

Harmonizing Case Validation for Vaccine Safety in Real-World Data: The VAC4EU Experience

Amirreza Dehghan Tarazjani¹, Daniel Weibel^{1,2}, Taylor Aurelius^{3,4}, Laura C Zwiers^{5,6}, Jesse M Van den Berg⁷, Lina Pérez-Breva⁸, Antonio Gimeno-Miguel^{9,10}, Luca Stona¹¹, Martín Solórzano^{1,12}, Anteneh Assefa Desalegn¹³, Beatriz Poblador-Plou^{9,10}, Thom S Lysen⁷, Jannik Wheler¹⁴, Mahmoud Zidan¹³, Juan José Carreras^{8,15}, Felipe Villalobos¹², Vera Ehrenstein¹⁴, Kathryn Morton^{3,4}, Cristina Rebordosa¹⁶, Fariba Ahmadizar^{1,17}, Joan Fortuny¹⁶, Alejandro Arana¹⁶, Miriam Sturkenboom^{1,2}

¹Department of Data Science and Biostatistics, Julius Global Health, University Medical Center Utrecht, Utrecht, The Netherlands; ²Vaccine Monitoring Collaboration for Europe (VAC4EU), Brussels, Belgium; ³Drug Safety Research Unit (DSRU), Southampton, UK; ⁴University of Portsmouth, Portsmouth, UK; ⁵Julius Global Health, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; ⁶Julius Clinical, Zeist, The Netherlands; ⁷PHARMO Institute for Drug Outcomes Research, Utrecht, The Netherlands; ⁸Vaccine Research Department, Foundation for the Promotion of Health and Biomedical Research in the Valencian Region (FISABIO – Public Health), Valencia, 46020, Spain; ⁹Research Network on Chronicity, Primary Care, and Health Promotion (RICAPPS), Institute of Health Carlos III (ISCIII), Madrid, 28029, Spain; ¹⁰EpiChron Research Group, Aragon Health Sciences Institute (IACS), Aragon Health Research Institute (IIS Aragón), Miguel Servet University Hospital, Zaragoza, 50009, Spain; ¹¹Fondazione Penta ETS, Padua, Italy; ¹²Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain; ¹³Pharmacoepidemiology and Drug Safety Research Group, Department of Pharmacy, Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Norway; ¹⁴Department of Clinical Epidemiology, Center for Population Medicine, Department of Clinical Medicine, Aarhus University and Aarhus University Hospital, Aarhus, Denmark; ¹⁵Biomedical Research Consortium of Epidemiology and Public Health (CIBER-ESP), Instituto de Salud Carlos III, Madrid, Spain; ¹⁶Department of Epidemiology and Risk Management, RTI Health Solutions, Barcelona, Spain; ¹⁷General Practice Department, Amsterdam University Medical Center, Amsterdam, The Netherlands

Correspondence: Miriam Sturkenboom, Department of Data Science and Biostatistics, Julius Global Health, University Medical Center Utrecht, Utrecht, The Netherlands, Email M.C.J.Sturkenboom@umcutrecht.nl

Purpose: Applying standardized Brighton Collaboration (BC) case definitions retrospectively to heterogeneous real-world data (RWD) is challenging due to inconsistent clinical detail and data structures across settings. To address this, the Vaccine Monitoring Collaboration for Europe (VAC4EU) developed a structured validation pipeline that operationalizes BC case definitions for vaccine safety outcomes in RWD in a harmonized, scalable, and reusable manner.

Methods: VAC4EU developed a systematic, stepwise approach to validate vaccine safety outcomes. BC case definitions were utilized when available, adapted for RWD as needed, and newly developed for outcomes without an existing BC definition. The approach involves: 1) Critical review and adaptation of BC definition by clinical and RWD experts; 2) Creation of dummy cases based on published reports; 3) Creation of REDCap electronic data collection forms (eDCF) incorporating decision logic to assigned levels of certainty (LOC); 4) Iterative testing of decision logic; and 5) comprehensive training of abstractors with real-time feedback. A dedicated task force assigned reference LOCs for dummy cases. Inter-rater reliability was measured using Fleiss' kappa (κ) by comparing abstractor LOCs to the reference standard.

Results: The pipeline was applied to 16 COVID-19 vaccine safety outcomes, of which 13 had existing BC definition. In total, 78 dummy case descriptions were developed across the outcomes for training of abstractors, and 15 REDCap eDCFs were created. Myocarditis and pericarditis shared an eDCF. Across 33 abstractors, 747 dummy case abstractions were completed. Agreement analysis showed 93 discrepancies (12.4%) and moderate overall concordance ($\kappa = 0.55$, across all outcomes), with the lowest for thrombosis with thrombocytopenia syndrome ($\kappa = -0.05$).

Conclusion: The VAC4EU validation pipeline provides a standardized framework for training and validating vaccine safety outcome in RWD. By adapting BC case definitions and dedicated training of abstractors, we will reduce variability in outcome validation in post-authorization safety studies.

Keywords: validation, vaccine safety, real-world data, Brighton collaboration

Introduction

Vaccines prevent millions of deaths and diseases annually.¹ Monitoring their benefit-risk balance is essential throughout the product lifecycle.² As vaccines are administered to generally healthy individuals, the tolerance for safety issues is much lower compared with therapeutic products.³ In this context, real-world evidence (RWE) derived from real-world data (RWD) plays a crucial role in evaluating the safety and effectiveness of vaccines in the post-licensure phase.

Health authorities such as the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA) now widely support the use of RWD in vaccine safety studies, provided that outcomes are rigorously validated to ensure data reliability, relevance, and traceability.⁴ In particular, the FDA's recent guidance on RWD and RWE underscores the critical role of validation processes in regulatory decision-making, particularly when using electronic health records (EHRs) and medical claims data.⁴ The guidance emphasizes the importance of data reliability, defined as the accuracy and consistency of data over time; relevance, referring to how well the data elements reflect the concepts of interest; and traceability, which allows verification of the data source and transformations. It prioritizes comprehensive data capture, mitigation of missing information, and effective data linkage. The guidance also recommends evaluating whether data sources adequately represent study populations, exposures, and outcomes. Suggested validation approaches include complete verification or performance assessments through sampling strategies, tailored to specific study contexts.⁴

The European Network of Centers in Pharmacoepidemiology & Pharmacovigilance (ENCePP) Methods Guide highlights the potential impact of misclassification of exposure, outcomes and covariates on study results and advocates for estimating sensitivity, specificity, and positive predictive values.⁵ Misclassification can introduce bias into effect estimates and validation of outcomes can provide an estimate of the error and bias, manual validation against medical charts remains the gold standard for outcome validation; however, algorithm-based or indirect approaches may be acceptable alternatives when chart review is not feasible.^{6–8}

The EMA's Guideline on Good Pharmacovigilance Practices (GVP) emphasizes the use of routinely collected healthcare data, such as electronic health records, disease registries, hospital discharge databases, or insurance claims, for large-scale studies, despite challenges like incomplete data and limited follow-up. The EMA stresses selecting data sources that balance validity and efficiency while adhering to privacy regulations.⁹

The Brighton Collaboration (BC) aims to harmonize vaccine safety data collection and has created more than 60 case definitions for various outcomes.¹⁰ The BC case definitions aim to provide an ordinal Level of Certainty (LOC) scale that classifies cases by diagnostic certainty based on the strength of available evidence. They are well recognized by regulators and are intended to be applicable in different types of data collections: prospectively in clinical trials, during passive reporting, and retrospectively in surveillance systems or RWD.¹¹ BC case definitions are typically developed by clinical experts with a prospective mindset, and implementation for retrospective use in heterogeneous RWD is more challenging because the required clinical details are inconsistently recorded, moreover the level of recording and access to source data varies largely across different RWD systems. Although the BC criteria are explicit, absence of information in RWD may be interpreted differently by different abstractors. During the H1N1 pandemic vaccine safety studies, application of BC case definitions across countries yielded high variability in LOC for Guillain-Barré syndrome (GBS) and narcolepsy cases, across settings.^{12–14} A Danish study also described that traditional trial-based definitions may not directly fit real-world data.¹⁵

The Vaccine Monitoring Collaboration for Europe (VAC4EU) is a non-for-profit association of institutions that aim to generate high-quality evidence on vaccines to support regulatory decision-making worldwide. The VAC4EU validation pipeline was set up because of the need of a harmonized approach to validation in the context of five multi-site Post-Authorization Safety Studies (PASS) on COVID-19 vaccines. This paper describes how we adapted BC case definitions for RWD, and trained abstractors for outcome validation using dummy case descriptions for each outcome. It does not show the results of the validation of cases that were identified in the actual studies, as these will be reported in the study reports.

Methods

Setting

VAC4EU facilitates the harmonized implementation of vaccine studies with RWD in a distributed manner across different organizations who may have different roles. Roles include lead scientific center (LSC), lead operating center and data expert and access providers (DEAPs). The validation pipeline was designed based on 5 regulatory required PASS studies on COVID-19 vaccines which were registered in the EU PAS registry (EUPAS numbers 41623, 47708, 45362, 105009, and 43556) which were conducted by different LSCs and lead operating centers. A total of 8 different DEAPs were contracted to conduct validation as part of one or more studies and each of them designated one or more clinical abstractors to validate identified cases in a retrospective manner. The validation task was coordinated by a VAC4EU validation task force composed of principal investigators of the different studies and medically trained experts. VAC4EU uses the ConcePTION common data model (CDM) for common analytics across sites and DEAPs need to transform their data to this CDM.¹⁶ DEAPs process the data in accordance with the local implementation of the General Data Protection Regulation (GDPR) and data stay local.¹⁷ Each study site was responsible for obtaining all required ethics committee or Institutional Review Board (IRB) approvals for the validation activity, which was monitored by the LOC.

The VAC4EU validation task force developed a stepwise process to allow for implementation of harmonized validation of the identified potential cases during 1 April 2023 and 30 October 2024.

The step-by-step workflow of this pipeline is illustrated in [Figure 1](#) and described below.

Definitions of Outcome Variables

The first step in the validation process is to identify case definitions for outcomes that require validation. VAC4EU adopted the principle of using BC case definitions, because of their regulatory adoption. If BC definitions were not available, the VAC4EU validation task force created definitions based on a literature review. A total of 16 outcomes required validation in one or more of the PASS, for 13 of these safety outcomes, a BC definition was available. Outcomes included GBS, narcolepsy, idiopathic thrombocytopenic purpura (ITP), thrombosis with thrombocytopenia syndrome (TTS), thrombocytopenia with bleeding, myocarditis, pericarditis, transverse myelitis, anaphylaxis, major congenital anomalies, encephalopathy (including acute disseminated encephalomyelitis, ADEM), deep vein thrombosis (DVT), pulmonary embolism (PE), hemorrhagic stroke, non-hemorrhagic stroke, and cerebral venous sinus thrombosis (CVST).

Review and Adaptation of Data Extraction Forms

The VAC4EU validation task force conducted a comprehensive review of publicly available data extraction forms from the BC case definition and adapted them, where necessary, to better align with data likely to be available in an RWD setting ([supplemental material 1](#)). A key challenge addressed by the task force during this phase was to clarify how abstractors should handle the absence of information. When clinical records are accessed retrospectively, absence of information can mean either absence of the condition or that the presence of the condition has not been recorded. In general, in RWD, there is information on which procedures, such as diagnostic tests, imaging, or treatments, were done and which diagnoses were made, but limited information on which conditions were actively excluded, or which procedures were not performed. To reduce heterogeneity in interpretation by different abstractors, the VAC4EU validation task force aimed to make the data extraction questionnaires as unambiguous as possible. Simplified question formats were used to minimize complexity, and specific guidance documents were provided to assist abstractors.

REDCap Data Collection Tool

VAC4EU uses the REDCap system for data collection to deploy a common data collection structure while keeping data local.^{18,19} The platform was centrally configured by Vall d'Hebron Hospital, which served as the VAC4EU REDCap managing center during the development and testing phases. Dedicated REDCap projects with outcome-specific electronic data collection forms (eDCFs) were developed and subsequently cloned for local implementation at participating sites. A quality review of the eDCF was conducted by the VAC4EU validation task force using the dummy cases.

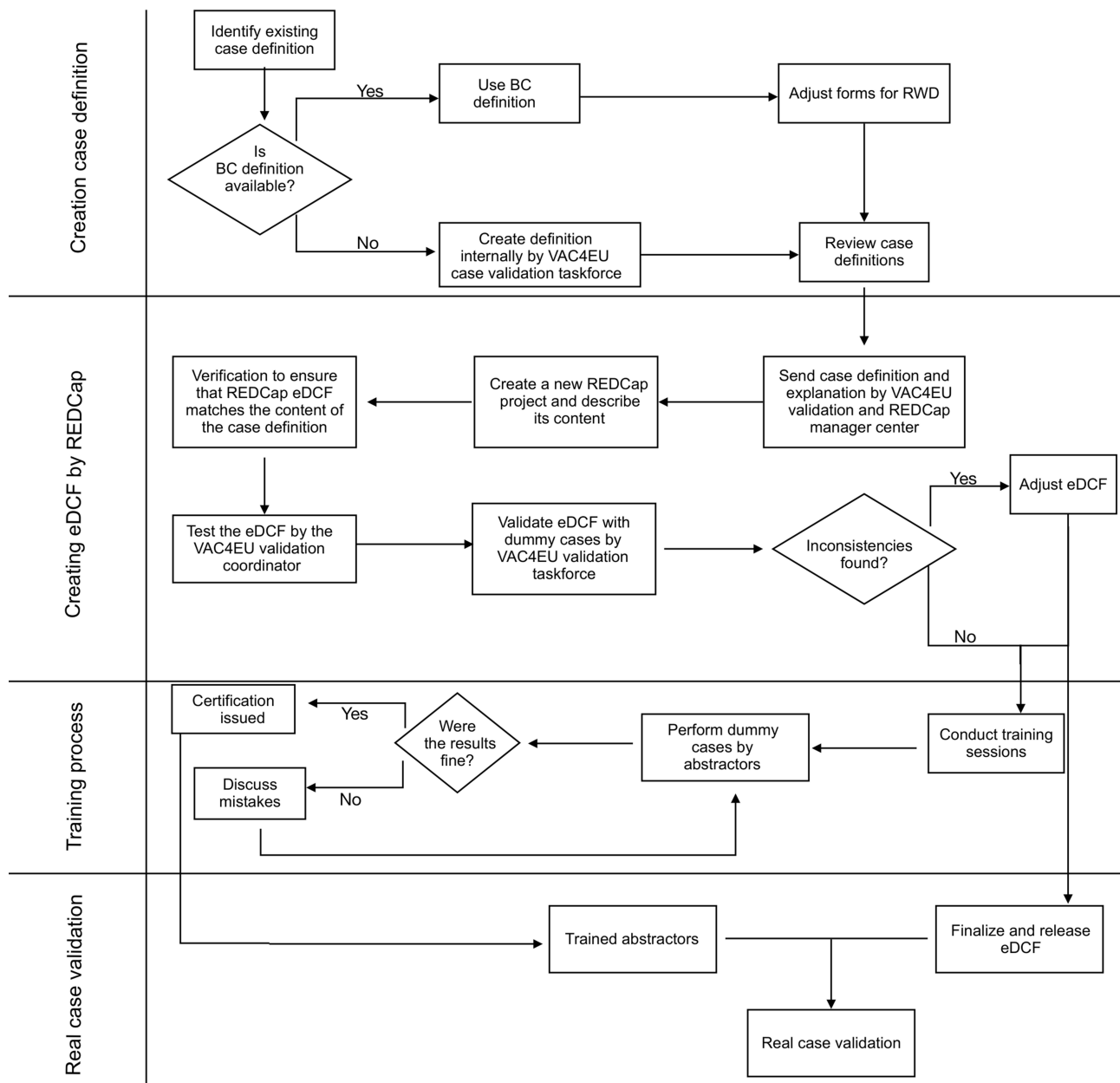


Figure 1 VAC4EU Case Validation Workflow.

Abbreviations: BC, Brighton Collaboration; RWD, real-world data; REDCap, Research Electronic Data Capture; eDCF, electronic data collection form; VAC4EU, Vaccine Monitoring Collaboration for Europe.

Dummy Cases

The VAC4EU validation task force created a set of five dummy case narratives for each outcome to train the abstractors and collect feedback on the eDCFs. These cases were adapted from published case reports and modified to ensure blinding to vaccination status. Each case was designed to represent a different LOC, allowing full coverage of all LOC categories during training.

Training of Abstractors

The VAC4EU validation coordinator conducted training sessions to train validation abstractors from each DEAP on how to use the data collection forms and guidelines, discuss complex items, and resolve questions through examples. Each

abstractor was required to extract the dummy cases and enter the information in the eDCF. The VAC4EU validation coordinator (AD) reviewed the results and discrepancies in the assigned LOCs were discussed.

Individual training certificates were issued after successfully completing the training. A follow-up meeting was held to discuss any doubts or challenges encountered during the dummy cases validation, particularly for cases where discrepancies were observed between the abstractors' results and the expected outcomes defined by the VAC4EU validation task force.

Finalization of eDCF

Based on insights gained during the training, the VAC4EU validation task force reviewed the abstractors' feedback. Any fields in the eDCF or logic that caused confusion or produced varied responses were discussed, and necessary adjustments were made to improve clarity and accuracy.

After incorporating these refinements, the VAC4EU validation task force released the final eDCF for use in the PASS case validation.

Support and Quality Control During Study Case Validation

During the study-specific case validation, the VAC4EU validation task force provides ongoing support to abstractors, addressing their questions and uncertainties if they arise. To maintain consistency and ensure alignment across DEAPs, the task force has periodic meetings to discuss common challenges and review emerging issues.

Analysis

Based on the analysis of the abstraction of the dummy cases by new abstractors we assessed the agreement between abstractors (inter-rater reliability) across all outcomes using Fleiss' Kappa,²⁰ suitable in settings with more than two abstractors per outcome. For each outcome, a total of 4 to 22 abstractors were involved across all participating DEAPs combined.

The observed agreement (P_o) was calculated for each case as the proportion of abstractors who selected the correct LOC classification. The overall observed agreement for each dataset was then computed as the average of the observed agreements across all cases.

$$P_o = \frac{\text{Total correct validated cases}}{\text{Total cases}}$$

The expected agreement (P_e) was calculated across all abstractors and cases as follows:

$$P_e = P(\text{Agreed cases})^2 + P(\text{Not agreed cases})^2$$

The Fleiss' Kappa statistic was then computed using the following formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

A higher Fleiss' Kappa indicated stronger agreement between abstractors than expected by chance. Kappa values were categorized as ≤ 0 (no agreement), 0.01–0.20 (slight agreement), 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost high agreement.²¹

To estimate the precision of the Fleiss' Kappa statistic, the standard error (SE) was calculated using the formula:

$$SE = \sqrt{\frac{P_o(1 - P_o)}{N}}$$

Where P_o is the observed agreement and N is the number of cases in the dataset. The 95% confidence interval (CI) for Fleiss' Kappa was then calculated by multiplying the standard error by the Z-score (Z) corresponding to a 95% confidence level (1.96) and adding this value from the Kappa estimate:

$$CI = \kappa \pm (Z \times SE)$$

All calculations were performed using R (version 4.3.0) Fleiss' κ was calculated with the "irr" package (version 0.84.1) in R, using an unweighted approach (treating LOC categories as nominal).^{21,22}

Results

Identification of Case Definitions

For GBS, narcolepsy, ITP, TTS, myocarditis, pericarditis, transverse myelitis, anaphylaxis, major congenital anomalies, encephalopathy (including ADEM), DVT, PE, and CVST, a BC definition was available. For the remaining 3 outcomes (hemorrhagic stroke, non-hemorrhagic stroke, and thrombocytopenia with bleeding), data collection forms were developed based on literature review.

Review of eDCF: [Table 1](#) summarizes how BC definitions were adapted for use in the VAC4EU eDCFs. For each outcome, the total number of criteria implemented reflects both original (unmodified) BC criteria and those either newly introduced or adapted to suit RWD settings. The "Adapted BC Criteria" column indicates how many original BC items required changes for clarity, relevance, or feasibility in RWD abstraction, while the "New Criteria" column captures additional items created by the task force. One such example is the criterion "Outcome reported by a specialist, but without accompanying clinical details", which the task force added to some case definitions (eg, Myocarditis/pericarditis, GBS and TTS) to capture situations where a diagnosis is documented by a specialist in the record but without the expected clinical documentation to fully assess it. Most importantly, the task force introduced a distinction in level 4 between diagnoses made by a specialist without further details and cases where there was insufficient information to meet

Table 1 Adaptations to Case Definitions, Introduction of New Criteria by the VAC4EU Validation Task Force

Outcome/Case Definition	Brighton Collaboration Used	Number of New Criteria Introduced by VAC4EU task force	Number of Adapted BC Criteria	Total Number of Criteria Used in VAC4EU eDCF
Myocarditis and Pericarditis ^a	Yes ²³	1 ("Reported by specialist without clinical details")	12	14
GBS	Yes ²⁴	1 ("Reported by specialist without clinical details")	1	10
Encephalopathy including ADEM	Yes ²⁵	0	1	7
TTS	Yes ²⁶	2 ("Reported by specialist without clinical details" and "Alternative diagnosis")	6	8
Transverse Myelitis	Yes ²⁷	1 ("Reported by specialist without clinical details")	3	5
ITP	Yes ²⁸	1 ("Reported by specialist without clinical details")	1	5
Anaphylaxis	Yes ²⁹	1 ("Reported by specialist without clinical details")	1	7
Narcolepsy	Yes ³⁰	0	1	4
PE	Yes ³¹	2 ("Reported by specialist without clinical details" and "Alternative diagnosis")	2	6

(Continued)

Table 1 (Continued).

Outcome/Case Definition	Brighton Collaboration Used	Number of New Criteria Introduced by VAC4EU task force	Number of Adapted BC Criteria	Total Number of Criteria Used in VAC4EU eDCF
DVT	Yes ³¹	2 (“Reported by specialist without clinical details” and “Alternative diagnosis”)	1	5
CVST	Yes ³¹	1 (“Alternative diagnosis”)	0	3
Major congenital anomalies	Yes ³²	2 (“remove conditions about confirming by medical record review and diagnostic codes”)	0	5
Hemorrhagic stroke	No (Created Internally)	–	–	6
Non hemorrhagic stroke	No (Created Internally)	–	–	5
Thrombocytopenia with bleeding	No (Created Internally)	–	–	7

Notes: *Myocarditis and pericarditis were combined into a single eDCF due to overlapping clinical features and a shared Brighton case definition. “Alternative diagnosis” refers to documentation of a plausible alternative cause for the patient’s clinical presentation, used in the case validation process to rule out the primary outcome when appropriate.

Abbreviations: GBS, Guillain-Barré syndrome; ADEM, acute Disseminated Encephalomyelitis; TTS, Thrombosis with Thrombocytopenia syndrome; PE, Pulmonary Embolism; DVT, Deep Vein Thrombosis; CVST, Cerebral Venous Sinus Thrombosis; ITP, Idiopathic Thrombocytopenic Purpura.

the case definition. This was achieved by adding a subgroup to level 4, labeling it as 4a, and classifying cases with insufficient information to meet the case definition as 4b.

Dummy Cases

For the 16 different outcomes, 78 dummy case narratives were created. Case narratives and their LOC are available in the [supplemental materials](#) and are also publicly available on Zenodo.³³

Training of Abstractors

During the study period 33 abstractors were trained across eight DEAPs. Most abstractors were trained only on a subset of outcomes depending on study contracts and requirements. Consequently, the number of abstractors that were trained per outcome ranged between 4 to 22 (Table 2). Altogether, 747 individual case abstractions were entered in REDCAP by the abstractors on the dummy cases in the training set. Discrepancies between the abstractors’ results and the expected LOC defined by the VAC4EU task force were observed for 93 dummy cases (12.40%). The overall inter-rater agreement was moderate ($\kappa = 0.55$). This κ represents concordance across the entire pooled sample of dummy cases and abstractors. Table 2 presents the per-outcome Fleiss’ κ values with their 95% CIs. The highest agreement among abstractors (Fleiss’ Kappa = 1.00) was observed for transverse myelitis, thrombocytopenia with bleeding, pulmonary embolism, and non-hemorrhagic stroke. In contrast, the highest number of discrepancies was found for the TTS dummy cases, with 53.75% discrepancies and a Kappa of -0.05 (95% CI: -0.48 to 0.38), indicating no agreement. Also the GBS dummy case abstraction showed 21.50% discrepancies, with a Kappa of 0.36 (95% CI: 0.00 to 0.72), suggesting only moderate agreement amongst abstractors. The remaining cases, such as narcolepsy, myocarditis, pericarditis, ITP, encephalitis including ADEM, hemorrhagic stroke, CVST, anaphylaxis, and major congenital anomalies, showed moderate to fair agreement.

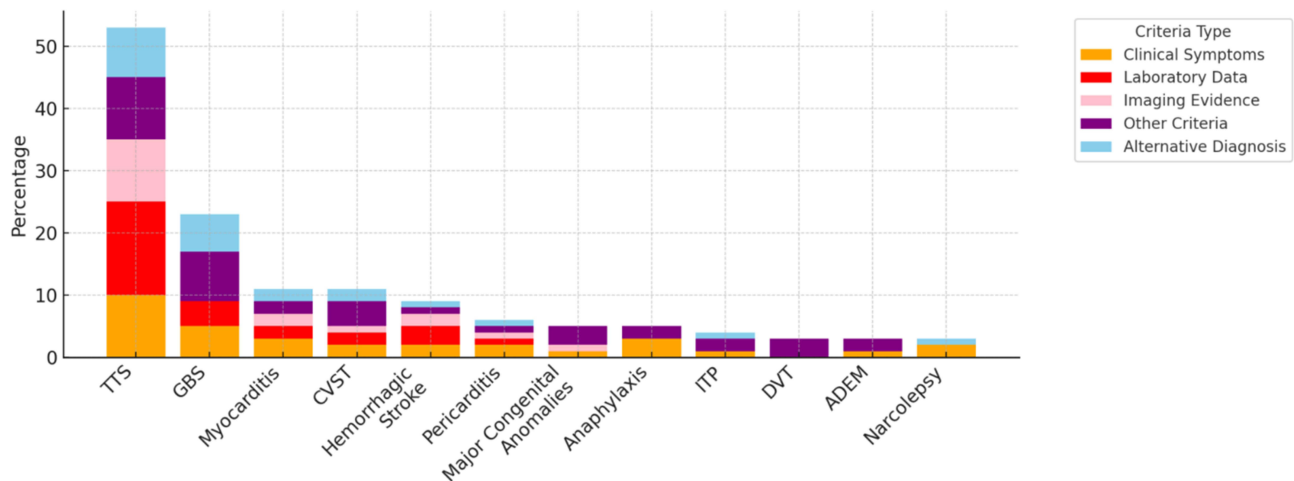
Figure 2 illustrates the distribution of type of discrepancies across five major criteria categories: clinical symptoms, laboratory data, imaging evidence, other criteria, and alternative diagnosis. The most substantial number of discrepancies

Table 2 The Distribution of Abstractors Across Different Outcomes with the Distribution of Discrepancies Across Abstracted Dummy Cases in the Training

Outcomes	Abstractors	Abstracted dummy cases	Discrepancies: Abstractor vs Expected (%)	Fleiss' Kappa (95% CI)
GBS	13	65	21.50%	0.36 (0.00,0.72)
Narcolepsy	13	65	4.61%	0.47 (0.37, 0.58)
ITP	15	75	5.33%	0.47 (0.27, 0.67)
TTS	16	80	53.75%	-0.05 (-0.48, 0.38)
Myocarditis	22	66	11.00%	0.44 (0.31, 0.56)
Pericarditis	22	66	6.00%	0.46 (0.25, 0.67)
Transverse myelitis	13	65	0.00%	1.00
Encephalitis - including ADEM	5	25	8.00%	0.45 (0.21, 0.69)
Thrombocytopenia with bleeding	4	20	0.00%	1,00
DVT	5	25	8.00%	0.45 (0.21, 0.69)
PE	4	20	0.00%	1.00
Hemorrhagic stroke	4	20	10.00%	0.44 (0.18, 0.70)
Non hemorrhagic stroke	5	25	0.00%	1.00
CVST	4	20	15.00%	0.37 (0.02, 0.72)
Anaphylaxis	10	50	8.00%	0.45 (0.21, 0.69)
Major congenital anomalies	12	60	8.30%	0.44 (0.19, 0.69)

Abbreviations: GBS, Guillain-Barré syndrome; ITP, idiopathic thrombocytopenic purpura; TTS, thrombosis with thrombocytopenia syndrome; ADEM, acute disseminated encephalomyelitis; DVT, deep vein thrombosis; PE, pulmonary embolism; CVST, cerebral venous sinus thrombosis.

was observed in the laboratory data category, predominantly driven by TTS cases, where laboratory-related errors accounted for 17.60% of total discrepancies. These discrepancies were often due to inconsistent interpretation of laboratory values, for example, variation in D-dimer reporting units such as fibrinogen equivalent units (FEU) or milligrams per milliliter.

**Figure 2** Distribution of validation discrepancies across criteria for different outcomes.

Abbreviations: TTS, thrombosis with thrombocytopenia syndrome; GBS, Guillain-Barré syndrome; CVST, cerebral venous sinus thrombosis; ITP, idiopathic thrombocytopenic purpura; DVT, deep vein thrombosis; ADEM, acute disseminated encephalomyelitis.

Errors in the alternative diagnosis and other criteria categories were more prominent in myocarditis, pericarditis, and GBS cases. Notably, myocarditis and pericarditis each had over five discrepancies attributed to alternative diagnoses. The GBS cases also had a high contribution from alternative diagnosis errors, totaling 12.00% of discrepancies for that event. Discrepancies related to clinical symptoms were generally lower, but showed some variation, especially in TTS, Narcolepsy, and Anaphylaxis. For instance, Anaphylaxis showed 8.00% discrepancies due to errors in identifying clinical features. Additionally, Major Congenital Anomalies had 8.30% discrepancies, primarily attributed to misclassification under the “other criteria” category, such as incorrect assignment of the event date or misclassification of the anomaly type, including internal, external, or functional anomalies.

Increasing Harmonization Based on Training Results

For cases with low or no concordance, such as in TTS, additional training meetings were scheduled to address the issues identified during the dummy case abstraction. These sessions included a review of misclassified cases, clarification of complex criteria, for instance, interpretation of D-dimer values or alternative diagnoses, and discussion of recurring misunderstandings. Minor discrepancies, such as overlooking one of several required clinical symptoms or entering an incorrect event date due to a typographical error, were communicated via Email with specific guidance to reduce errors. In total, three abstractors were not certified due to repeated errors and insufficient medical knowledge.

Real Case Validation Outcomes

Following dummy-based training and reliability assessment, abstractors were certified for validation of real identified cases across five PASS. At this moment a total of 153 cases of GBS, 83 narcolepsy, 297 ITP, 243 TTS, 1570 myocarditis, 2250 pericarditis, 41 transverse myelitis, 16 cases of acute disseminated encephalomyelitis including ADEM, 50 DVT and PE, 39 hemorrhagic stroke, 50 non-hemorrhagic stroke, 12 CVST, 115 thrombocytopenia with bleeding, 66 anaphylaxis, and 237 cases of major congenital anomalies have been entered in the REDCAP systems.

Discussion

This study described how a validation process of outcomes was implemented in the VAC4EU network, with the aim to reduce heterogeneity across sites and studies. A key finding is that harmonization through guidance documents, and substantive training is needed because initial inter-rater agreement (overall $\kappa \sim 0.55$) based on a standard set of dummy cases was moderate. It is important to note that this inter-rater agreement reflects reliability, the consistency among abstractors given the same information, and does not by itself indicate diagnostic validity. While BC case definitions and companion guides offer a required tool for harmonized vaccine safety assessment this alone does not eliminate heterogeneity when applied in RWD. BC definitions are very clinical and rely on detailed physical examination findings, specific laboratory parameters, and imaging results, which are often unavailable, inconsistently recorded, or entirely absent in some RWD sources. This mismatch affects abstractors' ability to apply case definitions uniformly, especially when evaluating critical diagnostic elements such as laboratory values (eg, variation in D-dimer units) or imaging findings. Our results also provide insight into which outcomes were more challenging for abstractors and why. TTS had the highest level of disagreement, with a low Kappa value, indicating no agreement between abstractors. Overall, the distribution of discrepancies suggests that alternative diagnosis and other criteria introduce the greatest challenges for abstractors.

Examination of the errors showed that application of the TTS definition requires complex judgments about laboratory data, imaging evidence, and especially about ruling out alternative causes. Abstractors varied in how they interpreted these criteria in the dummy cases. To ensure more concordance in validation of the cases in the PASS we organized additional focused training for TTS, clarifying definitions of key laboratory thresholds and what constituted an alternate explanatory diagnosis. Also, GBS had only moderate agreement between abstractors, partly because abstractors differed on interpreting certain neurological exam details in the narratives. To reduce that heterogeneity, we provided supplemental guidance on those points that could be used in the PASS. In contrast, other outcomes (transverse myelitis, thrombocytopenia with bleeding, PE, non-hemorrhagic stroke) achieved perfect agreement ($\kappa = 1.0$) in dummy cases.

The training on dummy cases and review of errors was very instructive and showed that certain types of errors were common. In particular, interpretation of laboratory results (eg, how to handle a slightly ambiguous D-dimer value for TTS) or determining whether a potential alternative diagnosis explained the case's symptoms (especially in outcomes like TTS and GBS), created differences. These patterns highlight where additional guidance was needed. The moderate to poor agreement observed for some outcomes reflect the challenges in achieving harmonized results in situations where clinical judgment and alternative diagnoses play a significant role. These insights provide important recommendations also for other large distributed research networks that implement case validation; for example, the Vaccine Safety Datalink (VSD) also highlights the need for thorough clinical review to ensure the accuracy of diagnoses, especially in cases where automated data may not fully capture complex clinical histories and uses BC definitions.³⁴ The Global Vaccine Data Network also validates cases, using BC definitions, and deployed training on dummy cases, based on VAC4EU recommendations.³⁵ Based on this experience VAC4EU provides structured feedback to the Brighton Collaboration as part of a living lab.

Strengths and Limitations

A major strength of this study is the application of a harmonized and scalable validation framework across a large, multinational network using RWD. The use of standardized training materials, structured data collection tools, and systematic adjudication processes will increase consistency and comparability across diverse settings.

However, several limitations should be noted. The dummy cases used for training were narrative-based and may not fully reflect the structure of actual data extracts encountered during study implementation, this means that the heterogeneity between abstractors may actually be greater. Furthermore, differences in data availability, language, and healthcare practices across DEAPs may require further local feedback.

We designed the validation process to be broadly applicable across different data sources and settings. Training materials and forms were developed in English to support broad accessibility across language settings, but minor interpretation differences may have arisen for non-native English abstractors. Differences in data availability may further create heterogeneity. For instance, not all DEAPs collect the same laboratory details (some may not record a troponin level needed for a myocarditis definition; others might not have imaging reports readily accessible). This variation highlights the importance of clear data abstraction tools for RWD.

The validation process required considerable effort into creation REDCap forms and verification of the logic. Making such forms and logic available in a digitized format, whether through the Brighton Collaboration or other networks, could reduce redundancy across networks, and improve consistency in applying the definitions. Moreover, incorporating insights from field implementation, including training feedback and validation challenges, would enhance the practical applicability of the BC tools in RWD settings.

The reported process and training highlight several opportunities for further improvement. First, it will be important to capture the learning curve among abstractors, for example, did agreement rates improve after the initial round of dummy case training, or with additional practice and feedback? Quantifying any improvement over successive training iterations could provide insight into how much training is "enough" and which outcomes needed extra reinforcement (we suspect outcomes like TTS would show significant improvement from round 1 to round 2 of training). Second, we can use alternative agreement metrics such as Gwet's AC1 in addition to Fleiss' κ . Gwet's AC1 which is more stable in situations with very high or very low prevalence of a category. Third, this study does not capture information on the heterogeneity of actual validation in PASS, further validation of this within the boundaries of the GDPR should be explored.

Conclusion

In conclusion, the VAC4EU validation pipeline offers a scalable and adaptable implementation model for harmonized case validation in distributed RWD studies. VAC4EU has set a standard for implementation of validation according to Brighton Collaboration criteria in retrospective use of RWD.

By adapting BC case definitions to the realities and constraints of RWD, and by implementing comprehensive abstractor training with ongoing feedback, this approach has the potential to improve the consistency and reliability of

outcome adjudication across multiple sites and countries. Overall, this validation process supports more harmonized robust post-marketing vaccine safety monitoring and may serve as a replicable model for future multinational collaborative studies leveraging heterogeneous RWD sources.

Acknowledgments

The authors would like to thank all members of the VAC4EU network who contributed to the development, testing, and implementation of the validation pipeline. We are especially grateful to the members of the validation task force for their input on adapting case definitions, creating training materials, and coordinating abstraction activities across participating data sources.

We would also like to thank the following abstractors for their invaluable contributions to the manual validation process across the various outcomes and data sources included in this study:

Rachel Aakerøy, Ashi Sarfraz Ahmad, Olaug Marie Bruheim Reiakvam, Juanjo Carreras, Costanza Di Chiara, María Díaz López, Edison Daniel Valenzuela Cumba, Sandeep Dhanda, Miranda Davies, Pablo Daniel Estrella Porter, Sílvia Fernández García, María Luisa Gil Canela, María Gine, Nils Erik Gilhus, Anyuli Gracia Gutiérrez, Eva Jara Castillejo, Christopher Steph Inchley, Laurits Juhl Heinsen, Parinaz Heydari, Thom Lysen, Aida Moreno Juste, Kathryn Morton, Marta Pastor Sanz, Sergio Pascual Vicedo Mata, Roberto Peribañez García, Lina Pérez Breva, Oscar Oelrich Rosenkrantz, Martín Solórzano, Esther Soriano García, Victor Hejgaard Sørensen, Christian Thaulow, Jesse Van den Berg, and Jannik Wheler.

Their careful review and commitment to consistency played a key role in achieving the quality and reproducibility of the validation process.

This study is based in part on data from the Clinical Practice Research Datalink (CPRD) (study protocol numbers: 21_000714, 21_000535, 23_003471) obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author(s) alone.

This paper has been uploaded to medRxiv as a preprint: <https://www.medrxiv.org/content/10.1101/2025.08.25.25334384v1/https://scity.org/articles/activity/10.1101/2025.08.25.25334384>

Author Contributions

MS conceptualized the study and supervised all stages of the validation pipeline. AD developed the manuscript, conducted the case validation pipeline, and coordinated co-author feedback. All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Disclosure

TL and JB are employees of the PHARMO Institute for Drug Outcomes Research, an independent research organization that conducts funded studies for governmental authorities and pharmaceutical companies. JJC and LP-B have no personal conflicts of interest. Their institution, FISABIO, receives funding from pharmaceutical companies solely for conducting scientific research; this institutional support had no influence on the content of this work. LZ was previously employed by Julius Clinical Ltd., an academic contract research organization (CRO) that receives funding from Moderna Ltd. for research on COVID-19 vaccine safety. JW and VE are salaried employees of institutions that have received financial support from Moderna Ltd. CR is an employee of RTI-HS. RTI-HS has participated in several vaccine safety studies in Europe and US. FA reports personal fees from Johnson and Johnson, during the conduct of the study. AA is an employee at RTI-HS, an independent institute member of the VAC4EU consortium. MS reports grants from AstraZeneca, Janssen, Pfizer, and European Medicines Agency, outside the submitted work. All other authors declare no conflicts of interest in relation to this publication.

References

- Shattock AJ, Johnson HC, Sim SY, et al. Contribution of vaccination to improved survival and health: modelling 50 years of the expanded programme on immunization. *Lancet*. 2024;403(10441):2307–2316. doi:10.1016/S0140-6736(24)00850-X
- Centers for Disease Control and Prevention (CDC). CDC's vaccine safety monitoring program. Available from: <https://www.cdc.gov/vaccine-safety-systems/about/cdc-monitoring-program.html>. Accessed August 7, 2025.
- Dodd C, Andrews N, Petousis-Harris H, Sturkenboom M, Omer SB, Black S. Methodological frontiers in vaccine safety: qualifying available evidence for rare events, use of distributed data networks to monitor vaccine safety issues, and monitoring the safety of pregnancy interventions. *BMJ Glob Health*. 2021;6(Suppl 2):e003540. PMID: 34011501; PMCID: PMC8137251. doi:10.1136/bmjgh-2020-003540
- U.S. Food and Drug Administration (FDA). Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. guidance for industry. Available from: <https://www.fda.gov/media/152503/download>. Accessed August 7, 2025.
- European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP). Guide on methodological standards in pharmacoepidemiology. Available from: http://www.encepp.eu/standards_and_guidances. Accessed August 7, 2025.
- Jurek AM, Greenland S, Maldonado G, Church TR. Proper interpretation of non-differential misclassification effects: expectations vs observations. *Int J Epidemiol*. 2005;34(3):680–687. doi:10.1093/ije/dyi060
- Ehrenstein V, Hellfritsch M, Kahlert J, et al. Validation of algorithms in studies based on routinely collected health data: general principles. *Am J Epidemiol*. 2024;193(11):1612–1624. doi:10.1093/aje/kwae071
- Weinstein EJ, Ritchey ME, Lo Re V. Core concepts in pharmacoepidemiology: validation of health outcomes of interest within real-world healthcare databases. *Pharmacoepidemiol Drug Saf*. 2023;32(1):1–8. doi:10.1002/pds.5537
- European Medicines Agency (EMA). Guideline on Good Pharmacovigilance Practices (GVP) module VIII – post-authorisation safety studies. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-good-pharmacovigilance-practices-gvp-module-viii-post-authorisation-safety-studies-rev-3_en.pdf. Accessed August 7, 2025.
- Brighton Collaboration. Case definitions. Available from: <https://brightoncollaboration.org/case-definitions-table-view/>. Accessed December 28, 2025.
- Kohl KS, Bonhoeffer J, Chen R, et al. The Brighton Collaboration: enhancing comparability of vaccine safety data. *Pharmacoepidemiol Drug Saf*. 2003;12(4):335–340. PMID: 12812014. doi:10.1002/pds.851
- Weibel D, Sturkenboom M, Black S, et al. Narcolepsy and adjuvanted pandemic influenza A (H1N1) 2009 vaccines - multi-country assessment. *Vaccine*. 2018;36(41):6202–6211. PMID: 30122647; PMCID: PMC6404226. doi:10.1016/j.vaccine.2018.08.008
- Gadroen K, Straus SMJM, Pacurariu A, Weibel D, Kurz X, Sturkenboom MCJM. Patterns of spontaneous reports on narcolepsy following administration of pandemic influenza vaccine; a case series of individual case safety reports in Eudravigilance. *Vaccine*. 2016;34(41):4892–4897. PMID: 27577558. doi:10.1016/j.vaccine.2016.08.062
- Romio S, Weibel D, Dieleman JP, et al. Guillain-Barré syndrome and adjuvanted pandemic influenza A (H1N1) 2009 vaccines: a multinational self-controlled case series in Europe. *PLoS One*. 2014;9(1):e82222. PMID: 24404128; PMCID: PMC3880265. doi:10.1371/journal.pone.0082222
- Ording AG, Cronin-Fenton D, Ehrenstein V, et al. Challenges in translating endpoints from trials to observational cohort studies in oncology. *Clin Epidemiol*. 2016;8:195–200. doi:10.2147/CLEP.S97874
- Thurin NH, Pajouheshnia R, Roberto G, et al. From inception to ConcePTION: genesis of a network to support better monitoring and communication of medication safety during pregnancy and breastfeeding. *Clin Pharmacol Ther*. 2022;111(1):321–331. doi:10.1002/cpt.2476
- European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Off J Eur Union*. 2016; L119:1–88.
- Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377–381. doi:10.1016/j.jbi.2008.08.010
- Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208. doi:10.1016/j.jbi.2019.103208
- Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: measures of agreement. *Perspect Clin Res*. 2017;8(4):187–191. doi:10.4103/picr.PICR_123_17
- R Core Team. R: a Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: <https://www.R-project.org/>. Accessed August 7, 2025.
- Gamer M, Lemon J, Fellows I, Singh P. irr: various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. 2019. Available from: <https://CRAN.R-project.org/package=irr>. Accessed August 7, 2025.
- Law B. AESI case definition companion guide: myocarditis and pericarditis. Zenodo. 2022. doi: 10.5281/zenodo.6668895.
- Law B. AESI case definition companion guide: guillain barré and Miller Fisher syndrome. 2021. doi: 10.5281/zenodo.6668879.
- Law B. AESI case definition companion guide: acute disseminated encephalomyelitis. Zenodo. 2021. doi: 10.5281/zenodo.6668857.
- Chen R, Buttery J. DRAFT - TTS: case definition & guidelines for data collection, analysis, and presentation of immunization safety data. Zenodo. 2021. doi: 10.5281/zenodo.6697333.
- Law B. AESI case definition companion guide: myelitis. Zenodo. 2021. doi: 10.5281/zenodo.6668655.
- Law B. AESI case definition companion guide: thrombocytopenia. Zenodo. 2021. doi: 10.5281/zenodo.6668865.
- Law B. AESI case definition companion guide: anaphylaxis version 2. Zenodo. 2022. doi: 10.5281/zenodo.7248919.
- Poli F, Overeem S, Lammers GJ, et al. Narcolepsy as an adverse event following immunization: case definition and guidelines for data collection, analysis and presentation. *Vaccine*. 2013;31(6):994–1007. PMID: 23246545. doi:10.1016/j.vaccine.2012.12.014
- Gollamudi J, Sartain SE, Navaei AH, et al. Thrombosis and thromboembolism: brighton collaboration case definition and guidelines for data collection, analysis, and presentation of immunization safety data. *Vaccine*. 2022;40(44):6431–6444. PMID: 36150973. doi:10.1016/j.vaccine.2022.09.001
- DeSilva M, Munoz FM, McMillan M, et al. Congenital anomalies: case definition and guidelines for data collection, analysis, and presentation of immunization safety data. *Vaccine*. 2016;34(49):6015–6026. doi:10.1016/j.vaccine.2016.03.047

33. The Vaccine Monitoring Collaboration for Europe (VAC4EU). VAC4EU Case definitions and data collection forms. Zenodo. 2026. doi:10.5281/zenodo.18141231.
34. Meil MM, Gee J, Weintraub ES, et al. The vaccine safety datalink: successes and challenges monitoring vaccine safety. *Vaccine*. 2014;32(42):5390–5398. doi:10.1016/j.vaccine.2014.07.073
35. Mònica S, Judit R-A, Elena B, et al. Validation of Guillain-Barré syndrome case identification in four heterogeneous vac4eu real-world data sources in spain and the united kingdom using the brighton collaboration criteria. doi: 10.2139/ssrn.5376906.

Clinical Epidemiology

Publish your work in this journal

Clinical Epidemiology is an international, peer-reviewed, open access, online journal focusing on disease and drug epidemiology, identification of risk factors and screening procedures to develop optimal preventative initiatives and programs. Specific topics include: diagnosis, prognosis, treatment, screening, prevention, risk factor modification, systematic reviews, risk & safety of medical interventions, epidemiology & biostatistical methods, and evaluation of guidelines, translational medicine, health policies & economic evaluations. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use.

Submit your manuscript here: <https://www.dovepress.com/clinical-epidemiology-journal>

Dovepress
Taylor & Francis Group