Design and validation of algorithms to identify Venous Thromboembolism in the French National Healthcare Database

Nicolas Thurin[1], Angela Grelaud[1], Adeline Grolleau[1], Marie-Agnès Bernard[1], Emmanuelle Bignon[1], Patrick Blin[1], Régis Lassalle[1], and Cécile Droz-Perroteau[1]

[1]Centre d'Investigation Clinique Plurithematique de Bordeaux

April 28, 2023

Design and validation of algorithms to identify Venous Thromboembolism in the French National Healthcare Database

Authors

Nicolas H Thurin[1], Angela Grelaud[1], Adeline Grolleau[1], Marie-Agnès Bernard[1], Emmanuelle Bignon[1], Patrick Blin[1], Régis Lassalle[1], Cécile Droz-Perroteau[1]

[1] Bordeaux PharmacoEpi, INSERM CIC-P 1401, Univ. Bordeaux, Bordeaux , France

Corresponding author:
Nicolas H Thurin
nicolas.thurin@u-bordeaux.fr
146 rue Léo Saignat – case 41
CS 61292 - 33076 Bordeaux cedex - France
Tel : +33 557 57 92 09

Ethics Statement

The work presented in this paper was conducted in the framework of the B023 post-authorization safety study (EUPAS #32271, EUPAS#34204), sponsored by Eli Lilly and Company, and for which access to the SNDS has been granted by the CNIL in France (reference 919392).

Conflict of Interest Statement

Co-authors are researchers at Bordeaux PharmacoEpi, a research platform of the University of Bordeaux and its subsidiary the ADERA, which performs financially supported studies for public and private partners, in compliance with the ENCePP Code of Conduct.

Abstract

Purpose

This paper aims to introduce the algorithm designed to identify Venous Thromboembolism (VTE) in the French National healthcare database (SNDS) and to estimate its positive predictive value.

Methods

A case-identifying algorithm was designed using SNDS inpatient and outpatient encounters, including hospital stays with discharge diagnoses, imaging procedures, and drug dispensing. An intra-database validation study was then conducted, drawing 150 cases identified as VTE by the algorithm and requesting 4 vascular specialists to assess them. Patient profiles used to conduct the case adjudication were reconstituted from de-identified pooled and formatted SNDS data with a 6-month look-back period prior to the supposed VTE onset and a 12-month follow-up period after. The positive predictive value (PPV) with its 95% confidence interval [95%CI] was calculated as the number of experts-confirmed VTE divided by the number of algorithm-identified VTE. The PPV and its 95%CI were then recomputed for the same patients once the VTE definition was updated based on expert recommendation.

Results

On the 150 patients meeting the VTE first definition, the adjudication committee confirmed 92 cases, resulting in a PPV of 61% (95%CI = [54-69]). The final definition including expert suggestions showed a PPV of 92% (95%CI = [86-98]) with a total of 87 algorithm-identified cases, including 80 retrieved from the 92 confirmed by experts

Conclusion

The identification of VTE in the SNDS is possible with a good PPV.

Key points

- Identification of Venous Thromboembolisms is possible in the French National healthcare database with a 92% positive predictive value
- Re-identification of patients in the French National Healthcare database is forbidden, making it impossible to conduct classical validation studies
- Anonymized patient profiles reconstituted from claims data are relevant materials to conduct intra-database case adjudication, in the absence of alternative
- The presence of a clearly identifiable care sequence around the event to be validated is essential to the success of an intra-database validation study
- Feedback from experts following the validation of cases is an important element for improving case-validation algorithms

Purpose

Venous Thromboembolisms (VTE) is often included as an outcome of interest in the framework of Post-Authorization Safety Studies (PASS) required by the health authorities or volunteered by Marketing Authorization Holder. As for all outcomes of interest, the identification of VTE in a data source may vary according to the nature of the data captured (e.g., the presence or absence of a diagnosis code and/or related procedures and/or related drug dispensing), and the ontology used to characterize them, when available. The healthcare settings (outpatient or inpatient) may also impact the way outcomes are coded and recorded (1). As a consequence, accurately identifying an outcome, such as VTE, in a data source requires the implementation of a specific algorithm to either refine a diagnosis code or overcome its absence.

Although the awareness of the risks related to outcome misclassification is high (2), the implementation of classical validation studies relying on the linkage between claims records and patients' charts is not always possible for financial, technical, or legal issues. A previous work conducted in data from the French National Healthcare Database (*Système National des Données de Santé* – SNDS) echoed by a German study demonstrated that anonymized patients profiles reconstituted from claims data (*i.e.*, reconstituted electronic health records – rEHR) were a relevant material to conduct case adjudication (3,4).

The objective of this paper is to introduce the algorithm designed in the French SNDS to identify VTE in outpatient and inpatient settings, as well as the results of the intra-database validation conducted to assess its performance. This algorithm was developed, implemented, and validated in the framework of an international PASS study focused on the safety of baricitinib in patients with rheumatoid arthritis (5).

Methods

Data source

This validation study was 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 (6). 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. Therefore, 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. Reimbursed dispensed drugs can be identified at the product level with the form and dosage in outpatient settings, and in inpatient settings when billed in addition to the hospital stays. This is also the case for medical devices. Performed laboratory tests and procedures are available but without their results. They are respectively recorded using the *Nomenclature des Actes de Biologie Médicale* (NABM) and the *Classification Commune des Actes Médicaux* (CCAM). Registration for Long Term Disease (LTD) – a status that ensures 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). Diagnoses made by outpatient practitioners are not recorded.

Access to the SNDS data for this PASS was approved by the French national data protection agency (*Commission Nationale de l'Informatique et des Libertés* – CNIL).

Population

As per the PASS study protocol the study population extracted from the SNDS consists of French patients aged at least 18 years old with a prior hospital discharge diagnosis of rheumatoid arthritis to whom baricitinib or Tumor Necrosis Factor Inhibitors (TNFi) were dispensed between September 1st, 2017, and December 31st, 2018.

VTE Case-identifying algorithm

A first version of the VTE-identifying algorithm was developed and implemented in the population. This first definition relied on the following indicators: (a) a hospital stay with a VTE primary discharge diagnosis, or (b) a hospital stay with a VTE-associated diagnosis followed by a dispensing of an anticoagulant with curative dosage within 31 days, or (c) an imaging procedure surrounded by a dispensing of an anticoagulant with curative dosage (±2 days).

This first perfectible definition was refined following the adjudication exercise to arrive at the final definition:

- (a) a hospital with a VTE primary discharge diagnosis, or
- (b+c) an imaging procedure or a hospital stay with a VTE-associated diagnosis,
    - AND an anticoagulant dispensing with curative dosage within 3 days before or after the encounter, without more than one dispensing of the same drug in the previous 6 months
    - AND an anticoagulant coverage ≥30 days over the subsequent 120 days (substituted by a coverage ≥50% for patients who died), or a hospital stay with a VTE diagnosis (any position) in the subsequent 120 days.

All codes used are presented in Supplementary Material.

Validation approach

The overall validation process followed the methodology introduced by Thurin NH *et al. (3).* Medical information available in the SNDS was de-identified pooled and formatted to generate rEHRs with a 6-month look-back period prior to the supposed VTE onset and a 12-month follow-up period after. 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 the outcome outset, location details were deleted and only age classes were displayed.

The adjudication committee was constituted of 2 pairs of vascular specialists. Each pair blindly adjudicated the status of 75 patients drawn among those identified by the algorithm according to the VTE first definition (*i.e.,* a total of 150 out of 1 103 patients). In case of discrepancy within a pair, the case was discussed by the 4 experts. The positive predictive value (PPV) with its 95% confidence interval [95%CI] was calculated as the number of experts-confirmed VTE divided by the number of algorithm-identified VTE. The PPV and its 95%CI were then recomputed for the same patients once the VTE definition was updated. Experts were also requested to categorize patients according to 4 clinical definitions: pulmonary embolism, lower extremity deep vein thrombosis, upper extremity deep venous thrombosis, deep venous thrombosis with unspecified localization, or unspecified thromboembolic event.

Results

On the 150 patients meeting the VTE first definition, the adjudication committee confirmed 92 cases, resulting in a PPV of 61% (95%CI = [54-69]). Out of the 92 cases confirmed, experts were able to identify 38 pulmonary embolisms, 65 lower extremity deep vein thrombosis, 6 upper extremities deep venous thrombosis, 4 deep venous thromboses (unspecified localization), and 1 unspecified thromboembolic event. Twenty-two cases presented both pulmonary embolism and lower extremity deep vein thrombosis.

The final definition including expert suggestions showed a PPV of 92% (95%CI = [86-98]) with a total of 87 algorithm-identified cases, including 80 retrieved from the 92 confirmed by experts (Table 1). False positives identified by experts were mainly related to the management in outpatient settings of other cardiovascular conditions requiring anticoagulation such as strokes, atrial fibrillation, but also superficial venous thrombosis.

Table 1. Contingency table for the final Venous Thromboembolism (VTE) definition

		Experts		
		VTE +	**VTE -**	**Total**
Algorithm	**VTE +**	80	7	87
	VTE -	12	51	63
	Total	92	58	150

Conclusion

This paper introduced an algorithm designed to identify VTE in outpatient and inpatient settings in the SNDS. The intra-database validation study showed that the final version of the algorithm had a 92% PPV (95%CI = [86-98]). The validation approach used is not without its limitations. Despite the richness and completeness of the information available to reconstitute patient profiles and conduct the case adjudication, claims data lack of clinical information such as laboratory test values or other diagnostics results. The same limitations apply to the algorithm itself, especially in outpatient settings for which the SNDS does not capture diagnosis codes but just procedure codes, medical visits, and dispensed drugs. However, experts reported that the sequence of care set in motion to manage VTE was most of the time specific enough to assess the reliability of the case with few clinical elements. Other examples of validated VTE-identifying algorithms with a PPV ranging from 75.5% when relying on encounters from in and outpatient settings, to 95% for outcomes identified exclusively in hospital settings are reported in the recent literature (5,7). Although these numbers may not be directly comparable because they are not derived from similar validation approaches (real patient charts *versus* rEHRs), the performances estimated in the present work appear consistent.

Finally, all these elements show the ability to identify VTE in the SNDS with a good PPV.

References

1. Willame C, Dodd C, Durán C, Elbers R, Gini R, Bartolini C, et al. Background rates of 41 adverse events of special interest for COVID-19 vaccines in 10 European healthcare databases - an ACCESS cohort study. Vaccine. nov 2022;S0264410X22014293.

2. Hall GC, Lanes S, Bollaerts K, Zhou X, Ferreira G, Gini R. Outcome misclassification: Impact, usual practice in pharmacoepidemiology database studies and an online aid to correct biased estimates of risk ratio or cumulative incidence. Pharmacoepidemiol Drug Saf. nov 2020;29(11):1450-5.

3. Thurin NH, Bosco-Levy P, Blin P, Rouyer M, Jové J, Lamarque S, et al. Intra-database validation of case-identifying algorithms using reconstituted electronic health records from healthcare claims data. BMC Med Res Methodol. 1 mai 2021;21(1):95.

4. Platzbecker K, Voss A, Reinold J, Elbrecht A, Biewener W, Prieto-Alhambra D, et al. Validation of Algorithms to Identify Acute Myocardial Infarction, Stroke, and Cardiovascular Death in German Health Insurance Data. Clin Epidemiol. 10 nov 2022;14:1351-61.

5. Salinas CA, Louder A, Polinski J, Zhang TC, Bower H, Phillips S, et al. Evaluation of VTE, MACE, and Serious Infections Among Patients with RA Treated with Baricitinib Compared to TNFi: A Multi-Database Study of Patients in Routine Care Using Disease Registries and Claims Databases. Rheumatol Ther [Internet]. 13 nov 2022 [cité 17 nov 2022]; Disponible sur: https://link.springer.com/10.1007/s40744-022-00505-1

6. Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, et al. The national healthcare system claims databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf. août 2017;26(8):954-62.

7. Molander V, Bower H, Askling J. Validation and characterization of venous thromboembolism diagnoses in the Swedish National Patient Register among patients with rheumatoid arthritis. Scand J Rheumatol. mars 2023;52(2):111-7.