological purposes (15). In 2017, its "infectious diseases" working group published a narrative review on infections studied through French medico-administrative databases and on the characteristics of the algorithms developed and/or used to identify patients with these infections (16). However, the performances of these algorithms remain a major concern for epidemiological studies based these databases as only few have reported a validation process using a gold standard and their methodological process was never evaluated.

To assess both the validity of French medico-administrative databases for epidemiological purposes in infectious diseases research and the methodological quality of studies that conducted validation of infectious diseases identification algorithms, we undertook a scoping review with the following specific objectives: (1) identify topics where efforts have been made to assess the completeness and validity of these databases; (2) identify and describe the methods and resources used to carry out these validations.

2. Methods

We undertook this review based on the PRISMA extension for Scoping Review (PRISMA-ScR) guidelines (17). The protocol of this study was not registered.

2.1. Types of study considered in this review

To address the objectives of this review, two methodological frameworks were considered: validation study and capture-recapture study.

Validation studies are commonly used in medical sciences to evaluate the predictive ability of screening and diagnostic procedures by comparing the predictions of these tests to a reference classification. Since diagnostic tests can be conceptually assimilated to classification algorithms, the methodology used to evaluate them can be transposed in a quasi-identical manner to the analysis of the performance of disease targeting algorithms in medical-administrative databases.

The capture-recapture method was originally developed in the field of ecology to estimate the size of animal populations. Adapted to epidemiology, its principle is to cross-reference several databases derived from the same population and containing information on diseased individuals in order to identify common cases. Using the number of cases reported by each source and the number of common cases, it is possible under certain conditions to estimate the total number of affected individuals in the source population and thus the completeness of each database. If one of the databases involved is considered to be exhaustive and the identification of cases within it is based on an algorithmic procedure, then the estimate of its completeness derived from the capture-recapture procedure can be interpreted as the sensitivity of the targeting algorithm.

2.2. Search strategy

We searched PubMed, Embase and Web of Science for articles published in English or French up to the end of 2021. To identify relevant studies, our search strategy consisted in associating the concepts "infectious disease" and "French medico-administrative database" using the Boolean operator "AND" (Supplementary 1). Since the terminology used to refer to French medico-administrative databases is not always explicit, we used a broad search lexicon to ensure the identification of studies as complete as possible (18). We also searched for articles in the documentary databases of three French institutions routinely using the SNDS (*Assurance Maladie*, *Santé publique France*, *EPI-PHARE*) (19–21).

The concepts of algorithm validation and database completeness assessment were not integrated into the search algorithm to avoid missing studies where these would have consisted in secondary objectives.

2.3. Studies selection

To be considered for inclusion in the study, articles had to satisfy four criteria: (1) to follow the "Introduction, Methods, Results, and Discussion" structure, (2) inclusion criteria and/or the main objective of the study directly related to an infectious disease and/or an anti-infective agent, (3) data used had to originate at least partly from a French medico-administrative database, (4) the study had to include at least one evaluative aspect either related to the completeness of the information sources via a capture-recapture method; or related to the performances of the algorithm employed to define the infectious disease and/or the anti-infective drug of interest.

For validation studies, we only considered research comparing medico-administrative data to a reference standard at the individual-patient level. Ecological validations (i.e., comparisons of aggregate statistics across studies) were exclude as they do not allow the calculation of algorithms accuracy indicators and carry too high risk of bias (22).

Using Google Scholar, we reviewed the bibliographies and citations of the articles that satisfied the first three inclusion criteria to find additional research papers of interest.

Abstracts were excluded from analysis.

2.4. Data extraction and quality assessment

For studies meeting at least the first three inclusion criteria, we used a standardized abstraction form to describe the research scope and methods: (1) condition of interest (i.e., nature of the infectious disease(s) and/or anti-infective treatment studied), (2) data source(s), (3) year of publication, (4) types of information used in the main algorithm (i.e., algorithm targeting the condition of interest), (5) geographical scope, (6) years of study, (7) number of other health conditions targeted by an algorithm, (8) whether the article described the main algorithm in a reproducible way.

For validation studies, additional information collected were: (1) nature of reference standard, (2) recruitment criteria for the validation sample, (3) sample size, (4) study design, (5) performance parameters.

Based on the works by Benchimol et al. (23) and Widdifield et al. (24) we used a 35-items checklist to evaluate the quality of reported information for research identified as validation studies (23, 24). For each study, the expected number of items to be carried forward was calculated excluding uncertain and non-applicable items.

For studies using capture-recapture methods, we also collected: (1) complementary sources of cases used, (2) matching strategy between sources of information, (3) completeness estimator used, (4) whether the study has undertaken an assessment of the method's assumptions, (5) completeness estimate and its 95% confidence interval given in the manuscript or recalculated from the available data.

2.5. Performance indicators definitions

In the epidemiological area, sensitivity, specificity, positive predictive value, and negative predictive value are the four indicators commonly presented to describe the performance of a targeting algorithm.

Positive predictive value (PPV) informs us about the capacity of the algorithm to avoid the generation of falsely positive individuals among positive individuals and thus to discriminate only true cases. Negative predictive value (NPV), the counterpart of PPV, characterizes the capacity of the algorithm to avoid generating false negatives among negative subjects and therefore to discriminate individuals who are free of the disease. Sensitivity (Se) informs us about the ability of the algorithm to avoid the generation of false negatives among infected persons and therefore to identify all the cases. Specificity (Sp), counterpart of Se, characterizes the capacity of the algorithm to avoid generating false positive among non-infected subjects and thus informs us about the ability of the algorithm to identify all disease-free subjects.

Other performance indicators may be reported in epidemiological study such as likelihood ratio (25). Positive likelihood ratio is the ratio of the probability that the algorithm classifies a diseased person as positive (Se) to the probability that it classifies a disease-free person as positive (1-Sp). In contrast, the negative likelihood ratio is the ratio of the probability that the algorithm classifies a diseased person as negative (1-Se) to the probability that it classifies a disease-free person as negative (Sp).

3. Results

3.1. Studies selection

The methodical search resulted in the identification of 204 distinct studies. The analysis of their bibliographic references allowed the finding of 37 additional articles (Figure 1) among which five were not referenced by medical literature databases and four mentioned the infectious concept in their abstract only in very specific terms (abscess, dengue, malaria, and gastroenteritis). For the 28 other studies, the abstract made no mention to medico-administrative databases or referred to them with unusual terms.

From these 241 studies fulfilling the first three inclusion criteria (Supplementary 2), we finally identified 19 studies that evaluated the quality of definition algorithms and nine studies that estimated the completeness of French medico-administrative databases using a capture-recapture method. These studies looked at 16 different infectious diseases or infectious concepts including bone and joints infections, nosocomial infections, endocarditis, pneumonia, influenza and bronchiolitis, mucormycosis, herpetic and meningococcal invasive infections, gastro-enteritis and *Clostridium difficile*, urinary tract infections, hantavirus, malaria and dengue (Tables 1, 3).

3.2. Validation studies

3.2.1. Methodological framework

For 17 of the 19 studies identified, validation covered in-hospital events based on algorithms built from the hospital discharge database (PMSI) alone (Table 1). Only two studies used primary care reimbursement data (DCIR) to evaluate the performance of an algorithm identifying cases of medicalised acute gastroenteritis (26, 27).

Almost all the studies based on the PMSI (16 out of 17) used medical expert reviewing of patients' hospital records as gold standard. We also identified other methods used to constitute the gold standard, based on data from hospital microbiology laboratories or on data from a nosocomial infection control center. In the two studies that investigated primary care reimbursement data, the reference data was based on patients' self-report of their treatment indication (26, 27).

The inter-quantile range (IQR) of validation sample sizes varied from 193 to 1,028 individuals. One study was able to

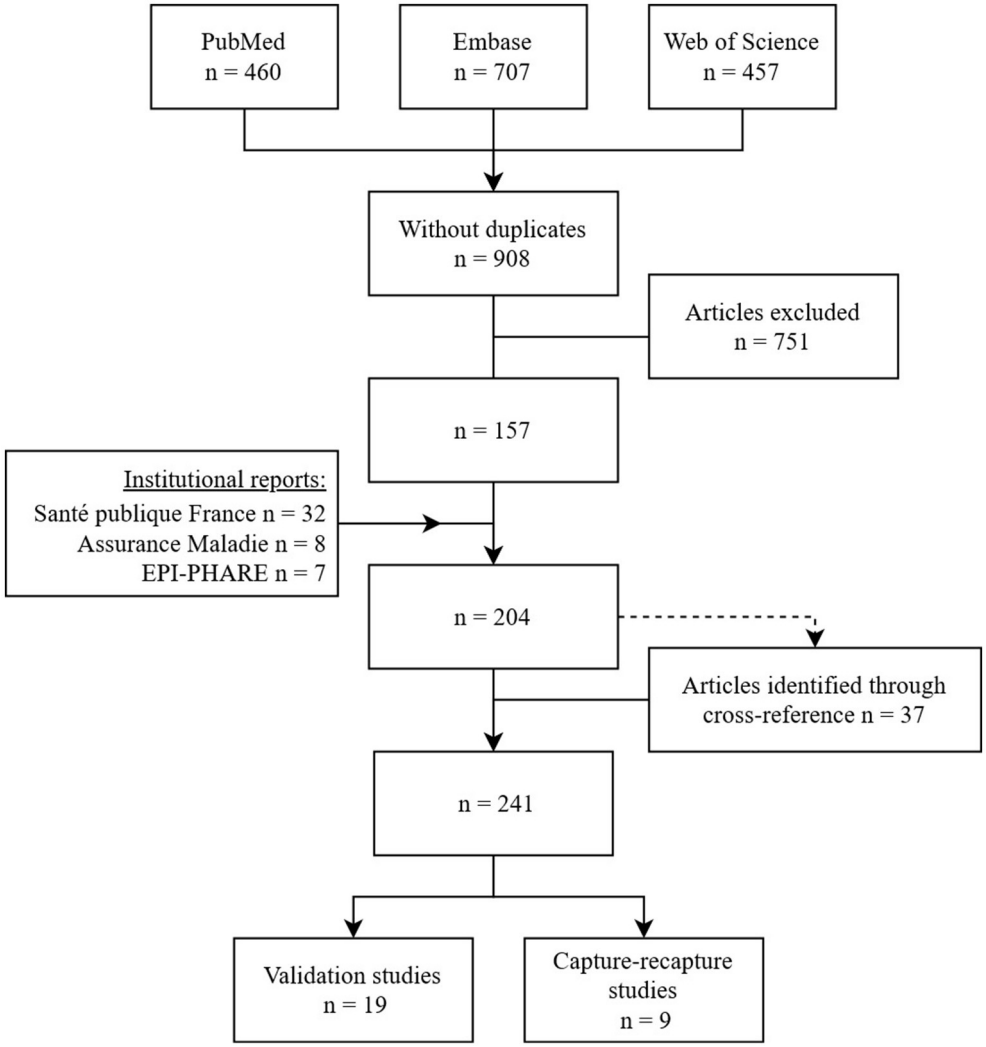

FIGURE 1
Flow diagram of articles selection studying epidemiology of infectious diseases using French medico-administrative databases published in English or French up to the end of 2021.

include 4,400 patients using data routinely collected by a hospital infection control team as the gold standard (28). One study used microbiology data as reference and was able to include up to 317,033 patients (29).

For each study, a median of three performance indicators were reported (IQR: 2–4). PPV was the most frequently reported indicator (17 out of 19 studies). NPV was reported in nine studies. Se of the algorithm was described in 14 studies and Sp was reported for 12 studies. Two other less common performance indicators were also identified. One study reported positive likelihood ratio, and two studies calculated the kappa coefficient, which is an indicator of the concordance between the classification made by the algorithm and the classification made by the gold standard (29, 30).

Based on the inclusion criteria of validation samples and according to the procedure steps of patients' classification into these samples, we determined four methodological approaches used to conduct the different studies (Figure 2).

Eleven studies used approaches 1 and 2 in which subjects were included and sampled independently of the classification results produced by the algorithm being evaluated. These first two approaches only differed by the order in which the algorithm and gold-standard classifications were applied. However, in seven of the 11 studies, the lack of clear indications about classification steps did not allow us to identify which of these two approaches has been applied. The calculation of all the principal performance parameters of the algorithm (Se, Sp, PPV, and NPV) is theoretically possible for both approaches assuming the sampling modes preserve the prevalence of the disease with respect to the inclusion criteria.

In approach 3A, used by four studies, sampling was carried out according to the result of the classification algorithm. This approach results in the constitution of two samples (patients

TABLE 1 Characteristics of validation studies.

Study	Subject	Database	Algorithm evaluated	Reference standard	Source population criteria	Sample size	Approach[a]	Number of algorithms evaluated	Measures of accuracy
Bitar et al. (37)	Mucormycosis	PMSI	ICD-10 codes	Medical expert review	Positive algorithm classification	179	3B	1	PPV = 0.36
Bounoure et al. (26)	mAGE	DCIR	Dispensed drugs + time lag between consultation and dispensation + patient's age	Patient's declaration	Targeted drug in the prescription	557	1	1	Se = 0.89 Sp = 0.89
Bounoure et al. (27)	mAGE	DCIR	Dispensed drugs + time lag between consultation and dispensation + patient's age	Patient's declaration	Targeted drug in the prescription	1 308	1	1	Se = 0.9 PPV = 0.82
de Lafforest et al. (59)	Urinary tract infections	PMSI	ICD-10 codes	Medical expert review	Algorithm classification, hospitalization unit	1 122	3A	1	Se = 0.95 Sp = 0.76 PPV = 0.70 NPV = 0.98
Dely et al. (25)	Preventable readmissions of CAP	PMSI	ICD-10 codes + time lag between hospitalisations + type of hospitalization admission	Medical expert review	Targeted ICD-10 code	415	2	5	Se = 0.31-0.5 Sp = 0.95-1 PPV = 0.36-0.66 PLR = 8.2-308
Gerbier et al. (60)	Nosocomial infections	PMSI	ICD-10 codes	Medical expert review	Surgical procedure	446	1 or 2	2	Se = 0.26-0.79 Sp = 0.66-1 PPV = 0.18-0.83 NPV = 0.94-0.97
				Reports to the center for the control of nosocomial infections	Stay in intensive care unit	1 499		4	Se = 0-0.59 Sp = 0.87 PPV = 0.09 NPV = 0.98
				Medical expert review and reports to the center for the control of nosocomial infections	Delivery in an obstetric unit	1 081		1	Se = 0.43 Sp = 0.786-1 PPV = 0-0.36 NPV = 0.88-0.98
Grammatico-Guillon et al. (61)	Vertebral osteomyelitis	PMSI	ICD-10 codes	Medical expert review	Positive algorithm classification	90	3B	1	PPV = 0.94
Grammatico-Guillon et al. (62)	Pneumococcal pneumonia	PMSI	ICD-10 codes	Medical expert review	Positive algorithm classification	45	3B	1	PPV = 0.82
				Laboratory results	positive pneumococcal sample	54	1 or 2		Se = 0.33 Sp = 1

(Continued)

10.3389/fpubh.2023.1161550

TABLE 1 (Continued)

Study	Subject	Database	Algorithm evaluated	Reference standard	Source population criteria	Sample size	Approach[a]	Number of algorithms evaluated	Measures of accuracy
Grammatico-Guillon et al. (63)	BJI	PMSI	ICD-10 codes + surgical procedure	Medical expert review	Positive algorithm classification	100	3B	1	PPV = 0.84
					Surgical procedure	205	1 or 2		Se = 0.95 Sp = 0.99 PPV = 0.98 NPV = 0.99
Grammatico-Guillon et al. (64)	Pediatric BJI	PMSI	ICD-10 codes + surgical procedure	Medical expert review	Algorithm classification, orthopedic fracture	398	3A	1	Se = 1 Sp = 0.8 PPV = 0.81 NPV = 1
Grammatico-Guillon et al. (65)	HKAI	PMSI	ICD-10 codes + surgical procedure	Medical expert review	Algorithm classification	1 010	3A	3	Se = 0.97-098 Sp = 0.71-0.95 PPV = 0.63-0.87 NPV = 0.98-0.99
Jones et al. (29)	Clostridium difficile infection	PMSI	ICD-10 codes	Laboratory results	Hospitalization	317 033	1 or 2	1	Se = 0.36 Sp = 1 PPV = 0.79 NPV = 1 κ = 0.49
Jouan et al. (66)	Herpes simplex encephalitis	PMSI	ICD-10 codes	Medical expert review	Algorithm classification, infection with neurological involvement	226	3A	1	PPV = 1 NPV = 1
Leclère et al. (28)	SSI	PMSI	ICD-10 codes + surgical procedure	Surveillance by the infection control team	Surgical procedure	4400	1 or 2	3	Se = 0.24-0.25 Sp = 0.98 PPV = 0.06-0.25 NPV = 0.98-1
Sahli et al. (67)	Various infections	PMSI	ICD-10 codes	Medical expert review	Positive algorithm classification	200	3B	2	PPV = 0.70-0.97
Soilly et al. (30)	RSV Bronchiolitis	PMSI	ICD-10 codes	Medical expert declaration	Hospitalization for bronchiolitis	302	1	1	Se = 0.55 Sp = 0.65 κ = 0.1
Sunder et al. (68)	Infective endocarditis	PMSI	ICD-10 codes	Medical expert review	Positive algorithm classification	198	3B	1	PPV = 0.87
					Surgical procedure	492	1 or 2		Se = 0.90 Sp = 1 PPV = 1 NPV = 0.99
Sunder et al. (69)	Infective endocarditis	PMSI	ICD-10 codes	Medical expert review	Positive algorithm classification	388	3B	1	PPV = 0.86
Tubiana et al. (70)	Oral streptococcal infective endocarditis	PMSI	ICD-10 codes	Medical expert review	Positive blood culture result for oral streptococci	130	1 or 2	1	Se = 0.54 PPV = 1

[a]see Figure 2 for a description of the different approaches. BJI, bone and joint infection; CAP, community-acquired pneumonia; DCIR, datamart de consummation inter-régime (national primary care consumption database); HKAI, hip or knee arthroplasty related infection; ICD-10, International Classification of Diseases 10th Revision; mAGE, medicalised acute gastroenteritis; NPV, negative predicted value; PLR, positive likelihood ratio; PMSI, programme de médicalisation des systems d'information (national hospital discharge summary database); PPV, positive predicted value; RSV, respiratory syncytial virus; Se, sensitivity; Sp, specificity; SSI, surgical site infection.

10.3389/fpubh.2023.1161550

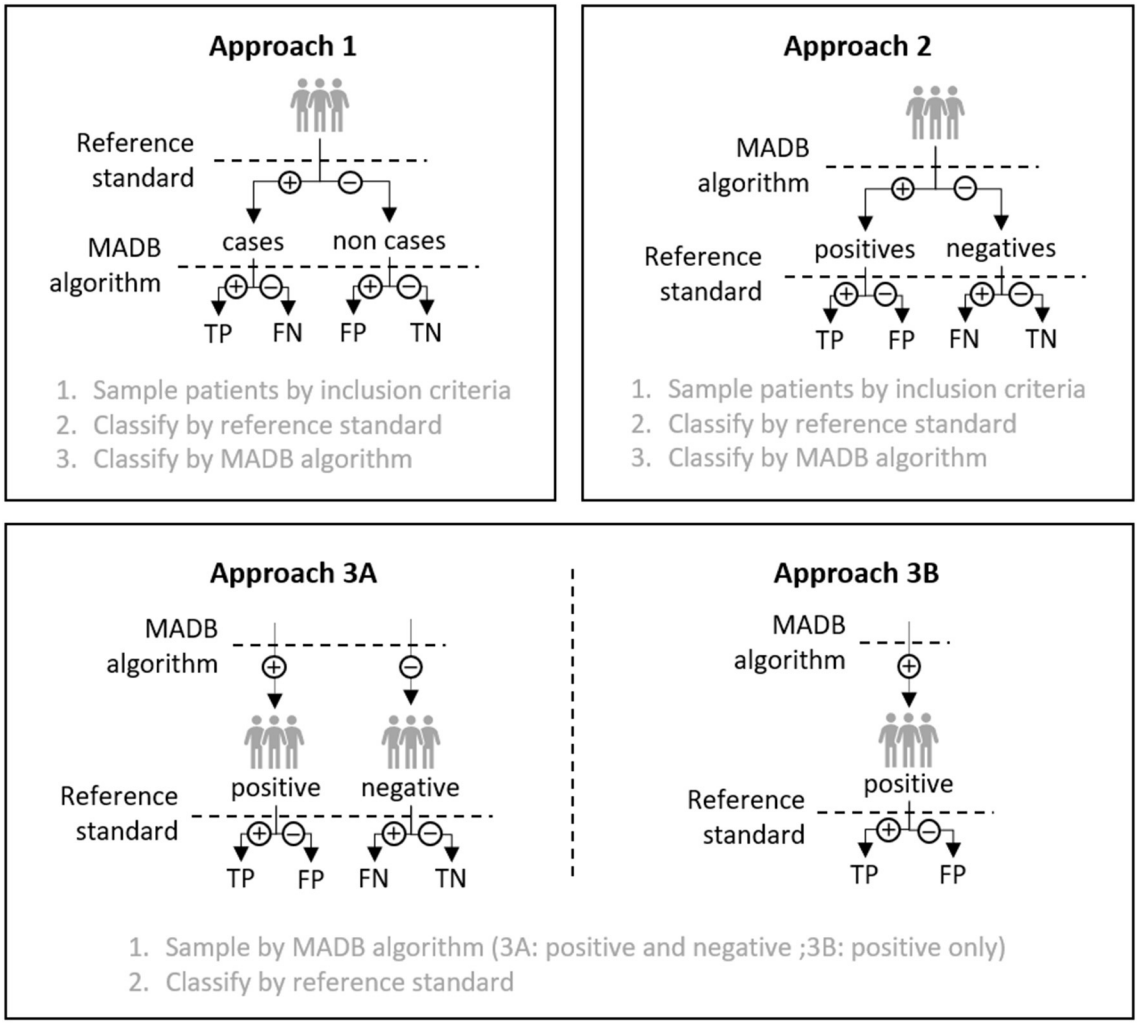

FIGURE 2
Approaches used to perform validation studies.

classified as positive and patients classified as negative by the algorithm) whose sizes were chosen arbitrarily. This approach allows the calculation of positive and negative predictive values of the algorithm but should not be used to calculate Se and Sp (31) Indeed, the arbitrary choice of the number of positive and negative patients in the validation sample is likely to distort the "natural" ratio of these marginal distributions, leading to a bias in the estimation of Se and Sp parameters (see Supplementary 3). The approach 3B, used in seven studies, differs in that only a group of positive patients is selected. Thus, it only allows an evaluation of the PPV of the algorithm.

We did not identify any studies that constituted a validation sample by selecting an arbitrary number of cases and non-cases as defined by gold standard classification. This method leads to the same type of issue as methods 3A and 3B since it leads to a distortion of the natural ratio of the disease prevalence. Thus, this method would not allow the calculation of the predictive values of the algorithm but only of its Se and Sp.

3.2.2. Quality of reported information

For all the studies (n = 19), out of the 35-items checklist a median of 56% of the expected items were reported (Q1-Q3, 40-65%) (Table 2). These statistics varied with the objectives of each study. For the 8 studies with algorithm validation as primary objective, a median of 66% of items were reported (Q1–Q3, 62–8%), compared with a median of 40% of items reported (Q1–Q3, 28–49%) for the studies where validation was a secondary objective.

Almost two third of the studies were identified as validation studies of medico-administrative databases (12 out of 19). In all studies, the data sources referred to the medico-administrative data

TABLE 2 Evaluation of validation studies [35-items checklist based on the works by Benchimol et al. (23) and Widdifield et al. (24)].

Section, criteria	Yes	No	Uncertain	NA
Title, Keywords, Abstract				
1 - Identifies article as study of assessing diagnostic accuracy	13	6		
2 - Identifies article as study of administrative data	18	1		
Introduction				
3 - States disease identification and validation as one of the goals of study	11	8		
Methods				
4 - Describes data sources	13	6		
5 - Describes medico-administrative algorithm	19	0		
6 - Describes inclusion/exclusion criteria	19	0		
7 - Reports enrolment dates	14	5		
8 - Describes sampling method	11	8		
9 - Describe data collection for the reference standard	11	8		
10 - Use of a split sample for revalidation	0	19		
11 - Describes number, training or expertise of persons reading reference standard	10	9		
12 - Reports a measure of concordance if > 1 persons reading the reference standards	1	5		13
13 - Readers of the reference standard were blinded to the results of the classification by administrative data	1	6	12	
14 - Describes explicit methods for calculating or comparing measures of diagnostic accuracy and the statistical methods used to quantify uncertainty	4	15		
Results				
Sample constitution				
15 - Reports number of patient satisfying inclusion/exclusion criteria	19	0		
16 - Reports study flow diagram	3	16		
17 - If patients are sampled by reference standard, reports the number of records unable to link	1	0		18
18 - Reports number of missing/incomplete medical records and/or the number of patients unwilling to participate	6	13		
Reports clinical and demographic characteristics of the validation sample				
19 - Age	4	15		
20 - Sex	3	16		
21 - Comorbid conditions	2	17		
Test results				
22 - Presents cross tabulation of the results of the index tests by the results of the reference standard	7	12		
23 - Reports explicit pretest prevalence in the validation sample	6	9		4
24 - Describes the characteristics of misclassified patients (false-positive and/or false-negative)	6	13		
25 - Reports results for any subgroup (age, comorbidity, sex, location…)	6	13		
Estimates				
26 - Sensitivity	14	5		
27 - Specificity	12	7		
28 - Positive predictive value	17	2		
29 - Negative predictive value	9	10		

(Continued)

TABLE 2 (Continued)

Section, criteria	Yes	No	Uncertain	NA
30 - Likelihood ratio	1	18		
31 - Kappa	2	17		
32 - Reports 95% confidence intervals	8	11		
33 - If PPV/NPV reported, number of case and control is not arbitrary defined?	16	0	1	2
34 – If Se/Sp reported, number of positive and negative is not arbitrary defined?	10	3	1	5
Discussion				
35 - Discusses the applicability of the study findings	18	1		

NA, not applicable; NPV, negative predicted value; PPV, positive predicted value; Se, sensitivity; Sp, specificity.

warehouse used, namely PMSI or DCIR. However, only 13 of the 19 studies identified and described the data sources used to constitute their gold standard.

All the studies reported a sufficiently detailed description of the algorithm evaluated and inclusion criteria for the validation sample to allow replication. In contrast, only half of the studies clearly stated the number and status of the persons who carried out the reference classification, and only one study explicitly stated that evaluators were blind to the results of the algorithm classification. Regarding the validation sample, only five studies reported at least one characteristic (age, sex, or comorbidity). Similarly, only six studies reported the specific characteristics of individuals misclassified by the algorithm. Eventually, only six studies reported having undertaken subgroup analyses to assess the influence of individual characteristics on the performance of the algorithm.

3.3. Capture-recapture studies

The nine studies that used a capture-recapture method were all based on PMSI data (Table 3). Except for one study using two complementary sources of information, all these studies matched PMSI data to a single other database.

These additional information sources can be categorized into four groups according to their origin and content. Data from the national register of death certificates (*CépiDC*) were used to estimate the number of deaths attributable to malaria and mucormycosis (32, 33). Two regional studies used mandatory declarations of meningococcal infection to adjust the disease incidence (34, 35). Three studies on mucormycosis, malaria, and haemorrhagic fevers with renal syndrome obtained clinical and biological data from the national reference centers (*Centers Nationaux de Réference*, CNR) related to these infections (32, 36, 37). Finally, two studies on severe influenza cases and one study on dengue fever benefited from the surveillance systems set up by the French public health agency (*Santé publique France*, SpF) (38–40).

To identify common patients between the PMSI and the complementary data sources, combinations of indirectly identifying data were used. Most frequently age, sex, and place of residence of the patients were used, along with location and dates of hospitalization. A single study included direct identifying data (surname and first name) (37).

The PMSI completeness was assessed by dividing the number of cases identified in this database by the total number of cases estimated using the capture-recapture method. This indicator varied according to the studied infection. For fever with renal syndrome and deaths related to mucormycosis the completeness was estimated at 37 and 43%, respectively (33, 36) In contrast, it was estimated that the PMSI recorded 82% of dengue cases on the Réunion island (40).

Of the nine studies reviewed, only four reported an assessment of assumptions involved in the capture-recapture method (validity of the case definition, completeness of case matching across sources, independence of sources and homogeneity of capture).

Validity of case definition was examined for mucormycosis by reviewing patient records and for malaria-related deaths by reviewing standardized hospital discharge summaries (32, 37). For severe influenza cases, sensitivity analysis with different VPP of the algorithm were carried out to investigate the impact of its validity on the estimation of the total number of cases (38).

The assumption of homogeneity of capture within each group implies that case identification within each source of information is equiprobable for all individuals regardless of their individual characteristics. To verify this hypothesis, one study on deaths associated with malaria and one study on severe influenza cases stratified their analysis according to three potential factors of heterogeneity: sex, age, and place of death for malaria and season, age, and place of residence for influenza (32, 38). Another study on invasive meningococcal infection also attempted to evaluate this hypothesis by comparing the distributions of cases between the two sources according to age and place of residence (34).

By linking three sources of information, the study on malaria-related deaths proposed to assess the dependency between these data sources by using log-linear models incorporating interaction terms between the different sources (32).

4. Discussion

This literature review identified epidemiological research on infectious diseases based on French medico-administrative databases, where validation efforts have been made, as well as the methods employed to that end. Showing strengths and

TABLE 3 Characteristics of capture-recapture studies.

Study	Subject	Database	Other sources	Matching strategy	Exhaustivity estimator	Database exhaustivity, 95%-CI	Assumption verifications
Belchior et al. (36)	Hantavirus haemorrhagic fever with renal syndrome	PMSI	NRC for haemorrhagic fever	Age, sex, hospitalization year, place of residence	Chapman	0.37 (0.34-0.41)	NA
Bitar et al. (33)	Mucormycosis (deaths)	PMSI	Death certificates	NA	Sekar	0.43 (NA)	NA
Bitar et al. (37)	Mucormycosis	PMSI	NRC for invasive mycosis	Name, age, sex, hospitalization date	Chapman	0.52 (0.45-0.63)	Case definition
Dubos et al. (34)	Meningococcal invasive infection	PMSI	mandatory reporting database	Age, sex, place of residence, hospital identity, date of infection	Sekar or Chapman	0.73 (0.71-0.74)	Homogeneity
Kendjo et al. (32)	Malaria (deaths)	PMSI	Death certificates and NRC for malaria	Sex, date of death, place of death, age at death	Log-linear model	0.64 (0.60-0.67)	Case definition, Independence, Homogeneity
Loury et al. (38)	Severe influenza	PMSI	SpF surveillance system	Hospital identity, admission date (-1 day to + 7 days tolerance), sex, age (± 1 year tolerance)	Chapman	0.73 (0.72-0.74)	Homogeneity, Case definition (sensitivity analysis)
Molinié et al. (35)	Meningococcal invasive infection	PMSI	mandatory reporting database	Sex, age, place of residence, place of hospitalization, date of infection	Chapman	0.74 (0.69-0.79)	NA
Pivette et al. (39)	Severe influenza	PMSI	SpF surveillance system	Hospital identity, admission date (-1 day to + 7 days tolerance), sex, age (± 1 year tolerance)	Chapman	0.78 (0.77-0.79)	NA
Verrier et al. (40)	Dengue fever	PMSI	SpF surveillance system	Age (± 1 year tolerance), sex, place of residence, hospitalization place and date	Chapman	0.82 (0.78-0.86)	NA

PMSI, programme de médicalisation des systèmes d'information (national hospital discharge database); NRC, national reference center; SpF, Santé publique France (French public health institute); NA, not available.

limits of these approaches, this scoping review highlighted that both design of the validation study and characteristics of the validation sample are crucial to the quality of estimation of the algorithmic performance.

4.1. Validation studies

This review first highlights that to date, only a small fraction (8%) of the studies on infectious diseases based on medico-administrative data undertook to evaluate their infectious diseases definition algorithms. Hence, most infectious diseases commonly studied using medico-administrative data (e.g., influenza, COVID-19, meningitis, or HIV), do not yet have any algorithm assessed and validated.

The second finding of this review is that most studies focused exclusively on hospital discharge data (PMSI). Only two studies evaluating the performance of a gastroenteritis detection algorithm were based on primary care data (DCIR) (26, 27). Moreover, no validation study using both data sources jointly was identified. This observation could be explained by the relative simplicity of building a gold standard for evaluating an algorithm based on PMSI data alone. Indeed, as PMSI data are generated locally by each hospital, it is relatively straightforward for in-hospital experts to link patients' PMSI data with their clinical and biological data archived within the institution. However, validation of PMSI data alone is largely insufficient. By the end of 2021, research on infectious diseases based solely on PMSI represented only half of the studies published to date. This proportion is expected to decrease over time due to the expansion of access to SNDS beyond national health agencies to academic research.

The third lesson of this review is the importance of the validation study designs. Choosing a particular design is far from trivial since it determines the type of indicators that can be estimated and whether they can be generalized. Since post-test probabilities (PPV and NPV) are conditioned by the prevalence of the disease, they can only be estimated when the sampling of the validation panel is not based on the gold standard classification results. Indeed, if in the validation sample, the proportion of infected subjects (i.e., prevalence of infection) is higher than in

the source population, PPV will be overestimated and NPV will be underestimated. For this reason, it is not possible to calculate these two parameters from a sample where the numbers of infected and non-infected patients are arbitrarily defined. Nevertheless, even when the study design allows for their calculation, these estimates should be interpreted with caution since they are conditioned by the prevalence of the infection in the source population. Thus, post-test probability estimators should always be considered conditional on the inclusion criteria chosen in the validation study. This dependence prohibits any generalization of these estimators. For instance, as the incidence of endocarditis in people with valve protheses is higher than in the general population, PPV and NPV of an algorithm targeting this infection estimated with a sample of people with valve protheses will not be generalisable to the general population.

While variability in post-test probabilities is usually well accepted, Se and Sp parameters are generally considered to be intrinsic properties of the test or algorithm being evaluated and therefore treated as constant values (41, 42). However, among the reviewed approaches of sample selection, some imply that the sampling of patients is done according to the classification results of the algorithm. By arbitrarily setting the ratio of the number of patients identified by the algorithm as infected and uninfected, a bias is introduced into the estimation of Se and Sp, whose magnitude depends on the real values of these parameters, the prevalence of the infection, and the chosen ratio of positive and negative patients (see Supplementary 3).

As with post-test probabilities, design of the validation study is not the only factor influencing Se and Sp. The concept of spectrum bias or spectrum effect is used to describe the variability in test performances depending on the characteristics of the validation samples (43). This effect, which is well characterized for biological tests of infectious diseases (44–46) should be particularly suspected when validating targeting algorithms in medico-administrative database. Unlike microbiological tests, which are based on biological markers, algorithms targeting infectious diseases in medico-administrative databases are based on medical parameters (e.g., diagnoses, anti-infectives prescriptions, or screening tests) which are the result of complex processes, interactions and decisions that are likely to vary broadly according to patient characteristics. For instance, a vulvar infection is more likely to be notified in the discharge summary of a pregnant or an immunocompromised woman than in the general population.

Thus, both design of the validation study and characteristics of the validation sample are critical for the quality of the estimates of algorithmic performance indicators. It therefore seems essential that authors of validation studies detail the design of their study and the estimation of valid performance indicators. Equally important, the precise reporting of the inclusion criteria and the description of the validation sample characteristics will allow to assess the scope of these performance indicators and consequently their transposability to future studies.

Eventually, it seems important to discuss the reuse of validated algorithm. In many cases, it appears that good performance parameters of an algorithm, in particular a high PPV, would alone justify its reuse and the robustness of the obtained results. However, even with satisfying performance, the few classification errors that an algorithm may generate are likely to induce significant bias in the results of a study. The magnitude and direction of this bias depend on the algorithm's performance, but also on its variability according to individual characteristics and the role of the algorithm in the study design (definition of an exposure/confounding, an outcome) (47). Thus, the reuse of algorithms for which performance indicators are available should always come along with a quantitative bias analysis to assess the impact of even minor misclassifications on the results of the research. Many quantitative bias analysis methods have been developed and implemented in most analysis software and are now extensively described (48).

4.2. Capture-recapture studies

Capture-recapture methods allow the simultaneous estimation of the infectious disease incidence and the completeness of data sources. However, these methods are limited by conditions of application which should be particularly questioned in the context of the use of medico-administrative database as information source (49).

The first criterion conditioning the validity of this method is that all the cases identified in the different sources are true cases. In the context of medico-administrative databases, this condition refers directly to the validity of the targeting algorithm and is therefore prone to be unsatisfied. Brenner demonstrated that in the case of a two-source model where only the medico-administrative database is affected by misclassification, and in absence of any correction for the status of misclassified patients, the total number of cases would always be overestimated by a factor equal to the inverse of the algorithm PPV (50). Two of the studies identified in this review took this issue into account through a systematic review of the medical records of all cases identified in the medico-administrative database to eliminate all false positive patients (32, 37). This strategy was feasible due to the scarcity inherent to the targeted events (mucormycosis and malaria-related deaths) resulting in a low volume of records to be reviewed (just over 200 in each of the two studies). Without necessarily proceeding to a systematic verification of cases identified, an evaluation of the PPV of the algorithm used in the medico-administrative database could allow correcting the overestimation induced thru misclassification by weighting the estimated total number of cases by the PPV value.

A second condition affecting the validity of the completeness estimators concerns the linkage of information sources. Indeed, capture-recapture implies that all cases common to the different sources are identified and that no case, in any source, is erroneously matched to a different case in another source. The relative impact of matching errors on the estimate of the total number of patients depends on the number of matching errors (erroneously matched cases and erroneously unmatched cases) and the actual number of cases shared by the information sources (see Supplementary 4) (51). Thus, if the overall matching error leads to an underestimation of the number of cases common to the databases, the estimate of the total number of patients will be overestimated and the completeness of the databases used will be underestimated. In this review, only one study was able to use a directly identifying variable (patient name) as a key to link the PMSI with the complementary data

source (37). The other studies had to use a series of indirectly identifying variables to match patients. Three other studies also reported introducing some tolerance on the values of patients' age and dates of care to reduce the bias associated with data entry errors (38–40). This relaxation of the matching rules reduces the risk of missing a valid match but simultaneously increases the risk of false matches. This type of manipulation falls within the complex field of probabilistic matching methods that often prove to be indispensable when dealing with medico-administrative databases (52, 53). However, evaluating their results becomes particularly challenging in the context of approaches such as capture-recapture, since the number of cases to be matched is not known in advance. A sensitivity analysis should at least be carried out to assess the impact of linkage bias on the recapture estimators.

A third assumption underlying capture-recapture models is the independence of the sources of information involved. Dependence between two sources emerges when the presence of a case in one of them affects the probability of the case being present in the other source. A positive dependence results in an underestimation of the number of cases, while a negative dependence has the opposite effect (54). Capture-recapture studies included in this review all used the PMSI as source of information and linked it to additional sources coming from death certificates, national reference centers, mandatory reporting registers or surveillance networks. Given that the occurrence of an infectious disease case in the PMSI largely relies on diagnosis coded by hospital practitioners who are often responsible for notifying the other data sources, a positive dependency between the PMSI and the other data sources should be strongly suspected. The assessment of dependency between n sources always requires the involvement of at least an additional source (n+1). For this reason, it is strongly recommended to carry out capture-recapture studies based on at least 3 data sources (55). This literature review identified only one study that used three databases (32). This study on malaria mortality used two methods to assess and consider the dependence between sources in the estimation of the total number of deaths. The first method consists in making a contingency table of the presence of cases in the first two data sources using only the cases identified in the third source (55). Thus, this method provides the number of cases not recorded in the first two sources and enables evaluating their independence by a χ^2 test or an odds ratio calculation. The second method consists to estimate the number of cases in the population using a log-linear model (49). Log-linear models have the advantage of allowing the integration of interaction terms to account for dependence between data sources. However, due to the impossibility to estimate the interaction of highest degree (i.e., the interaction involving all available sources simultaneously), log-linear methods cannot be used to evaluate dependency in a two sources capture-recapture study. Nevertheless, these methods can be used for sensitivity analysis in two sources capture-recapture study (56). Indeed, the integration of a parameter with a fixed value into the model allows to assess the impact of different degrees of dependency between sources on the estimated value of the total number of cases in the population.

A final constraint of capture-recapture methods concerns the homogeneity of captures within each source. In other words, the probability of inclusion within each data source must be identical for all cases, regardless of their individual characteristics (57). A first method to account for this potential bias is to stratify the recapture analysis on available variables. Two studies included in this review used this method as a sensitivity analysis by stratifying their analyses on patients' characteristics (32, 38). Another option is to consider covariates that are potentially sources of heterogeneity in the estimation of the total number of infectious disease cases in the population of interest by incorporating them into a log-linear model. It has been shown that this method also has the advantage of allowing the use of partially observed variables (i.e., variables not measured in all data sources) through the use of an expectation-maximization algorithm (58).

5. Conclusion

Despite many limitations, medico-administrative databases are sources of information that make possible the study of many health phenomena on considerable numbers of individuals. As their data supply is automated through health insurance systems, they constitute easily accessible and inexpensive sources of information that are highly complementary to more traditional sources of epidemiological data. For these reasons, their use for epidemiological purposes is expected to continue to grow. In the field of infectious diseases, they could therefore become a valuable resource for the monitoring of antimicrobial drug use, the surveillance of resistant or nosocomial infections, or the description of epidemic dynamics.

This literature review showed that despite its importance, quantitative evaluation of algorithms targeting infectious diseases in French medico-administrative databases is not yet a common practice for epidemiologists. This is undoubtedly linked to the fact that these evaluation studies, even though their seeming methodological simplicity, are in reality challenging to implement: constitution of a representative validation sample, definition and application of a reference classification, and above all regulatory and technical constraints linked to the matching of medico-administrative databases with other sources of information.

This literature review was also an opportunity to highlight the many risks of bias related to the underlying assumptions of these evaluation methods. These additional constraints should never be set aside by authors of evaluation studies but should be considered when planning their research work so that it can be compatible with analysis methods that allow the assessment of these assumptions or to take into account their violation. At the very least, these studies should include sensitivity analyses to assess the impact of breaching these assumptions on their estimates of performance parameters of the evaluated algorithms.

Despite all their limitations, medico-administrative databases are valuable sources of information for epidemiological research, especially if they are linked to other sources of information to enrich them with clinical and biological content. Linkages between medico-administrative databases and more conventional epidemiological databases should be encouraged and facilitated as they would allow both the implementation of more powerful observational studies and at the same time the evaluation and development of useful targeting algorithms when these medico-administrative databases are used alone.

Author contributions

M-FT, KS, and LG-G contributed to the review conception and design. M-FT performed the literature search, study selection, data extraction, and wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript, authors read, and approved the final manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2023.1161550/full#supplementary-material

References

1. Brand D, Moreau A, Cazein F, Lot F, Pillonel J, Brunet S, et al. Characteristics of patients recently infected with HIV-1 non-B subtypes in France: a nested study within the mandatory notification system for new HIV diagnoses. *J Clin Microbiol.* (2014) 52:4010–6. doi: 10.1128/JCM.01141-14

2. Barry A, Ahuka-Mundeke S, Ali Ahmed Y, Allarangar Y, Anoko J, Archer BN, et al. Outbreak of Ebola virus disease in the Democratic Republic of the Congo, April–May, 2018: an epidemiological study. *Lancet.* (2018) 392:213–21.

3. Lau JTF, Tsui H, Lau M, Yang X. SARS transmission, risk factors, and prevention in Hong Kong. *Emerg Infect Dis.* (2004) 10:587–92. doi: 10.3201/eid1004.030628

4. Guillon A, Laurent E, Duclos A, Godillon L, Dequin P-F, Agrinier N, et al. Case fatality inequalities of critically ill COVID-19 patients according to patient-, hospital- and region-related factors: a French nationwide study. *Ann Intensive Care.* (2021) 11:127. doi: 10.1186/s13613-021-00915-4

5. Piroth L, Cottenet J, Mariet A-S, Bonniaud P, Blot M, Tubert-Bitter P, et al. Comparison of the characteristics, morbidity, and mortality of COVID-19 and seasonal influenza: a nationwide, population-based retrospective cohort study. *Lancet Respir Med.* (2021) 9:251–9. doi: 10.1016/S2213-2600(20)30527-0

6. Li J, Huang DQ, Zou B, Yang H, Hui WZ, Rui F, et al. Epidemiology of COVID-19: a systematic review and meta-analysis of clinical characteristics, risk factors, and outcomes. *J Med Virol.* (2021) 93:1449–58. doi: 10.1002/jmv.26424

7. Gavrielov-Yusim N, Friger M. Use of administrative medical databases in population-based research. *J Epidemiol Community Health.* (2014) 68:283–7. doi: 10.1136/jech-2013-202744

8. Cadarette SM, Wong L. An introduction to health care administrative data. *Can J Hosp Pharm.* (2015) 68:232–7. doi: 10.4212/cjhp.v68i3.1457

9. Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc J-L, Sailler L. French health insurance databases: what interest for medical research. *Rev Med Interne.* (2015) 36:411–7. doi: 10.1016/j.revmed.2014.11.009

10. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. *Rev DÉpidémiologie Santé Publique.* (2017) 65:S149–67. doi: 10.1016/j.respe.2017.05.004

11. LOI. *2016-41 du 26 janvier 2016 de modernisation de notre système de santé.* (2016). Available online at: https://www.legifrance.gouv.fr/loda/id/JORFTEXT000031912641 (accessed June 06, 2021).

12. Virnig BA, McBean M. Administrative data for public health surveillance and planning. *Annu Rev Public Health.* (2001) 22:213–30. doi: 10.1146/annurev.publhealth.22.1.213

13. Grimes DA. Epidemiologic research using administrative databases: garbage in, garbage out. *Obstet Gynecol.* (2010) 116:1018–9. doi: 10.1097/AOG.0b013e3181f98300

14. Mazzali C, Duca P. Use of administrative data in healthcare research. *Intern Emerg Med.* (2015) 10:517–24. doi: 10.1007/s11739-015-1213-9

15. Goldberg M, Carton M, Doussin A, Fagot-Campagna A, Heyndrickx E, Lemaitre M, et al. Le réseau REDSIAM (Réseau données Sniiram) – Spécial REDSIAM. *Rev DÉpidémiologie Santé Publique.* (2017) 65:S144–8. doi: 10.1016/j.respe.2017.06.001

16. Fonteneau L, Le Meur N, Cohen-Akenine A, Pessel C, Brouard C, Delon F, et al. Apport des bases médico-administratives en épidémiologie et santé publique des maladies infectieuses. *Rev DÉpidémiologie Santé Publique.* (2017) 65:S174–82. doi: 10.1016/j.respe.2017.03.131

17. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med.* (2018) 169:467–73. doi: 10.7326/M18-0850

18. Martin-Latry K, Cougnard A. Terminologie utilisée concernant les bases de remboursement de l'assurance maladie en pharmaco-épidémiologie : une harmonisation nécessaire. *Therapies.* (2010) 65:379–85. doi: 10.2515/therapie/2010047

19. Études / publications | L'Assurance Maladie. (2021). Available online at: https://assurance-maladie.ameli.fr/etudes-et-donnees/etudes-publications (accessed February 21, 2022).

20. Portail documentaire Santé publique France - Cadic intégrale. (2021). Available online at: https://portaildocumentaire.santepubliquefrance.fr/exl-php/recherche/spf___internet_page_accueil (Accessed February 21, 2022).

21. EPI-PHARE : rapports d'études et publications. (2021). Available online at: https://www.epi-phare.fr/rapports-detudes-et-publications/ (accessed February 21, 2022).

22. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. *J Clin Epidemiol.* (2012) 65:126–31. doi: 10.1016/j.jclinepi.2011.08.002

23. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol.* (2011) 64:821–9. doi: 10.1016/j.jclinepi.2010.10.006

24. Widdifield J, Labrecque J, Lix L, Paterson JM, Bernatsky S, Tu K, et al. Systematic review and critical appraisal of validation studies to identify rheumatic diseases in health administrative databases. *Arthritis Care Res.* (2013) 65:1490–503. doi: 10.1002/acr.21993

25. Dely C, Sellier P, Dozol A, Segouin C, Moret L, Lombrail P. Preventable readmissions of "community-acquired pneumonia": Usefulness and reliability of an indicator of the quality of care of patients' care pathways. *Presse Medicale Paris Fr.* (2012) 41:e1-9. doi: 10.1016/j.lpm.2011.06.007

26. Bounoure F, Beaudeau P, Mouly D, Skiba M, Lahiani-Skiba M. Syndromic surveillance of acute gastroenteritis based on drug consumption. *Epidemiol Infect.* (2011) 139:1388–95. doi: 10.1017/S095026881000261X

27. Bounoure F, Mouly D, Beaudeau P, Bentayeb M, Chesneau J, Jones G. Syndromic surveillance of acute gastroenteritis using the french health insurance database: discriminatory algorithm and drug prescription practices evaluations. *Int J Environ Res Public Health.* (2020) 17: doi: 10.3390/ijerph17124301

28. Leclère B, Lasserre C, Bourigault C, Juvin M-E, Chaillet M-P, Mauduit N, et al. Matching bacteriological and medico-administrative databases is efficient for a computer-enhanced surveillance of surgical site infections: retrospective analysis of 4,400 surgical procedures in a French university hospital. *Infect Control Hosp Epidemiol.* (2014) 35:1330–5. doi: 10.1086/678422

29. Jones G, Tarright N, Boelle PY, Marty J, Lalande V, Eckert C, Barbut F. Accuracy of ICD-10 Codes for Surveillance of Clostridium difficile Infections, France.- *Emerging Inf. Dis. J.* CDC. (2012) 18, 188. doi: 10.3201/eid1806.111188

30. Soilly A-L, Ferdynus C, Desplanches O, Grimaldi M, Gouyon JB. Paediatric intensive care admissions for respiratory syncytial virus bronchiolitis in France: results of a retrospective survey and evaluation of the validity of a medical information system programme. *Epidemiol Infect.* (2012) 140:608–16. doi: 10.1017/S0950268811001208

31. Fox MP, Lash TL, Bodnar LM. Common misconceptions about validation studies. *Int J Epidemiol.* (2020) 49:1392–6. doi: 10.1093/ije/dyaa090

32. Kendjo E, Thellier M, Noël H, Jauréguiberry S, Septfons A, Mouri O, et al. Mortality from malaria in France, 2005 to 2014. *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull.* (2020) 25:579. doi: 10.2807/1560-7917.ES.2020.25.36.1900579

33. Bitar D, Van Cauteren D, Lanternier F, Dannaoui E, Che D, Dromer F, et al. Increasing incidence of Zygomycosis (Mucormycosis), France, 1997-2006. *Emerg Infect Dis.* (2009) 15:1395–401. doi: 10.3201/eid1509.090334

34. Dubos F, Maréchal I, Tilmont B, Courouble C, Leclerc F, Martinot A. Incidence des infections invasives à méningocoque de l'enfant dans le Nord–Pas-de-Calais : intérêt et limites du programme de médicalisation des systèmes d'information (PMSI) pour la correction des données des déclarations obligatoires. *Arch Pédiatrie.* (2009) 16:984–90. doi: 10.1016/j.arcped.2009.03.006

35. Molinie F, Le Tourneau B, Ilef D. Estimation de l'exhaustivité du système de surveillance des infections à méningocoque dans le Nord-Pas-de-Calais, 1997-1998. *Bull Epidemiol Hebd.* (2002) 41:203–5.

36. Belchior E, Zeller H, Nicolau J, Vaillant V, Capek I. La fièvre hémorragique avec syndrome rénal en France métropolitaine de 2002 à 2007 : données du PMSI et du CNR. Bull Epidemiol Hebd (2009) 22:233–6.

37. Bitar D, Morizot G, Van Cauteren D, Dannaoui E, Lanternier F, Lortholary O, et al. Estimating the burden of mucormycosis infections in France (2005-2007) through a capture-recapture method on laboratory and administrative data. *Rev Epidemiol Sante Publique.* (2012) 60:383–7. doi: 10.1016/j.respe.2012.03.007

38. Loury P, Jones G, Chappert J, Pivette M, Hubert B. Analyse de l'exhaustivité et de la qualité de la surveillance des grippes sévères, 2009-2013. *Saint-Maurice: Santé publique France.* (2017) 47:S123–4. doi: 10.1016/j.medmal.2017.03.298

39. Pivette M, Loury P. Analyse de l'exhaustivité de la surveillance des grippes sévères en France métropolitaine, saison 2017-2018. *Bull Epidemiol Hebd.* (2019) 28:571–2.

40. Verrier F, Etienne A, Vincent M, Vilain P, Lafont M, Mourembles G, et al. Sévérité de l'épidémie de dengue à La Réunion : données de surveillance des cas hospitalisés, avril 2017 à décembre 2018. *Bull Epidemiol Hebd.* (2019) 19:383–9.

41. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. *Contin Educ Anaesth Crit Care Pain.* (2008) 8:221–3. doi: 10.1093/bjaceaccp/mkn041

42. Gallagher EJ. The problem with sensitivity and specificity. *Ann Emerg Med.* (2003) 42:298–303. doi: 10.1067/mem.2003.273

43. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation *Ann Intern Med.* (2002) 137:598–602. doi: 10.7326/0003-4819-137-7-200210010-00011

44. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med.* (1992) 117:135–40. doi: 10.7326/0003-4819-117-2-135

45. DiMatteo LA, Lowenstein SR, Brimhall B, Reiquam W, Gonzales R. The relationship between the clinical features of pharyngitis and the sensitivity of a rapid antigen test: Evidence of spectrum bias. *Ann Emerg Med.* (2001) 38:648–52. doi: 10.1067/mem.2001.119850

46. Einhauser S, Peterhoff D, Niller HH, Beileke S, Günther F, Steininger P, et al. Spectrum bias and individual strengths of SARS-CoV-2 serological tests—a population-based evaluation. *Diagnostics.* (2021) 11:1843. doi: 10.3390/diagnostics11101843

47. Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. *Curr Epidemiol Rep.* (2014) 1:175–85. doi: 10.1007/s40471-014-0027-z

48. Petersen JM, Ranker LR, Barnard-Mayers R, MacLehose RF, Fox MP, A. systematic review of quantitative bias analysis applied to epidemiological research. *Int J Epidemiol.* (2021) 50:1708–30. doi: 10.1093/ije/dyab061

49. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev.* (1995) 17:243–64. doi: 10.1093/oxfordjournals.epirev.a036192

50. Brenner H. Effects of misdiagnoses on disease monitoring with capture—Recapture methods. *J Clin Epidemiol.* (1996) 49:1303–7. doi: 10.1016/0895-4356(95)00026-7

51. Gerritse SC, Bakker BF, van der Heijden PG. *The Impact of Linkage Errors and Erroneous Captures on the Population Size Estimator Due to Implied Coverage*. Statistics Netherlands (2017). Available online at: https://www.cbs.nl/en-gb/background/2017/39/impact-of-linkage-errors-and-erroneous-captures (accessed May 05, 2022).

52. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. *Int J Epidemiol.* (2016) 45:954–64. doi: 10.1093/ije/dyv322

53. Harron K, Dibben C, Boyd J, Hjern A, Azimaee M, Barreto ML, et al. Challenges in administrative data linkage for research. *Big Data Soc.* (2017) 4:2053951717745678. doi: 10.1177/2053951717745678

54. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent. *Epidemiology.* (1995) 6:42–8. doi: 10.1097/00001648-199501000-00009

55. Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. *J Chronic Dis.* (1974) 27:25–36. doi: 10.1016/0021-9681(74)90005-8

56. Gerritse SC, Heijden PGM. van der, Bakker BFM. Sensitivity of population size estimation for violating parametric assumptions in log-linear models. *J Off Stat.* (2015) 31:357–79. doi: 10.1515/jos-2015-0022

57. Hook EB, Regal RR. Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. *Am J Epidemiol.* (1993) 137:1148–66. doi: 10.1093/oxfordjournals.aje.a116618

58. Zwane EN, van der Heijden PGM. Analysing capture–recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations. *Stat Med.* (2007) 26:1069–89. doi: 10.1002/sim.2577

59. de Lafforest S, Magnier A, Vallée M, Bey E, Le Goux C, Saint F, et al. FUrTIHF: French urinary tract infections in healthcare facilities – five-year historic cohort (2014–2018). *J Hosp Infect.* (2021) 116:29–36. doi: 10.1016/j.jhin.2021.04.035

60. Gerbier S, Bouzbid S, Pradat E, Baulieux J, Lepape A, Berland M, et al. [Use of the French medico-administrative database (PMSI) to detect nosocomial infections in the University hospital of Lyon]. *Rev Epidemiol Sante Publique.* (2011) 59:3–14. doi: 10.1016/j.respe.2010.08.003

61. Grammatico-Guillon L, Baron S, Rusch E, Lepage B, Surer N, Desenclos JC, et al. Epidemiology of vertebral osteomyelitis (VO) in France: analysis of hospital-discharge data 2002-2003. *Epidemiol Infect.* (2008) 136:653–60. doi: 10.1017/S0950268807008850

62. Grammatico-Guillon L, Thiercelin N, Mariani S, Lecuyer A-I, Goudeau A, Bernard L, et al. [Study of hospitalizations for pneumococcal pneumoniae in Centre region, 2004-2008]. *Rev Epidemiol Sante Publique.* (2012) 60:1–8. doi: 10.1016/j.respe.2011.07.005

63. Grammatico-Guillon L, Baron S, Gettner S, Lecuyer A-I, Gaborit C, Rosset P, et al. Bone and joint infections in hospitalized patients in France, 2008: clinical and economic outcomes. *J Hosp Infect.* (2012) 82:40–8. doi: 10.1016/j.jhin.2012.04.025

64. Grammatico-Guillon L, Maakaroun-Vermesse Z, Baron S, Gettner S, Rusch E, Bernard L. Paediatric bone and joint infections are more common in boys and toddlers: a national epidemiology study. *Acta Paediatr Oslo Nor.* (2013) 102:e120–5. doi: 10.1111/apa.12115

65. Grammatico-Guillon L, Baron S, Gaborit C, Rusch E, Astagneau P. Quality assessment of hospital discharge database for routine surveillance of hip and knee arthroplasty-related infections. *Infect Control Hosp Epidemiol.* (2014) 35:646–51. doi: 10.1086/676423

66. Jouan Y, Grammatico-Guillon L, Espitalier F, Cazals X, François P, Guillon A. Long-term outcome of severe herpes simplex encephalitis: a population-based observational study. *Crit Care.* (2015) 19:345. doi: 10.1186/s13054-015-1046-y

67. Sahli L, Lapeyre-Mestre M, Derumeaux H, Moulis G. Positive predictive values of selected hospital discharge diagnoses to identify infections responsible for hospitalization in the French national hospital database. *Pharmacoepidemiol Drug Saf.* (2016) 25:785–9. doi: 10.1002/pds.4006

68. Sunder S, Grammatico-Guillon L, Baron S, Gaborit C, Bernard-Brunet A, Garot D, et al. Clinical and economic outcomes of infective endocarditis. *Infect Dis Lond Engl.* (2015) 47:80–7. doi: 10.3109/00365548.2014.968608

69. Sunder S, Grammatico-Guillon L, Lemaignen A, Lacasse M, Gaborit C, Boutoille D, et al. Incidence, characteristics, and mortality of infective endocarditis in France in 2011. *PLoS ONE.* (2019) 14:e0223857. doi: 10.1371/journal.pone.0223857

70. Tubiana S, Blotière P-O, Hoen B, Lesclous P, Millot S, Rudant J, et al. Dental procedures, antibiotic prophylaxis, and endocarditis among people with prosthetic heart valves: nationwide population based cohort and a case crossover study. *BMJ.* (2017) 358:j3776. doi: 10.1136/bmj.j3776