Intra-database Validation of Case-identifying Algorithms Using Reconstituted Electronic Health Records From Healthcare Claims Data

Nicolas Henri Thurin (nicolas.thurin@u-bordeaux.fr)
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi, Bordeaux, France https://orcid.org/0000-0003-3589-0819

Pauline Bosco-Levy
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Patrick Blin
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Magali Rouyer
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Jérémy Jové
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Stéphanie Lamarque
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Lignot Séverine
Univ.Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Régis Lassalle
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Abdelilah Abouelfath
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Bignon Emmanuelle
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Diez Pauline
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Marine Gross-Goupil
Départment of medical Oncology, Hôpital saint-André, CHU de Bordeaux

Michel Soulié
Department of Urology, University Hospital of Rangueil, CHU de Toulouse

Mathieu Roumiguié
Department of Urology, University Hospital of Rangueil, CHU de Toulouse

Sylvestre Le Moulec
Department of oncology, Clinique Marzet

Marc Debouverie
Departement of neurology, CHRU de Nacy, Université de Lorraine, EA 4360 APEMAC
Bruno Brochet
CRC SEP, Neurology department, CHU de Bordeaux ; Univ. Bordeaux, INSERM U1215, Neurocentre magendie
Francis Guillemin
Université de Lorraine, EA 4360 APEMAC ; CHRU de Nancy, INSERM CIC 1433 Epidémiologie Clinique
Céline Louapre
Sorbonne Université, Institut du cerveau, ICM, Hôpital de la Pitié Salpêtrière, INSERM S 1127, CNRS UMR 7225 ; Neurology Department, Pitié Salpetriere Hospital, APHP
Elisabeth Maillart
Neurology Department, Pitie Salpetriere Hospital, APHP
Olivier Heinzlef
department of Neurology, Hôpital CHI de Poissy/Saint-germain-en-Laye
Nicholas Moore
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi
Droz-Perroteau Cécile
Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi

Research article

Keywords: Validation study, case-identifying algorithm, claims database, reconstituted Electronic Health Record, Multiple Sclerosis, Prostate Cancer, Positive Predictive Value, Negative Predictive Value

Posted Date: January 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-145262/v1

Version of Record: A version of this preprint was published on May 1st, 2021. See the published version at https://doi.org/10.1186/s12874-021-01285-y.

Intra-database validation of case-identifying algorithms using reconstituted electronic health records from healthcare claims data

Authors: Nicolas H. Thurin[1†], Pauline Bosco-Levy[1†], Patrick Blin[1], Magali Rouyer[1], Jérémy Jové[1], Stéphanie Lamarque[1], Séverine Lignot[1], Régis Lassalle[1], Abdelilah Abouelfath[1], Emmanuelle Bignon[1], Pauline Diez[1], Marine Gross-Goupil[2], Michel Soulié[3], Mathieu Roumiguié[3], Sylvestre Le Moulec[4], Marc Debouverie[5,6], Bruno Brochet[7,8], Francis Guillemin[6,9], Céline Louapre[10,11], Elisabeth Maillart[11], Olivier Heinzlef[12], Nicholas Moore[1], Cécile Droz-Perroteau[1]

[†] Contributed equally

Affiliations:

[1] Univ. Bordeaux, INSERM CIC-P1401, Bordeaux PharmacoEpi, Bordeaux, France

[2] Department of Medical Oncology, Hôpital Saint André, CHU de Bordeaux, Bordeaux, France

[3] Department of Urology, University Hospital of Rangueil, CHU de Toulouse, Toulouse, France

[4] Department of oncology, Clinique Marzet, Pau, France

[5] Department of neurology, CHRU de Nancy, Nancy, France

[6] Université de Lorraine, EA 4360 APEMAC, Nancy, France

[7] CRC SEP, Neurology department, CHU de Bordeaux, Bordeaux, France

[8] Univ. Bordeaux, INSERM U1215, Neurocentre Magendie, Bordeaux, France

[9] CHRU de Nancy, INSERM CIC 1433 Epidémiologie clinique, Nancy, France

[10] Sorbonne Université, Institut du cerveau, ICM, Hôpital de la Pitié Salpêtrière, INSERM UMR S 1127, CNRS UMR 7225, Paris, France

[11] Neurology Department, Pitie Salpetriere Hospital, APHP, Paris, France,

[12] Department of Neurology, Hôpital CHI de Poissy/Saint-Germain-en-Laye, Paris, France

Corresponding author : Nicolas H. Thurin, nicolas.thurin@u-bordeaux.fr

Email addresses: Nicolas H. Thurin nicolas.thurin@u-bordeaux.fr; Pauline Bosco-Levy pauline.bosco-levy@u-bordeaux.fr; Patrick Blin patrick.blin@u-bordeaux.fr; Magali Rouyer magali.rouyer@u-bordeaux.fr; Jérémy Jové jeremy.jove@u-bordeaux.fr; Stéphanie Lamarque stephanie.lamarque@u-bordeaux.fr; Séverine Lignot severine.lignot@u-bordeaux.fr; Régis Lassalle regis.lassalle@u-bordeaux.fr; Abdelilah Abouelfath abdelilah.abouelfath@u-bordeaux.fr; Emmanuelle Bignon emmanuelle.bignon@u-bordeaux.fr; Pauline Diez pauline.diez@u-bordeaux.fr; Marine Gross-Goupil marine.gross-goupil@chu-bordeaux.fr; Michel Soulié soulie.m@chu-toulouse.fr; Mathieu Roumiguié roumiguie.m@chu-toulouse.fr; Sylvestre Le Moulec sylvestre.lemoulec@gmail.com; Marc Debouverie m.debouverie@chru-nancy.fr; Bruno Brochet bruno.brochet@chu-bordeaux.fr; Francis Guillemin francis.guillemin@chru-nancy.fr; Céline Louapre celine.louapre@aphp.fr; Elisabeth Maillart elisabeth.maillart@aphp.fr; Olivier Heinzlef Olivier.Heinzlef@ght-yvelinesnord.fr; Nicholas Moore nicholas.moore@u-bordeaux.fr; Cécile Droz-Perroteau cecile.droz@u-bordeaux.fr

ORCID

Nicolas H. Thurin 0000-0003-3589-0819; Pauline Bosco-Levy 0000-0001-9763-5974; Patrick Blin 0000-0003-4005-7928; Magali Rouyer 0000-0002-2560-4412; Régis Lassalle 0000-0001-6726-6215; Bruno Brochet 0000-0003-3824-2796; Céline Louapre 0000-0002-4987-1531; Elisabeth Maillart 0000-0001-7699-0328; Nicholas Moore 0000-0003-1212-2817; Cécile Droz-Perroteau 0000-0002-7697-1167

ABSTRACT

Background: Diagnosis performances of case-identifying algorithms developed in healthcare database are usually assessed by comparing identified cases with an external data source. When this is not feasible, intra-database validation can present an appropriate alternative.

Objectives: To illustrate through two practical examples how to perform intra-database validations of case-identifying algorithms using reconstituted Electronic Health Records (rEHRs).

Methods: Patients with 1) multiple sclerosis (MS) relapses and 2) metastatic castration-resistant prostate cancer (mCRPC) were identified in the French nationwide healthcare database (SNDS) using two case-identifying algorithms. A validation study was then conduct to estimate diagnostic performances of these algorithms through the calculation of their positive predictive value (PPV) and negative predictive value (NPV). To that end, anonymized rEHRs were generated based on the overall information captured in the SNDS over time (e.g. procedure, hospital stays, drug dispensing, medical visits) for a random selection of patients identified as cases or non-cases according to the predefined algorithms. For each diseases, an independent validation committee reviewed the rEHRs of 100 cases and 100 non-cases in order to adjudicate on the status of the selected patients (true case/ true non-case), blinded with respect to the result of the corresponding algorithm.

Results: Algorithm for relapses identification in MS showed a 95% PPV and 100% NPV and for mCRPC identification, a 97% PPV and 99% NPV.

Conclusion: The use of rEHRs to conduct an intra-database validation appears to be a valuable tool to estimate the performances of a case-identifying algorithm and assess its validity, in the absence of alternative.

KEYWORDS

Validation study; case-identifying algorithm; claims database; reconstituted Electronic Health Record; Multiple Sclerosis; Prostate Cancer; Positive Predictive Value; Negative Predictive Value

DECLARATIONS

Ethics approval and consent to participate: In both of the presented examples, rEHR adjudication was used to answer the main research question of the corresponding study for which the approval from the French data protection agency (*Commission Nationale Informatique & Libertés* – CNIL) was obtained. Moreover, data were fully anonymized before adjudication.

Consent for publication: Not applicable

Availability of data and material: As per law raw SNDS data cannot be shared. Access to SNDS data requires approval from the *Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé* (CESREES) in charge of assessing scientific quality of the project, and authorization from the *Commission Nationale de*

l'Informatique et des Libertés (CNIL) which is the French data protection authority, and then an agreement with the SNDS data holder (CNAM).

Competing interests:

M. G.-G. declares personal fees and non-financial support from Janssen, Sanofi, Astellas, Ipsen, Amgen and Pfizer.

M. S. and M. R. declare personal fees and non-financial support from Janssen, Sanofi, Astellas, Ipsen, Amgen, Ferring, and Astra-Zeneca.

E. M. declares personal fees and non-financial support from Biogen, Novartis, Roche, Merck, Sanofi-Genzyme.

B. B. declares personal fees and non-financial support from Biogen, Genzyme, Bayer, Medday, Actelion, Roche, Celgene, Novartis, Merck.

F. G. and M.D. declare personal fees and non-financial support from Biogen.

C. L. declares consulting or travel fees from Biogen, Novartis, Roche, Sanofi, Teva and Merck Serono, and research grant from Biogen

O. H. declares personal fees and non-financial support from Biogen, Merck, Novartis, Roche, Genzyme

All remaining authors have declared no conflicts of interest.

Funding: Presented examples were drawn from two studies funded by Janssen-Cilag, France and Biogen, France.

Authors' contributions: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by J.J., R.L. and A.A.. All authors discussed the results. The first draft of the manuscript was written by P. B.-L. and N. H. T.

and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Both of the presented examples were drawn from studies carried out by the Bordeaux PharmacoEpi platform in collaboration with Janssen-Cilag, France and Biogen, France and supervised by independent Scientific Committees. The authors thank all the members of these scientific committees for their support and advices. The authors thank ADERA for legal, human resource and management support that made these studies possible.

INTRODUCTION

For the last two decades, the use of healthcare databases has considerably increased in health research field [1]. This trend is fueled by the growing recognition that randomized clinical trials, while essential, are not the unique and exhaustive answer to therapeutic efficacy and safety issues. The wealth of information that healthcare databases contain, made them robust tool for many epidemiology-related fields of research, especially in pharmacoepidemiology, where epidemiologic approaches are applied to well-defined and/or large population to assess the use and the effects of drugs in real-world practice [2, 3]. The very large amount of data collected prospectively and systematically in extended period of times, mainly for billing purposes, enables the assessment of infrequent or delayed adverse events as well as therapeutic long-term effectiveness, which is complex to assess in classical randomized trials, field cohort or registry [4]. However, the use of secondary data collected for other purposes than epidemiologic research, is not devoid of significant limitations [5, 6]. Data quality is a major issue that may impact case identification by inducing a selection or misclassification bias. In studies conducted

on healthcare databases, the population or the health outcome of interest are generally identified using in- and/or out-patient diagnosis codes. To enhance accuracy, algorithms including multiple elements specifically related to the studied medical condition (*e.g.* medical procedures, drug dispensing, laboratory test or radiological exam) in addition to the diagnosis code, may also be developed and implemented [7]. Whatever the approach used, the coding quality may be nuanced in terms of how the codes are applied or how physician records are interpreted by the medical reviewer entering the codes. The financial pressure induced by activity based payment may also lead to encourage the income-maximizing coding of diagnoses and procedures in hospitals at the expense of clinical accuracy [8], although more and more quality audits are carried out to improve coding reliability [9–11]. The validity of algorithm used to identify health outcome in administrative and claims data has always been a matter of concern for researchers, especially in a context of active surveillance and assessment of marketed medical product [12, 13]. Several different types of validation studies may be conducted to assess the fidelity of the codes or algorithms used for cases identification. In all of them, cases identified by the algorithm are compared with a presumably more reliable external diagnostic source or gold standard [14][7]. These gold standards are most of the time the information that have originated the records in the database (e.g. medical charts or registries) and which contain measure of the disease status based on clinical, biological and/or imaging criteria. The performance of a case-identifying code or algorithm is commonly reported in term of positive predictive value, sensitivity and specificity. Although necessary, these validation studies are time-consuming and require significant resources and expertise to review diagnoses of clinical data sources. Setting up such a process is also not always possible since the access to the original data source is often complicated or even impossible because of technical or legal issues.

Healthcare databases, are constantly updated with all patient healthcare encounters – medical visits and procedures, lab tests or medical imaging, drugs dispensing, hospital stays, *etc.* – over

a considered period of time, or sometimes a lifetime. They may, by their richness and their depth, contain information not available in medical charts. Hence, they may provide a holistic overview of the patient's journeys in real-life settings. These longitudinal patient records could be seen as reconstituted Electronic Health Records (rEHRs) and so constitutes a valuable alternative to medical chart in validation study of case-identifying algorithms.

The objective of this paper is to illustrate through two examples of validation studies conducted in the French nationwide healthcare database, the *Système National des Données de Santé* (SNDS) [15], how to perform intra-database validations of case-identifying algorithm using anonymized rEHRs.

METHODS

Data source

Two validation studies were conducted using data from the SNDS, which currently covers more than 99% of the French population from birth (or immigration) to death (or emigration), even if a subject moves, changes occupation or retires [15, 16]. Using a unique pseudonymized identifier, the SNDS merges all reimbursed outpatient claims from all French healthcare insurance schemes with hospital-discharge summaries from public and private hospitals, and the national death registry. As a consequence, the SNDS contains information on all reimbursed medical and paramedical encounters. For each expenditure, the prescriber and caregiver specialties as well as the corresponding date are provided. The exact quantity of drug dispensed and reimbursed can be identified at the product level with the exact form and dosage. Performed laboratory test and procedures are available but without results. Registration for Long Term Disease (LTD) – status that ensures a full coverage for all related medical expenses – hospital

discharge diagnosis and cause of death are defined using codes from the International Classification of Diseases, 10th revision (ICD-10).

General method

In the frame of different projects approved by the French regulatory authorities (*Comité d'Expertise pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé,* CEREES and *Commission Nationale Informatique & Libertés,* CNIL), two algorithms were developed in the SNDS in collaboration with clinical experts of the field to identify: 1) multiple sclerosis (MS) relapses and 2) metastatic castration-resistant prostate cancer (mCRPC). The same methodology for intra-database validation was then applied to each of them in order to ascertain that patients identified as cases or non-cases by the algorithm were respectively true cases and true non-cases.

In a first step, anonymized longitudinal rEHRs were generated based on SNDS data for a random selection of 100 patients identified by the algorithm as case and 100 patients identified by the algorithm as non-case (Figure 1). To ensure that individual data contained in these rEHRs did not lead to patient re-identification, new patient identifiers were assigned, calendar dates were replaced by the delay elapsed since inclusion, location details were deleted and only age classes were displayed. In a second step, a validation committee consisting of medical experts of the field, proceeded to a double review of the rEHRs in order to adjudicate on the true case or true non-case status of the selected patients, blinded with respect to the algorithm result. In case of discrepancy, committee members discussed to reach a consensus. In a final step, experts conclusions were compared with algorithm results to estimate its diagnostic performance through the positive predictive value (PPV), the negative predictive value (NPV) and their corresponding 95% confidence intervals (95%CI). The formulae for PPV and NPV were:

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

where TP and FP are respectively true and false positives, and TN and FN true and false negatives. Corresponding formula for 95%CIs were:

$$[95\%CI]_{PPV} = PPV \pm z_{1-\frac{\alpha}{2}} * \sqrt{\frac{PPV(1-PPV)}{n_{positive}}}$$

$$[95\%CI]_{NPV} = NPV \pm z_{1-\frac{\alpha}{2}} * \sqrt{\frac{NPV(1-NPV)}{n_{negative}}}$$

where $n_{positive}$ and $n_{negative}$ are respectively the number of algorithm-based positive and negative assessed cases, and $z_{(1-\alpha/2)}$ the z-value for standard normal distribution with left-tail probability (1-α/2). Here $z_{(1-\alpha/2)} = 1.96$ for a type-1 error $\alpha = 0.05$.

The limitation of the number of rEHRs to be assessed per group to 100 allowed to estimate PPV and NPV with a margin of error <10% for values above 50%, and made the adjudication of cases possible by experts in less than 48h.

Following the validation study, experts' inputs were used to adjust algorithm settings and further improve its discriminatory ability. Overall estimated performances indicators were then updated.

Case examples

1) Relapses identification in Multiple Sclerosis (MS) patients

The algorithm for identifying relapse in MS was initially developed in the EVIDEMS study whose objective was to assess the effectiveness of dimethyl fumarate versus other MS drugs (*i.e.* teriflunomide, fingolimod or immunomodulatory injectable drugs) on relapses after treatment initiation [17]. The study cohort included all patients identified in the SNDS by a first dispensing of MS drug (i.e. dimethyl fumarate, indicated for MS) between July 2015 and December 2017, with 4.5-year history and 1 to 3.5 years of follow-up. Relapses were identified

using a complex algorithm including dispensing of high dose of corticosteroids (methylprednisolone or betamethasone) for outpatient, and hospitalizations with MS relapse diagnosis whether or not combined with high dose of corticosteroids.

2) *Patients with metastatic Castration-Resistant Prostate Cancer (mCRPC)*

The algorithm for mCRPC patients identification was initially developed in the CAMERRA study whose objectives were to assess mCRPC burden and describe mCRPC-specific treatment lines [18, 19]. The study cohort included all patients with a prostate cancer identified in the SNDS by a specific hospital discharge or LTD diagnosis code or a specific treatment dispensing between January 2009 and December 2014, with 5-year history and 3 years of follow-up. Patients with mCRPC were identified using an algorithm integrating time indicators related to metastases management and castration resistance. Both indicators relied on the detection of specific procedures (e.g. imaging, surgery or radiotherapy), drug dispensing (e.g. androgen deprivation therapy, metastases-targeted treatment, chemotherapy) or specific hospitalizations. A complete description of the algorithm and its validation are available elsewhere [18]. In the CAMERRA validation study, so as to ensure the presence of all categories of non-mCRPC patients, three groups of non-mCRPC patients were identified: 34 with non-metastatic hormone-sensitive prostate cancer, 33 with metastatic hormone-sensitive prostate cancer, and 33 with non-metastatic castration-resistant prostate cancer. A single NPV relying on the overall non-mCRPC population was then estimated for the algorithm by weighting false-negative cases according to the actual distribution of the 3 categories of non-mCRPC patient in the prostate cancer population. In a last stage, PPV, NPV and the observed prevalence of mCRPC among the study population (prostate cancer patients), were used to derived the sensitivity and specificity [20].

RESULTS

Diagnostic performance of the MS relapse algorithm

A sample of 200 patients was randomly selected from the initial study population; 100 of them had at least one relapse and 100 did not have any relapse according to the algorithm. The validation committee confirmed 95 patients with relapses (true cases) among the algorithm-identified cases and 96 without relapse among the algorithm-identified non-cases, resulting in a PPV of 95.0% (95%IC = [91; 99]) and a NPV of 96.0% (95%IC = [92; 100]) (Appendix 1-A). After the update of algorithm settings based on experts' conclusions, NPV reached 100.0% (TABLES

Table *1*).

Diagnostic performance of the mCRPC algorithm

A sample of 200 patients was randomly selected from the initial population with prostate cancer; 100 of them were identified as mCRPC and 100 as non-mCRPC according to the algorithm. Experts confirmed 92 of the 100 algorithm-identified mCRPC cases and 93 of the 100 algorithm-identified non-mCRPC cases, resulting in a PPV and NPV of respectively 92.0% (95%CI= [87; 97]) and 99.0% (95%CI= [98; 100]), after weighting according to non-mCRPC cases distribution (Appendix 1-B). Following the algorithm adjustment based on expert feedback, PPV reached 97% (95%CI= [93; 100]) (Table 2). Based on an observed proportion mCRPC/prostate cancer of 3.4%, sensitivity and specificity of the final algorithm were respectively estimated at 80% and 100% [18].

DISCUSSION

Based on two practical examples relying on the French nationwide healthcare database, this paper illustrates an innovative method to assess case-identifying algorithms conducting an intra-database validation study. In both examples, this validation study showed that algorithms had high diagnostic performances, with excellent PPV and NPV. To our knowledge, there are no previous examples of the use of rEHRs to assess the performances of algorithms for case identification, therefore results are difficult to compare with existing data. Because by law, returning to individual medical records from SNDS data is forbidden, most of the currently published French validation studies were limited to the comparison of hospital discharge codes extracted from local hospital databases – before their de-identification and integration to the SNDS – with traditional sources of information such as medical charts or registries, and leading to a PPV varying from 80 to 90% but tending to decrease according to the granularity of the required information [21–27].

In the present case, intra-database validation provides the opportunity to assess algorithms that rely on multiple elements from SNDS, enabling to improve discriminatory abilities compared to single identification criterion or to overcome the absence of a direct-identifying diagnostic code [28]. Experts of the validation committee reported that rEHRs proceeding from SNDS data were on certain points more informative than the usual medical charts and contained a high level of details as well as an accurate chronology regarding patients' journeys, which generally made the adjudication of the cases non-ambiguous. Clinicians insights also allowed to refine the algorithm, adjusting its settings to further improve its performances. This suggests that SNDS data are comprehensive enough to develop a complex algorithm and to validate it.

Validation studies based on medical charts review stay the best way to evaluate claims database algorithms. However, it requires a lot of human time and reliable significant funding, which are often missing, to be able most often to estimate only the PPV. Wherever feasible, validation studies relying on linkage between administrative databases and medical registries or electronic

medical record databases are a good alternative, but they are rarely fully representative of the whole database population, and stay quite long and expensive. Conversely, rEHR review offer a time- and cost-efficient way to conduct validation studies, using the data source accessed by the algorithm. Files to review are standardized and structured, allowing the assessment of hundreds of cases in a limited time: 50 cases per expert-day in the two presented examples. Moreover, as both cases and non-cases are accessible, this approach enables the calculation of other indicators than PPV (e.g. NPV, sensitivity, specificity), with a full representativeness of the population covered by the database.

We acknowledge that validating an algorithm in the same database that was used to develop it may be questionable. The suitability of using an unique data source to generate and evaluate a hypothesis has been previously discussed in the scientific literature, even if the scope was slightly different [29–32]. Walker AM., and Wang SV. and colleagues argued that for such an approach to be considered valid "test data need to be independent of hypothesis-generating data" [29, 30]. Though it is consensual that re-using data to perform quality check, reevaluate findings and strengthen hypotheses (e.g. sensitivity analyses) in the frame of pharmacoepidemiology studies belongs to good research practice, the fact that they can also be used to validate hypotheses is more challenged, especially because of the potential lack of argument to establish causality [31]. In hypothesis-evaluating treatment effectiveness studies, the reuse of data sources is usually not recommended upon the main argument that it leads to replication rather than confirmation [30, 32]. In the present work the lack of argument to establish causality is not an issue, as we do not seek it; the unique objective is to prove that cases identified by the algorithm are true cases. Moreover, here, the independence is ensured by the unrelated approaches used in the identification and the confirmation of the cases: to classify a case, the algorithm picks up information in the database as previously defined in a statistical analysis plan. When experts do so, they choose the relevant information for

themselves in the rEHR. Relying on the same data, the elements considered can be the same (or not), but the approach and the selection process are different, resulting in independent bodies of proof.

Obviously, preference must be given to external data source to conduct validation study, but when it is not feasible, the re-use of the original data source appears as a valuable alternative.

Moreover, it should be borne in mind that the decision whether or not to proceed with intra-database validation for case-identifying algorithm will strongly depend both on the nature of the outcome of interest and on the characteristics of the considered database. Two conditions must be fulfilled to ensure an effective application of the method: 1) the health outcome of interest must be managed by a specific sequence of cares and encounters; 2) the considered healthcare database must capture in an exhaustive way a sufficient number of medical elements in line with the outcome of interest.

Outcome validation should not rely on a unique diagnostic or procedure code but on several tangible elements. As a consequence, intra-database validation should only be considered for health outcomes that are managed in usual clinical practice by a well-defined chronological sequence of cares (procedures, drug dispensings, hospital stays, medical visits, *etc.*) since diagnostic evidence – such as images or laboratory results – may be absent of the database. The succession of healthcare encounters, that individually may be unspecific of the outcome when taken together, gives rise to a specific healthcare pathway. This is particularly true for serious outcomes, mobilizing large healthcare resources. Thus, chronic conditions such as MS (see example 1) or cancer (see example 2), or serious acute outcomes for which the management follows consensual and structured guidelines (e.g. myocardial infarction) [33] seem to be better suited to intra-database validation, compared to non-serious outcomes involving few and unspecific healthcare resources (e.g. acute sore throat) [34], or serious but rare diseases with no clinical practice guidelines [35]. Particular attention must be paid to the clinical guidelines

which were ongoing at the time of the study, since they drive patients journeys and thus, experts' judgment.

Furthermore, in order to ensure that rEHRs provide sufficient and reliable information to enable case adjudication, the underlying healthcare database must capture a sufficient number of medical elements in an exhaustive way over a suitable period of time. Data collected must be, at least, in line with the type of care involved in the management of the outcome of interest (e.g. validation of a myocardial infarction identification requires outpatient data). Ideally, outpatient and inpatient healthcare encounters should be included, and the quality of the captured information regularly assessed. Data completeness, at least over the study period, is mandatory to ensure that the absence of record is synonym of an absence of encounter. The SNDS is particularly well suited to this situation since it fulfills all these requirements: it includes in- and out-patient information of all reimbursed healthcare encounters, most of the time lifelong, and the quality of coding is ensured by regular internal and external audits [9–11].

CONCLUSION

Homogeneous healthcare databases such as the SNDS captures healthcare journey of patients lifelong. Although these data cannot replace the anamnesis and the clinical information reported in patient medical charts, this succession of healthcare records appears to be comprehensive enough to generate consistent rEHRs assessable by experts, allowing to conduct validation studies without using external information. It should be made clear that intra-database validation based on rEHRs review does not pretend to replace traditional methods of validation relying on medical charts review. However, as illustrated here through the MS relapse example and the mCRPC example, in the absence of alternative, such method appears to be a valuable tool to estimate the performances of a case-identifying algorithms and assess their validity. The

development in the coming years of data linkages allowing to gather claims data, registries, electronic health records, *etc.* [36, 37], will further enrich data available for experts to review in rEHRs and may blur the line between intra-validation and external medical chart review.

LIST OF ABREVIATIONS

CEREES: Comité d'Expertise pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé; CNIL: Commission Nationale Informatique & Libertés; FN: false negative; FP: false positive; ICD-10: International Classification of Diseases, 10th revision; LTD: Long Term Disease; mCRPC: metastatic castration-resistant prostate cancer; MS: Multiple Sclerosis; NPV: negative predictive value; PPV: positive predictive value; rEHR: reconstituted Electronic Health Record; SNDS: Système National des Données de Santé; TN: True negative; TP: True positive; 95%CI: 95% confidence interval

REFERENCES

1. Ray WA (2011) Improving Automated Database Studies. Epidemiology 22:302. https://doi.org/10.1097/EDE.0b013e31820f31e1
2. Gavrielov-Yusim N, Friger M (2014) Use of administrative medical databases in population-based research. J Epidemiol Community Health 68:283–287. https://doi.org/10.1136/jech-2013-202744
3. Strom BL (2019) What Is Pharmacoepidemiology? In: Pharmacoepidemiology. John Wiley & Sons, Ltd, pp 1–26
4. Hennessy S (2006) Use of Health Care Databases in Pharmacoepidemiology. Basic

420 Clin Pharmacol Toxicol 98:311–313. https://doi.org/10.1111/j.1742-7843.2006.pto_368.x
421 5. Hashimoto RE, Brodt ED, Skelly AC, Dettori JR (2014) Administrative Database
422 Studies: Goldmine or Goose Chase? Evid-Based Spine-Care J 05:74–76.
423 https://doi.org/10.1055/s-0034-1390027
424 6. Grimes DA (2010) Epidemiologic research using administrative databases: garbage in,
425 garbage out. Obstet Gynecol 116:1018–1019.
426 https://doi.org/10.1097/AOG.0b013e3181f98300
427 7. Lanes S, Brown JS, Haynes K, et al (2015) Identifying health outcomes in healthcare
428 databases. Pharmacoepidemiol Drug Saf 24:1009–1016. https://doi.org/10.1002/pds.3856
429 8. Georgescu I, Hartmann FGH (2013) Sources of financial pressure and up coding
430 behavior in French public hospitals. Health Policy 110:156–163.
431 https://doi.org/10.1016/j.healthpol.2013.02.003
432 9. Gilleron V, Gasnier-Duparc N, Hebbrecht G (2018) Certification des comptes: Une
433 incitation à la traçabilité des processus de contrôle. Revue Hospitaliere de France 582:6
434 10. Marescaux C (2011) Entre soin et contrôle de gestion : place du DIM dans
435 l'organisation hospitalière. Inf Psychiatr Volume 87:487–491
436 11. Caeyseele T, Bruandet A, Delaby F, Theis D (2016) Création d'un outil de gestion des
437 contrôles qualités du codage au DIM du CHRU de Lille. Rev DÉpidémiologie Santé Publique
438 64:S20. https://doi.org/10.1016/j.respe.2016.01.066
439 12. Carnahan RM (2012) Mini-Sentinel's systematic reviews of validated methods for
440 identifying health outcomes using administrative data: summary of findings and suggestions
441 for future research: HEALTH OUTCOME ALGORITHM SUMMARY. Pharmacoepidemiol
442 Drug Saf 21:90–99. https://doi.org/10.1002/pds.2318
443 13. Carnahan RM, Moores KG (2012) Mini-Sentinel's systematic reviews of validated
444 methods for identifying health outcomes using administrative and claims data: methods and

445 lessons learned: HEALTH OUTCOME ALGORITHM REVIEW METHODS.
446 Pharmacoepidemiol Drug Saf 21:82–89. https://doi.org/10.1002/pds.2321
447 14. van Walraven C, Bennett C, Forster AJ (2011) Administrative database research
448 infrequently used validated diagnostic or procedural codes. J Clin Epidemiol 64:1054–1059.
449 https://doi.org/10.1016/j.jclinepi.2011.01.001
450 15. Bezin J, Duong M, Lassalle R, et al (2017) The national healthcare system claims
451 databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology.
452 Pharmacoepidemiol Drug Saf 26:954–962. https://doi.org/10.1002/pds.4233
453 16. Tuppin P, Rudant J, Constantinou P, et al (2017) Value of a national administrative
454 database to guide public decisions: From the système national d'information interrégimes de
455 l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in
456 France. Rev Epidemiol Sante Publique 65 Suppl 4:S149–S167.
457 https://doi.org/10.1016/j.respe.2017.05.004
458 17. Bosco-Levy P, Debouverie M, Brochet B, et al Comparative effectiveness of dimethyl
459 fumarate versus other disease-modifying therapies in Multiple Sclerosis: a large population-
460 based cohort study. Under Submission
461 18. Thurin NH, Rouyer M, Gross-Goupil M, et al (2020) Epidemiology of metastatic
462 castration-resistant prostate cancer: A first estimate of incidence and prevalence using the
463 French nationwide healthcare database. Cancer Epidemiol 69:101833.
464 https://doi.org/10.1016/j.canep.2020.101833
465 19. Gross-Goupil M, Thurin NH, Rouyer M, et al (2020) Survival outcome in patients
466 with metastatic castration-resistant prostate cancer according to first-line treatment. J Clin
467 Oncol 38:5570–5570. https://doi.org/10.1200/JCO.2020.38.15_suppl.5570
468 20. Bollaerts K, Rekkas A, Smedt TD, et al (2020) Disease misclassification in electronic
469 healthcare database studies: Deriving validity indices—A contribution from the ADVANCE

project. PLOS ONE 15:e0231333. https://doi.org/10.1371/journal.pone.0231333

21. Bosco-Lévy P, Duret S, Picard F, et al (2019) Diagnostic accuracy of the International Classification of Diseases, Tenth Revision, codes of heart failure in an administrative database. Pharmacoepidemiol Drug Saf 28:194–200. https://doi.org/10.1002/pds.4690

22. Coureau G, Baldi I, Savès M, et al (2012) [Performance evaluation of hospital claims database for the identification of incident central nervous system tumors compared with a cancer registry in Gironde, France, 2004]. Rev Epidemiol Sante Publique 60:295–304. https://doi.org/10.1016/j.respe.2012.02.003

23. Giroud M, Hommel M, Benzenine E, et al (2015) Positive Predictive Value of French Hospitalization Discharge Codes for Stroke and Transient Ischemic Attack. Eur Neurol 74:92–99. https://doi.org/10.1159/000438859

24. Goueslard K, Cottenet J, Benzenine E, et al (2020) Validation study: evaluation of the metrological quality of French hospital data for perinatal algorithms. BMJ Open 10:e035218. https://doi.org/10.1136/bmjopen-2019-035218

25. Mezaache S, Derumeaux H, Ferraro P, et al (2017) Validation of an algorithm identifying incident primary immune thrombocytopenia in the French national health insurance database. Eur J Haematol 99:344–349. https://doi.org/10.1111/ejh.12926

26. Palmaro A, Gauthier M, Conte C, et al (2017) Identifying multiple myeloma patients using data from the French health insurance databases. Medicine (Baltimore) 96:. https://doi.org/10.1097/MD.0000000000006189

27. Prat M, Derumeaux H, Sailler L, et al (2018) Positive predictive values of peripheral arterial and venous thrombosis codes in French hospital database. Fundam Clin Pharmacol 32:108–113. https://doi.org/10.1111/fcp.12326

28. Fuentes S, Cosson E, Mandereau-Bruno L, et al (2019) Identifying diabetes cases in health administrative databases: a validation study based on a large French cohort. Int J Public

Health 64:441–450. https://doi.org/10.1007/s00038-018-1186-3

29. Walker AM (2010) Orthogonal predictions: follow-up questions for suggestive data. Pharmacoepidemiol Drug Saf 19:529–532. https://doi.org/10.1002/pds.1929

30. Wang SV, Kulldorff M, Glynn RJ, et al (2018) Reuse of data sources to evaluate drug safety signals: When is it appropriate? Pharmacoepidemiol Drug Saf 27:567–569. https://doi.org/10.1002/pds.4442

31. Gould AL (2010) Generating and confirming hypotheses. Pharmacoepidemiol Drug Saf 19:533–536. https://doi.org/10.1002/pds.1928

32. Berger ML, Sox H, Willke RJ, et al (2017) Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. Pharmacoepidemiol Drug Saf 26:1033–1039. https://doi.org/10.1002/pds.4297

33. Ibanez B, James S, Agewall S, et al (2018) 2017 ESC Guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevationThe Task Force for the management of acute myocardial infarction in patients presenting with ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J 39:119–177. https://doi.org/10.1093/eurheartj/ehx393

34. Pelucchi C, Grigoryan L, Galeone C, et al (2012) Guideline for the management of acute sore throat: ESCMID Sore Throat Guideline Group. Clin Microbiol Infect 18:1–27. https://doi.org/10.1111/j.1469-0691.2012.03766.x

35. Pavan S, Rommel K, Mateo Marquina ME, et al (2017) Clinical Practice Guidelines for Rare Diseases: The Orphanet Database. PLoS ONE 12:. https://doi.org/10.1371/journal.pone.0170365

36. Bradley CJ, Penberthy L, Devers KJ, Holden DJ (2010) Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future. Health Serv Res 45:1468.

https://doi.org/10.1111/j.1475-6773.2010.01142.x

37. Scailteux L-M, Droitcourt C, Balusson F, et al (2019) French administrative health care database (SNDS): The value of its enrichment. Therapie 74:215–223. https://doi.org/10.1016/j.therap.2018.09.072

TABLES

Table 1. Positive (PPV) and negative (NPV) predictive values of the final algorithm for the identification of relapse in multiple sclerosis

		Validation committee		
		Relapse +	**Relapse -**	**Total**
Algo-ithm	**Relapse +**	99	5	104
	Relapse -	0	96	96
	Total	99	101	200

PPV = 95% (95%CI = [91; 99])
NPV = 100%

Table 2. Positive (PPV) and negative (NPV) predictive values of the final algorithm for the identification of metastatic castration-resistant prostate cancer (mCRPC), adapted from Thurin NH, et al. 2020

		Validation committee		
		mCRPC +	**mCRPC -**	**Total**
Algo-ithm	**mCRPC +**	90	3	93
	mCRPC -	1.23*	105.77*	107
	Total	91.23	108.77	200

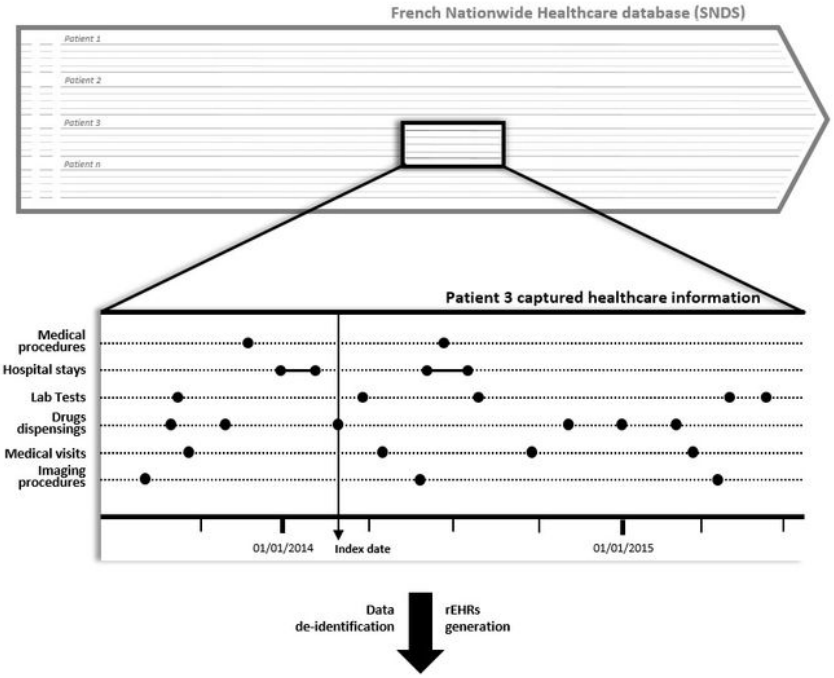
PPV = 97% (95%CI= [93; 100])
NPV = 99% (95%CI= [97; 100])

**After weighting*

FIGURE CAPTIONS

Fig. 1 Generation of anonymized reconstituted Electronic Health Records from the French Nationwide Healthcare database (SNDS) to conduct intra-database case-identifying algorithm validation

Figures



Anonymized Reconstituted Electronic Health Records (rEHR) of Patient n°3

Year of inclusion (index year): 05/03/2014 **Age class at the inclusion:** [65 - 70[years **Gender:** Male

Type of healthcare encounter	Encounter start date	Encounter end date	Description
...	...	...	...
Hospital stay	-1 month 5 days	-1 month days	ICD-10 principal (PD), related (RD) and associated (AD) discharge diagnoses *e.g. PD N10 - acute pyelonephritis / AS B962 - escherichia coli, as the cause of diseases classified elsewhere*
Drug dispensing	0		Drug name, dosage, form, quantity supplied, ATC code, INN *e.g. ZYTIGA 250 MG CPR 120 - ATC L02BX03 (ABIRATERONE)*
Lab test	+1 month 14 days		Code and description of the lab test *e.g. Prostate Specific Antigen (PSA): dosage*
Medical visit	+2 months 3 days		Physician specialty *e.g. General practioner*
Imaging procedure	+3 months 1 day		Code and description of the imaging procedure *e.g. ZCQH001 – Computed tomography of the abdomen and small pelvis, with intravenous injection of contrast material*
Hospital stay	+3 months 18 days	+3 months 28 days	ICD-10 principal (PD), related (RD) and associated (AD) discharge diagnoses *e.g. PD I702 - Atherosclerosis of native arteries of the extremities*
Medical procedure	+3 months 20 days		Code en description of the medical procedure *e.g EDFA007 - iliofemoral thromboendarterectomy, by inguinofemoral approach*
...	...	...	...

Figure 1

Generation of anonymized reconstituted Electronic Health Records from the French Nationwide Healthcare database (SNDS) to conduct intra-database case-identifying algorithm validation

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- APPENDIX1.pdf