ORIGINAL RESEARCH ARTICLE

Phenotype Algorithms for the Identification and Characterization of Vaccine-Induced Thrombotic Thrombocytopenia in Real World Data: A Multinational Network Cohort Study

Azza Shoaibi[1,2] · Gowtham A. Rao[1,2] · Erica A. Voss[1,2] · Anna Ostropolets[2,3] · Miguel Angel Mayer[4] ·
Juan Manuel Ramírez-Anguita[4] · Filip Maljković[5] · Biljana Carević[6] · Scott Horban[7] · Daniel R. Morales[7] ·
Talita Duarte-Salles[8] · Clement Fraboulet[9] · Tanguy Le Carrour[10] · Spiros Denaxas[11] · Vaclav Papez[11] · Luis H. John[12] ·
Peter R. Rijneek[12] · Evan Minty[13] · Thamir M. Alshammari[2,14] · Rupa Makadia[1,2] · Clair Blacketer[1,2] · Frank DeFalco[1,2] ·
Anthony G. Sena[1,2] · Marc A. Suchard[2,15] · Daniel Prieto-Alhambra[16] · Patrick B. Ryan[1,2]

Abstract

Introduction Vaccine-induced thrombotic thrombocytopenia (VITT) has been identified as a rare but serious adverse event associated with coronavirus disease 2019 (COVID-19) vaccines.

Objectives In this study, we explored the pre-pandemic co-occurrence of thrombosis with thrombocytopenia (TWT) using 17 observational health data sources across the world. We applied multiple TWT definitions, estimated the background rate of TWT, characterized TWT patients, and explored the makeup of thrombosis types among TWT patients.

Methods We conducted an international network retrospective cohort study using electronic health records and insurance claims data, estimating background rates of TWT amongst persons observed from 2017 to 2019. Following the principles of existing VITT clinical definitions, TWT was defined as patients with a diagnosis of embolic or thrombotic arterial or venous events and a diagnosis or measurement of thrombocytopenia within 7 days. Six TWT phenotypes were considered, which varied in the approach taken in defining thrombosis and thrombocytopenia in real world data.

Results Overall TWT incidence rates ranged from 1.62 to 150.65 per 100,000 person-years. Substantial heterogeneity exists across data sources and by age, sex, and alternative TWT phenotypes. TWT patients were likely to be men of older age with various comorbidities. Among the thrombosis types, arterial thrombotic events were the most common.

Conclusion Our findings suggest that identifying VITT in observational data presents a substantial challenge, as implementing VITT case definitions based on the co-occurrence of TWT results in large and heterogeneous incidence rate and in a cohort of patints with baseline characteristics that are inconsistent with the VITT cases reported to date.

✉ Azza Shoaibi
ashoaibi@its.jnj.com

Extended author information available on the last page of the article

△ Adis

Key Points

Our findings suggest that identifying vaccine-induced thrombotic thrombocytopenia (VITT) in observational data presents a substantial challenge.

Implementing VITT case definitions based on the co-occurrence of thrombosis with thrombocytopenia results in large and heterogeneous incidence rates in a cohort composed of patients with baseline characteristics that are different to the VITT cases reported after the coronavirus disease 2019 (COVID-19) vaccines.

We advise that further refinement of the case definition is needed before observational data can be reliably used to generate unbiased population-level effect estimates for VITT safety surveillance.

1 Introduction

Vaccine-induced thrombotic thrombocytopenia (VITT) has been reported after vaccination with adenovirus-based coronavirus disease 2019 (COVID-19) vaccines [1, 2]. VITT is characterized by exposure to one of the aforementioned vaccines 4–30 days prior to presentation, followed by thrombosis, mild-to-severe thrombocytopenia, and a positive platelet factor-4 (PF4)-heparin enzyme-linked immunosorbent assay (ELISA) [3]. Thrombotic events can include either venous or arterial thrombosis, and often involves atypical locations, including cerebral venous thrombosis and splanchnic vein thrombosis [3, 4].

Multiple conditions are associated with both thrombosis and thrombocytopenia, including heparin-induced thrombocytopenia (HIT), antiphospholipid syndrome, thrombotic thrombocytopenic purpura (TTP), disseminated intravascular coagulation (DIC), and some malignant tumors. However, the term VITT today (also termed thrombosis with thrombocytopenia syndrome [TTS]) [5] refers specifically to the syndrome associated with COVID-19 vaccines and is considered a new clinical syndrome. As such, there is currently no standard case definition for VITT accepted for use by all countries. On April 3, 2021, the British Society of Haematology published its Updated Guidance on Management (Version 1.0) with a case definition for possible, probable, and definite cases for VITT [6]. Additionally, the Brighton Collaboration drafted and published an interim case definition for possible, probable, and definite (level one) VITT cases oriented towards identification and treatment of cases [7]. Both definitions require a platelet

count of less than 150,000 per microliter to identify thrombocytopenia, and definite case (level one) criteria require confirmed thrombosis through laboratory, imaging, surgical, or pathology findings. In addition, the British Society of Haematology definition requires that antibodies to PF4 have been identified in an absence of heparin exposure. The Brighton definition further classifies the case into 1–3 "H" level depending on history of heparin exposure within 100 days. While the above case definitions provided guidance for health providers to identify and treat VITT patients, there is yet no consensus (or a clear guidance) on how to identify VITT cases in observational health data including claims and electronic health records (EHR).

There is a consensus that VITT is a new clinical phenomenon; however, estimates of the historical background rate of the co-occurrence of thrombosis with thrombocytopenia (TWT) are still needed to contextualize VITT safety signals. Specifically, estimating the number of patients that may have a co-occurrence of TWT that would typically be observed in the absence of vaccinations is required to understand the risk. More importantly, properly identifying cases from historical data provides an idea about the profile of patients who had TWT in the past. This can help inform a consideration of whether the profiles of individuals with VITT after a vaccination against COVID-19 differ from those of individuals who have historically had similar events.

Retrospective observational data can also be used to estimate the historical frequency at which patients with a thrombotic event have platelet counts measured (before the emergence of VITT as a phenomenon). Establishing this background frequency can help measure and account for the surveillance bias that is likely to occur when estimating the relative risk of VITT. Exploring TWT definitions in real world data can provide insight into whether such definitions can be used as a proxy for VITT case identification, for conducting observational safety outcome research, for case finding in the context of safety surveillance activities and epidemiological studies, and for accurate historical background rate estimation.

In this study, we implement the VITT Brighton Collaboration case definitions as standardized cohorts that can be applied across disparate observational data sources to empirically examine alternative TWT definitions (phenotypes). As illustrated in Fig. 1, we address the following questions: (1) What is the estimated background rate of TWT and how does it vary across alternative TWT definitions? (2) What are the baseline characteristics of TWT identified patients and how comparable are they with known VITT patient profiles? (3) What specific events make up thrombosis among the TWT cohorts (deep venous thrombosis [DVT], myocardial infarction [MI], splenic thrombosis, etc.)? (4) How do we capture thrombocytopenia across various data sources (for example, using
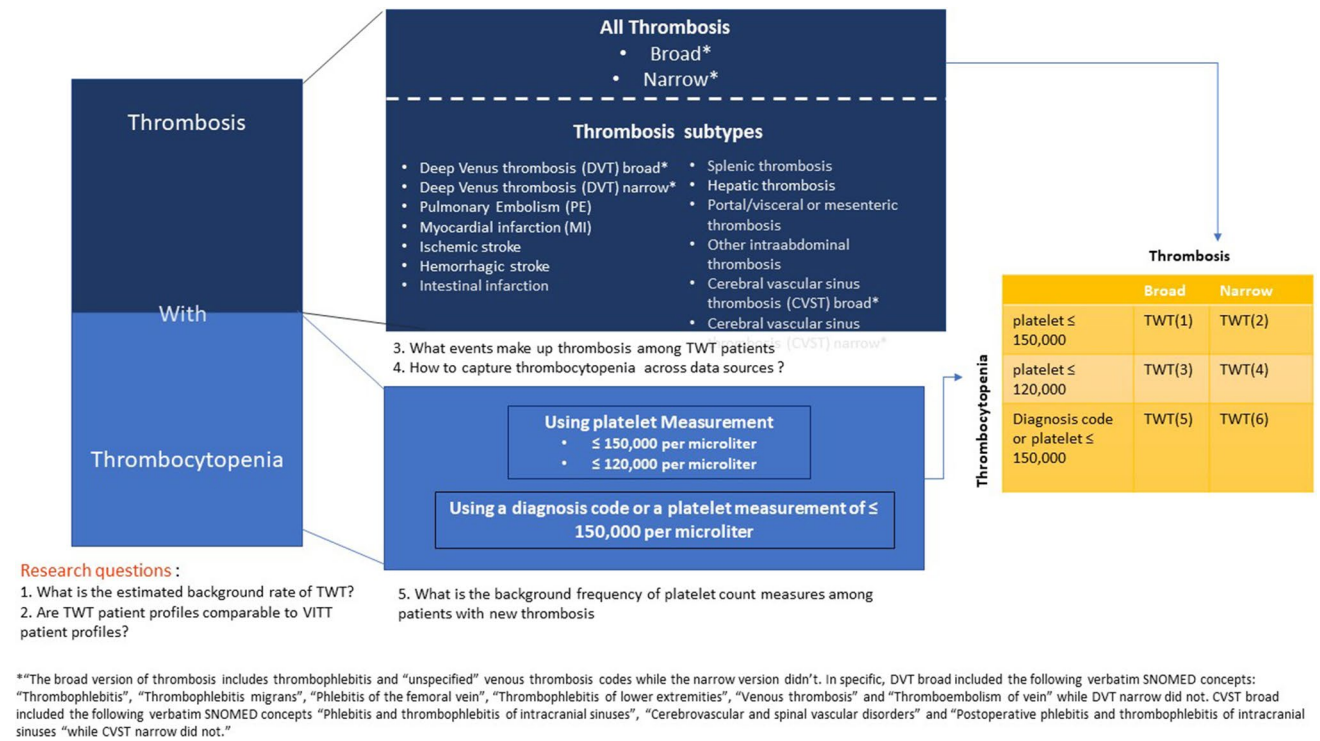
Fig. 1 Research questions and thrombosis with thrombocytopenia (TWT) outcome definitions; schematic of the research questions addressed in the article and the TWT definitions considered in the analysis

diagnosis code or using platelet measure value)? And finally, (5) What is the background frequency of platelet count measures among patients with new thrombosis? All analyses were run across data sources within the Observational Health Data Sciences and Informatics (OHDSI) program [available from: https://www.ohdsi.org/] and European Health Data & Evidence Network (EHDEN) [available from https://www.ehden.eu/], including administrative claims and EHR sources across the United States (US), Europe, and Asia-Pacific regions.

2 Methods

2.1 Study Design

We conducted an international network cohort study using routinely collected primary care and hospital patient records from across the US, Australia, Japan, and Europe. Data were previously mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [8, 9], which allowed for the study to be run in a distributed manner, with common analytic codes run by each site without the need to share patient-level data between sites.

2.2 Data Sources

We included 17 data sources (Table 1) from ten countries, of which four were administrative health claims, one was a biobank registry, and the rest were EHR data sources.

The EHR data sources were:

- IQVIA® Australia Longitudinal Patient Data (IQVIA_ Australia), general practitioner (GP) data from Australia.
- Integrated Primary Care Information (IPCI) [10] GP data from the Netherlands.
- IQVIA® Disease Analyser Germany (IQVIA_Germany), GP and medical center data from Germany.
- Information System for Research in Primary Care (SID-IAP), primary care records linked to hospital admissions of Conjunt Mínim Bàsic de Dades d'Alta Hospitalària from Catalonia, Spain.
- Clinical Practice Research Datalink (CPRD), GP data from the United Kingdom (UK).
- Columbia University Irving Medical Center (CUMC), hospital records from New York-Presbyterian Hospital/ Columbia University Irving Medical Center in the US.
- Optum® de-identified Electronic Health Record Dataset (Optum_EHR), an EHR repository derived from dozens of healthcare provider organizations in the US.

Table 1 Data sources type, country, and data element availability

Data source	Country	Type	Explicit observation period	Inpatient diagnosis	Outpatient diagnosis	Inpatient measurements	Outpatient measurements	Outpatient drug	Inpatient drug
CCAE	USA	Claims	Available	Available	Available	Not Available	Partially available	Available	Not Available
Optum_Extended_DoD	USA	Claims	Available	Available	Available	Not Available	Partially available	Available	Not Available
MDCD	USA	Claims	Available	Available	Available	Not Available	Not Available	Available	Not Available
MDCR	USA	Claims	Available	Available	Available	Not Available	Partially available	Available	Not Available
Optum_EHR	USA	EHR	Not Available	Available	Partially available	Available	Partially available	Partially available	Available
CUMC	USA	Hospital EHR	Not Available	Available	Partially available	Available	Available	Partially available	Available
CPRD	UK	GP records	Available	Not Available	Available	Not Available	Available	Available	Not Available
Biobank_UK	UK	Registry	Available	Partially available	Available	Not Available	Available	Available	Not Available
HIC	Scotland	Hospital EHR	Available	Available	Available	Available	Available	Available	Not Available
IPCI	Netherlands	GP records	Available	Not Available	Available	Not Available	Available	Available	Not Available
APHM	France	Hospital EHR	Available	Available	Not Available	Available	Available	Not Available	Available
SIDIAP	Spain	EHR	Available	Available	Available	Not Available	Available	Available	Not Available
FIMIM-IMASIS	Spain	Hospital EHR	Available	Available	Partially available	Available	Not Available	Not available	Available
IQVIA_Germany	Germany	GP records	Not Available	Not Available	Available	Not Available	Partially available	Available	Not Available
CC_Serbia	Serbia	Hospital EHR	Not Available	Available	Available	Available	Available	Not available	Available
IQVIA_Australia	Australia	GP records	Not Available	Not Available	Available	Not Available	Available	Available	Not Available
JMDC	Japan	Claims	Available	Available	Available	Not Available	Not Available	Available	Available

APHM Health Data Warehouse of Assistance Publique—Hopitaux de Marseille, *Biobank_UK* UK Biobank, *CC_Serbia* University Clinical Center of Serbia, *CCAE* IBM® MarketScan® Commercial Claims and Encounters Database, *CPRD* Clinical Practice Research Datalink, *CUMC* Columbia University Irving Medical Center, *EHR* electronic health records, *FIMIM-IMASIS* Information System of Parc Salut Mar Barcelona, *GP* general practitioner, *HIC* Health Informatics Centre from University of Dundee, *IPCI* Integrated Primary Care Information, *IQVIA_Australia* IQVIA® Australia Longitudinal Patient Data, *IQVIA_Germany* IQVIA® Disease Analyser Germany, *JMDC* Japan Medical Data Center, *MDCD* IBM® MarketScan® Multi-State Medicaid Database, *MDCR* IBM® MarketScan® Medicare Supplemental and Coordination of Benefits Database, *Optum_EHR* Optum® de-identified Electronic Health Record Dataset, *Optum_Extended_DoD* Optum® De-Identified Clinformatics® Extended Data Mart Database—Date of death, *SIDIAP* Information System for Research in Primary Care

- Health Data Warehouse of Assistance Publique—Hopitaux de Marseille (APHM), a hospital-based EHR data from France.
- Information System of Parc Salut Mar Barcelona (FIMIM-IMASIS), a hospital-based EHR from Barcelona, Spain.
- University Clinical Center of Serbia (CC_Serbia), a hospital-based EHR data from Serbia.

- Health Informatics Centre (HIC) from the University of Dundee, an EHR database from Scotland, containing laboratory measurements from both primary and secondary care. The biobank registry data source was UK Biobank [11], a large longitudinal biobank study of 500,000 middle aged adults from the UK (England, Scotland and Wales) and with extensive health outcome linkages established to primary care, hospitalization,

cancer registration and death EHR sources (Biobank_UK).

The claims-based data sources were:

- The Japan Medical Data Center (JMDC) [12] and four US administrative claims data sources
- IBM® MarketScan® Commercial Claims and Encounters Database (CCAE)
- IBM® MarketScan® Medicare Supplemental and Coordination of Benefits Data source (MDCR)
- IBM® MarketScan® Multi-State Medicaid Data source (MDCD)
- Optum® De-Identified Clinformatics® Extended Data Mart Data source—Date of death (Optum_Extended_DoD)

All data sources vary in their time duration, patient coverage, and density of lab measurement records and results. EHR data sources cover lab measurement results that occurred in a particular setting (inpatient or outpatient), while claims data sources have either partial or no lab measurements. A detailed description of the data sources can be found in Appendix Table 1 (see the electronic supplementary material). Table 1 summarizes the availability of data elements by data source.

2.3 Study Participants and Time at Risk for Background Incidence Rate Estimates

The study cohort consisted of individuals present in a data source as of January 1, 2017, 2018, or 2019. These dates were used as the index date for all study participants. Individuals were required to have a minimum of 1 year of history available in the data source prior to their index date. Time at risk was defined as 365 days (from 0 days to 365 days following the index date January 1). Patients contributed time at risk from the index date until the earliest of 365 days after the index, their observation period end date, or the start date of a TWT event. Persons with prior thrombosis or thrombocytopenia events did not begin to contribute time at risk until the 365-day clean window requirement was satisfied for thrombosis and 90 days for thrombocytopenia.

2.4 TWT Outcome Definitions (Phenotypes)

Following the principles of existing VITT clinical definitions, TWT was defined as patients with a diagnosis of embolic or thrombotic arterial or venous events and a diagnosis or measurement of thrombocytopenia within 7 days (thrombocytopenic events can occur 7 days before or after the thrombotic event). Figure 1 illustrates the TWT outcome definitions that were considered.

Based on the approach used to define thrombosis and thrombocytopenia, six alternative TWT definitions were considered. The occurrence of thrombosis was identified using diagnosis codes (Supplemental Table 1, see the electronic supplementary material) with a broad set of codes (that included thrombophlebitis and other generic venous thrombosis codes) representing a more sensitive approach and a narrow set representing a more specific approach. The additional set of concepts that were part of the broad set but not in the narrow set is summarized in Supplemental Table 2 using the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) and the corresponding ICD-10-CM and Read codes. The occurrence of thrombocytopenia was defined as follows: (1) having a measurement of ≤ 150,000 platelets per microliter of blood regardless of the presence/absence of a diagnosis code, (2) having a measurement of ≤ 120,000 platelets per microliter regardless of the presence/absence of a diagnosis code, or (3) having a diagnosis code or a platelet measurement of ≤ 150,000 per microliter. The combination of these three alternatives to define thrombocytopenia and the two alternatives to define thrombosis led to six different TWT phenotypes (yellow matrix in Fig. 1). To implement the requirement of "new" thrombosis and thrombocytopenia in observational data, we required a clean window of 365 days for thrombosis (of any kind) and 90 days for thrombocytopenia in all six TWT definitions. In all TWT definitions, patients enter the study population (i.e., classified as having the TWT event) at the date of the new thrombotic event.

2.5 Specific TWT Subtypes

As represented in the dark blue box in Fig. 1, we explored 13 additional specific TWT phenotypes based on thrombosis type as follows: (1) DVT using a broad version (i.e., sensitive) that includes phlebitis diagnosis codes and other generic venous thrombosis codes; (2) DVT using a narrow version (i.e., specific); (3) pulmonary embolism (PE); (4) MI; (5) ischemic stroke; (6) hemorrhagic stroke; (7) hepatic thrombosis; (8) splenic thrombosis; (9) intestinal infarction; (10) portal/visceral or mesenteric thrombosis; (11) other intra-abdominal thrombosis (vena cava thrombosis, iliac artery, abdominal aorta, and trunk thrombosis and others); (12) cerebral venous sinus thrombosis (CVST) using a broad version that includes phlebitis of intracranial sinuses; and (13) CVST using a narrow version that did not include phlebitis diagnosis codes. As summarized in Supplemental Table 2 (see the electronic supplementary material), the concepts that were contained in the broad DVT definition but were omitted from the narrow were as follows: "thrombophlebitis," "thrombophlebitis migrans," "phlebitis of the femoral vein," "thrombophlebitis of lower extremities," "venous thrombosis" and "thromboembolism

of vein." Similarly, "phlebitis and thrombophlebitis of intracranial sinuses," "cerebrovascular and spinal vascular disorders" and "postoperative phlebitis and thrombophlebitis of intracranial sinuses" were contained in the broad CVST definition but omitted in the narrow.

For simplicity, in all specific TWT subtype cohorts presented in this paper; we identified thrombocytopenia using a diagnosis code or a measurement of ≤ 150,000 per microliter. We also required a clean window of 365 for thrombosis (of any kind) and 90 days for thrombocytopenia in all 13 TWT subtypes.

Supplemental Table 1 provides the list of included diagnosis codes for all types of thrombosis and thrombocytopenia using SNOMED CT vocabulary. All study outcomes were identified using code lists reviewed by a panel of epidemiologists and clinicians (AS, PR, DPA, GR, AO, EM). A previous version of this code list had been used in previous published studies [13] and reviewed by a hematologist and a neurologist. These definitions were reviewed using the aid of the CohortDiagnostics R package [14] so as to identify additional diagnosis codes of interest and to remove those highlighted as irrelevant based on feedback from regulators (e.g., puerperium and pregnancy-related disease) through an iterative process during the initial stages of analyses. A detailed description of the definitions used to identify the outcomes of the study is provided at https://data.ohdsi.org/Covid19VaccineAesiDiagnostics/. This application summarizes the codes used to identify outcomes and their frequency in the data sources used in the study, the overlap between cohorts in the data sources, and a detailed summary of the profiles of all the individuals with a code of interest in each of the data sources.

2.6 Cohort Characterization

We applied the definitions described above across the 17 data sources using all available historical data and summarized the following characterizations:

1. *The index event breakdown*: To explore what event makes up thrombosis among TWT patients, we provide the distribution of the clinical events—thrombosis diagnosis codes—that were observed on index and qualified individuals for cohort entry.
2. *Cohort characterization*: The distribution of baseline characteristics including demographics, clinical conditions, and heparin and other drug use. We identified medical conditions in the data sources present 1 year before or at the index date, using inpatient and outpatient diagnosis codes. We report drug use in the last 30 days before and including index.

All descriptive analysis was done through CohortDiagnostics R package [14].

2.7 Statistical Methods

The profiles of the study cohorts and those with an outcome of interest were summarized, with median and interquartile range (IQR) used for continuous variables and counts and percentages used for categorical variables. Incidence rates (IR) were tabulated as the number of outcomes, divided by the person-time at risk, and summarized per 100 person-years. The rates were also calculated by stratifying by age decades and gender. Incidence proportions are additionally reported and were calculated as the number of persons with the outcome, divided by the number of persons with time-at-risk, summarized per 100 persons. Age- and gender-specific rates are reported

The proportional difference (% change) in incidence of two cohort definitions A and B is computed as (incidence of A—incidence of B) divided by incidence of B. For example, the % change in overall IR of TWT when thrombocytopenia is defined using a measurement of ≤ 120,000 platelets per microliter of blood compared to when thrombocytopenia is defined using measurement of ≤ 150,000 per microliter was calculated as (overall IR using measurement of ≤ 120,000— overall IR using measurement of ≤ 150,000)/overall IR using measurement of ≤ 150,000.

The incidence proportion of platelet measures was calculated as the number of patients with new diagnosis of thrombosis who had a new platelet measurement (that is of any value) within 7 days divided by the number of persons with a new diagnosis of thrombosis (of any kind). We also report the incidence proportion of patients with low platelet by dividing the number of patients with new diagnosis of thrombosis who had a new platelet measurement that is ≤ 150,000 per microliter within 7 days by the number of patients with new diagnosis of thrombosis who had a new platelet measurement (that is of any value). For these calculations, we identified thrombosis using the narrow version (not including phlebitis). In keeping with the intent of this descriptive characterization analysis, no formal statistical tests of comparisons between definitions or outcomes were performed.

All analytical codes used for the analysis are open source and have been made publicly available at https://github.com/ohdsi-studies/Covid19VaccineAesiIncidenceRate and https://github.com/ohdsi-studies/Covid19VaccineAesiDiagnostics.

3 Results

Results are openly available through https://data.ohdsi.org/Covid19VaccineAesiDiagnostics/.

3.1 Background Incidence Rate of TWT Events

Figure 2 summarizes the age- and gender-specific IRs of TWT, defined as patients with a new diagnosis of thrombosis (identified using the narrow set of diagnosis codes) and a new diagnosis of thrombocytopenia (identified either by a diagnosis code or a platelet measurement ≤ 150,000 per microliter) within 7 days. Age- and gender-specific IRs generated using the remaining five different TWT phenotype variants are summarized in Fig. 3 (fourth row). As illustrated in Fig. 2, the overall TWT IR ranged from 1.62 (in CPRD) to 150.65 (in MDCR) per 100,000 person-years. Figure 1 illustrates substantial heterogeneity across data sources and by age group and sex within the same data source. However, similar age and sex trends were observed in most data sources, where higher rates of TWT were observed among men of older age groups.

Figure 3 summarizes the age- and gender-specific IRs of TWT subtypes. The most common TWT subtypes were as follows: DVT with thrombocytopenia (IR ranged from 0.53 to 34.31 per 100,000 person-years), hemorrhagic stroke with thrombocytopenia (IR ranged from 0.06 to 18.46 per 100,000 person-years), ischemic stroke with thrombocytopenia (IR ranged from 0.05 to 49.85 per 100,000 person-years), and MI with thrombocytopenia (IR ranged from 0.39 to 56.17 per 100,000 person-years). On the other hand,

CVST with thrombocytopenia was only observed in nine data sources, where the IR ranged from 0.01 to 0.20 per 100,000 person-years. Splenic thrombosis (IR ranged from 0.05 to 1.09 per 100,000 person-years) and hepatic thrombosis (IR ranged from 0.01 to 0.24 per 100,000 person-years) were also very rare.

3.2 Proportional Difference (% Change) of TWT Incidence Rate by Phenotype

Supplementary Figures 1–3 reports the proportional difference (% change) in IR across alternative TWT definitions (see the electronic supplementary material).

As illustrated in Supplementary Fig. 3, using a broad version of diagnosis codes to identify thrombosis led to a minor (less than 5%) change in the estimated TWT IR across most data sources. The highest % change was observed in IQVIA_Germany data, with a 44.8% increase in overall TWT IR when using the broad version of thrombosis compared to the narrow.

Due to a lack of or incomplete laboratory data, relying on platelet measurements only to define thrombocytopenia was not at all possible in some data sources, such as MDCD and JMDC. This also led to artificially small IRs in others, such as CCAE, Optum_Extended_DoD, and Biobank_UK (Supplementary Fig. 1).

When measurements are available, using a lower threshold of 120,000 per microliter to define thrombocytopenia led to a large reduction in IR compared to that estimated by using a threshold of 150,000 per microliter. As illustrated in Supplementary Fig. 2, the % change in TWT IR was over
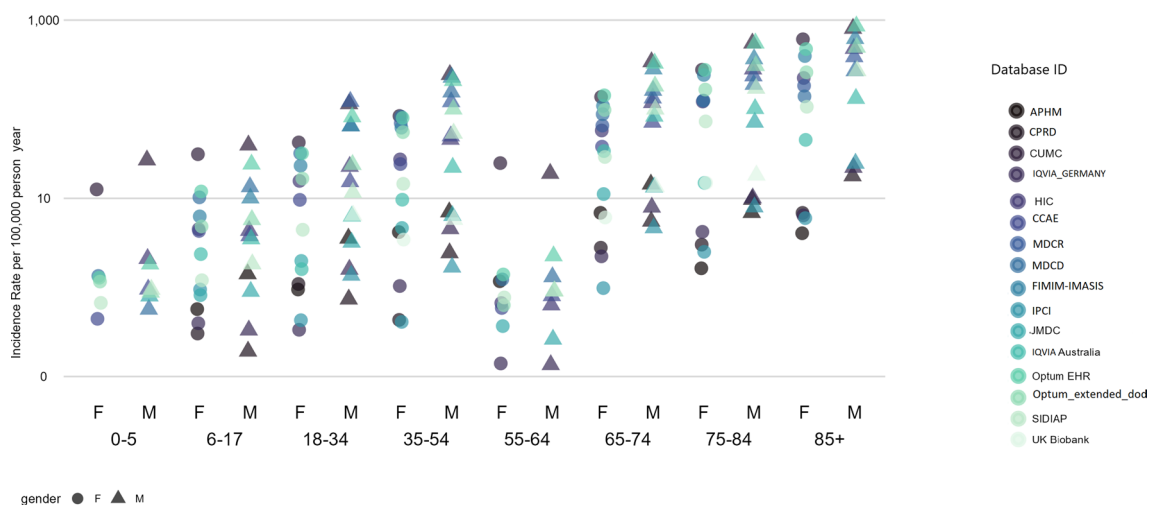
Fig. 2 Age- and gender-specific incidence rates of thrombosis with thrombocytopenia (TWT); we report the incidence rate per 1000 person-years for TWT, defined as patients with a new diagnosis of thrombosis (identified using the narrow set of diagnosis codes) and a new diagnosis of thrombocytopenia (identified either by a diagnosis code or a platelet measurement ≤ 150,000 per microliter) within 7 days
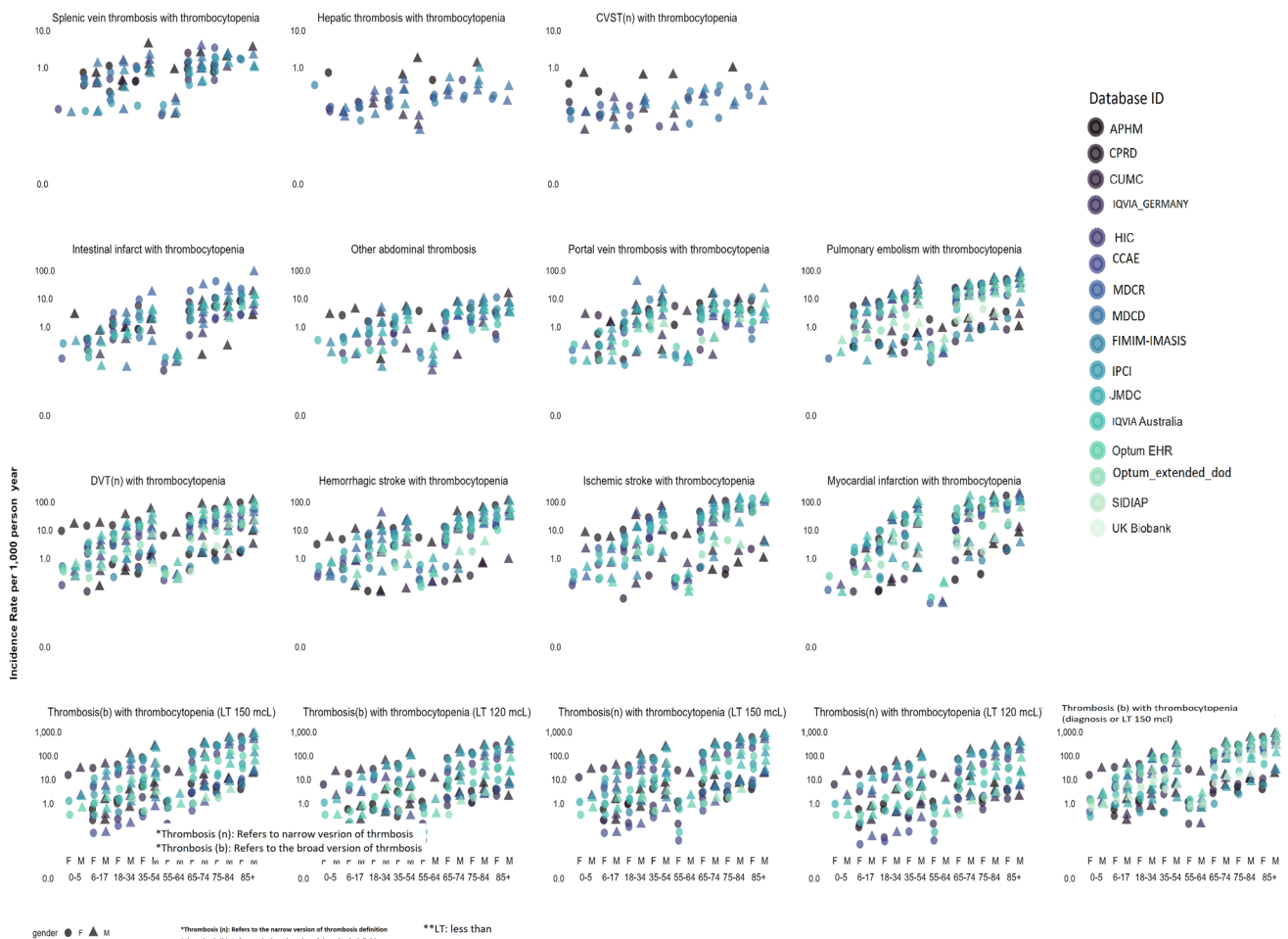
Fig. 3 Age- and gender-specific incidence rates of thrombosis with thrombocytopenia (TWT) subtypes and alternative definitions. In the *first three rows*, we report the incidence rate per 1000 person-years for TWT subtypes, defined as patients with a new diagnosis of a given thrombosis subtype (such as myocardial infarction, deep venues thrombosis, hepatic thrombosis) and a new diagnosis of thrombocytopenia (identified either by a diagnosis code or a platelet measurement ≤ 150,000 per microliter) within 7 days. In the *last row*, we report on the incidence rate per 1000 person-years for TWT alternative definition

50% when using 120,000 compared to 150,000 platelets per microliter in most data sources.

3.3 Baseline Characteristics

Table 2 illustrates selected baseline characterization of patients with TWT—defined as patients with a new diagnosis of thrombosis (identified using the narrow set of diagnosis codes) and a new diagnosis of thrombocytopenia (identified either by a diagnosis code or a platelet measurement ≤ 150,000 per microliter) within 7 days. The full sets of distributions of 1325 characteristics among all TWT cohorts can be found at https://data.ohdsi.org/Covid19Vac cineAesiDiagnostics/.

TWT patients across all data sources were likely to be men of older age with various comorbidities. Specifically, less than 8% of TWT patients were less than 40 years old and only around 39.22% were females (ranged from 27.8% in IQVIA_Australia to 49.4% in MDCD). An average of 66.65% had heart disease at baseline (ranged from 36.1% in CPRD to 70.8% in CC_Serbia and MDCD). Around 50% of the patients had hypertensive disorder in almost all data sources. Chronic liver disease ranged from 1.3% in CPRD to 15.1% in MDCD. Renal impairment ranged from 3.5% in IQVIA_Australia to 48.9% in Optum_Extended_DoD. Malignant neoplastic disease ranged from 8.3% in IQVIA_Australia to 36.50% in Biobank_UK. Finally, heparin use in the last 30 days ranged from 0% in Biobank_UK to 54.7% in Optum_EHR.

Similar trends were observed when alternative TWT definitions were examined. Covariant distribution using all 1325 characteristics were comparable when thrombocytopenia is defined using a measurement of ≤ 150,000 platelets per microliter of blood compared to when

Table 2 Baseline characteristics of thrombosis with thrombocytopenia patients

Characteristic	CCAE	Optum_ Extended_ DoD	MDCD	MDCR	Optum_EHR	CUMC	CPRD	Biobank_UK	HIC
Age group (years)									
0–19	2.50%	0.50%	3.50%	0.00%	1.00%	4.60%	0.00%	0.00%	0.20%
20–39	10.80%	2.60%	11.00%	0.00%	5.10%	5.90%	4.20%	0.00%	4.20%
40–59	53.50%	14.10%	37.70%	0.50%	22.00%	20.50%	18.10%	17.30%	13.50%
60–79	33.00%	52.00%	35.90%	54.80%	48.80%	44.00%	49.60%	80.40%	45.00%
80 +	0.00%	30.90%	12.00%	44.70%	23.50%	24.90%	27.50%	1.80%	37.20%
Gender = female	39.40%	41.00%	49.40%	40.30%	37.60%	41.70%	34.70%	39.10%	39.00%
Medical history									
Chronic liver disease	11.10%	8.50%	15.10%	5.00%	6.00%	6.10%	1.30%	5.20%	1.60%
Diabetes mellitus	26.80%	38.30%	35.00%	34.10%	29.50%	15.90%	3.00%	15.50%	14.90%
Hyperlipidemia	40.00%	63.90%	36.90%	51.50%	47.70%	18.40%	0.80%	21.40%	3.60%
Hypertensive disorder	55.60%	76.90%	65.20%	72.40%	60.40%	35.70%	3.60%	46.50%	24.10%
Renal impairment	27.30%	48.90%	42.90%	38.00%	37.00%	20.90%	8.20%	27.70%	20.00%
Heart disease	66.00%	79.90%	73.40%	83.90%	74.50%	69.60%	36.10%	55.70%	59.40%
Obesity	16.40%	17.50%	15.10%	7.80%	16.30%	4.10%	0.50%	6.30%	1.40%
Malignant neoplastic disease	23.50%	26.80%	17.90%	30.70%	15.80%	14.90%	15.00%	36.50%	14.50%
Heparin use in the last 30 days	8.60%	8.10%	6.80%	5.00%	54.80%	28.70%	9.40%	0.00%	1.20%
Anti-platelets (platelet aggregation inhibitors excluding heparin) use in the last 30 days	4.90%	8.00%	5.00%	11.20%	54.00%	31.60%	30.30%	9.90%	20.10%
Clopidogrel use in the last 30 days	3.30%	6.60%	2.10%	9.30%	15.20%	9.8%	5.6%	0.018%	4.70%
Aspirin use in the last 30 days	1.00%	0.60%	3.00%	1.00%	51.00%	30.00%	32.00%	9.00%	16.00%

| Characteristic	IPCI	IQVIA_Germany	SIDIAP	FIMIM-IMASIS	IQVIA_Australia	JMDC	CC_Serbia	APHM
Age group (years)								
0–19	0.00%	0.00%	0.70%	0.00%	7.00%	9.80%	0.00%	6.00%
20–39	>5.2%	3.70%	2.10%	2.60%	< 10.5%	13.60%	3.30%	11.60%
40–59	19.6%	17.40%	15.70%	16.50%	17.50%	43.10%	18.50%	27.10%
60–79	43.90%	52.00%	44.80%	48.70%	54.20%	33.60%	56.20%	39.30%
80 +	34.90%	26.40%	36.80%	32.10%	< 20.8%	0.00%	21.20%	18.70%
Gender = female	32.30%	30.00%	32.70%	36.10%	27.80%	41.50%	36.60%	37.30%
Medical history								
Chronic liver disease	> 2.6%	4.30%	6.10%	7.70%	< 3.5%	14.90%	2.30%	NA
Diabetes mellitus	18.00%	24.10%	22.90%	25.90%	6.30%	41.10%	13.30%	NA
Hyperlipidemia	4.80%	25.10%	34.20%	27.20%	20.80%	39.60%	10.60%	NA
Hypertensive disorder	18.00%	48.80%	44.50%	45.50%	33.30%	58.60%	49.60%	NA
Renal impairment	6.9%	14.40%	26.70%	19.70%	3.50%	20.20%	15.40%	NA
Heart disease	45.50%	60.60%	58.70%	63.90%	83.30%	64.10%	63.00%	NA
Obesity	> 2.6%	9.30%	12.00%	5.40%	< 3.5%	0.90%	1.10%	NA
Malignant neoplastic disease	13.80%	16.40%	16.50%	12.10%	8.30%	30.80%	10.30%	NA
Heparin use in the last 30 days	9.50%	6.10%	4.50%	15.00%	0.00%	2.40%	60.80%	2.00%
Anti-platelets (platelet aggregation inhibitors excluding heparin) use in the last 30 days	20.60%	16.80%	30.50%	10.10%	6.10%	11.80%		0.90%
Clopidogrel use in the last 30 days	3.20%	3.60%	6.10%	5.20%	0.038%	2.40%		0
Aspirin use in the last 30 days	18.00%	14.00%	25.70%	9.00%	5.00%	11.00%		1.00%

We report the proportion (%) of selected based-line characteristics for patients with a new diagnosis of thrombosis (identified using the narrow set of diagnosis codes) and a new diagnosis of thrombocytopenia (identified either by a diagnosis code or a platelet measurement ≤ 150.000 per microliter) within 7 days

APHM Health Data Warehouse of Assistance Publique—Hopitaux de Marseille, *Biobank_UK* UK Biobank, *CC_Serbia* University Clinical Center of Serbia, *CCAE* IBM® MarketScan® Commercial Claims and Encounters Database, *CPRD* Clinical Practice Research Datalink, *CUMC* Columbia University Irving Medical Center, *FIMIM-IMASIS* Information System of Parc Salut Mar Barcelona, *HIC* Health Informatics Centre from University of Dundee, *IPCI* Integrated Primary Care Information, *IQVIA_Australia* IQVIA® Australia Longitudinal Patient Data, *IQVIA_ Germany* IQVIA® Disease Analyser Germany, *JMDC* Japan Medical Data Center, *MDCD* IBM® MarketScan® Multi-State Medicaid Database, *MDCR* IBM® MarketScan® Medicare Supplemental and Coordination of Benefits Database, *Optum_EHR* Optum® de-identified Elec-

Table 2 (continued)

tronic Health Record Dataset, *Optum_Extended_DoD* Optum® De-Identified Clinformatics® Extended Data Mart Database—Date of death, *SIDIAP* Information System for Research in Primary Care

NA Not aviable
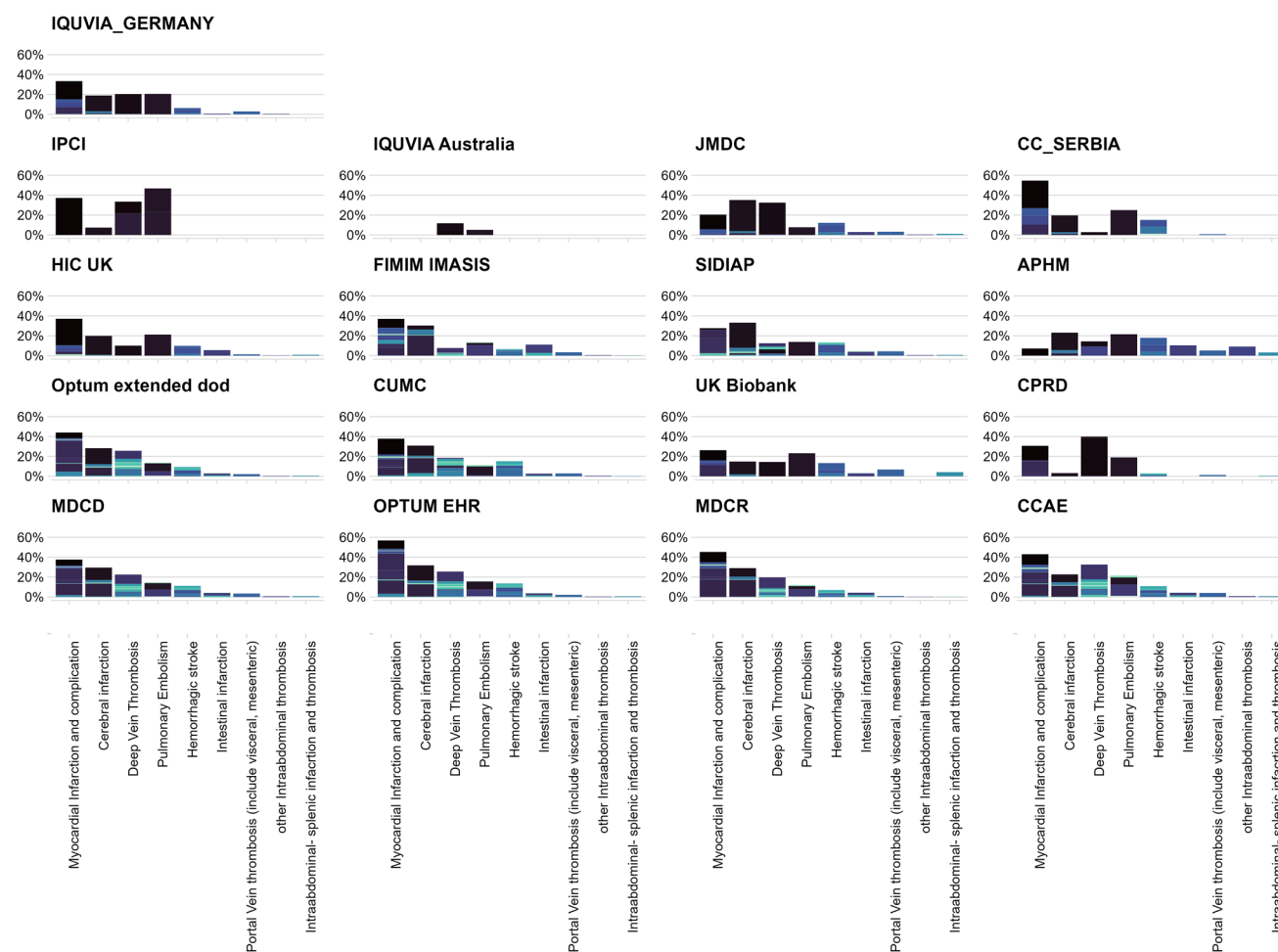
Fig. 4 The distribution of thrombotic events among thrombosis with thrombocytopenia, defined as patients with a new diagnosis of thrombosis (identified using the narrow set of diagnosis codes) and a new diagnosis of thrombocytopenia (identified either by a diagnosis code or a platelet measurement ≤ 150,000 per microliter) within 7 days, by data source. The thrombotic events are represented by single SNOMED CT concept and grouped by thrombosis subtypes. A single SNOMED CT concept is indicated by specific different color on each histogram, and each histogram represents a specific thrombosis subtype. For example, in Optum_Extended_dod, the color *dark purple* in the first histogram (myocardial infarction) is the SNOMED CT concept "Acute non-ST segment elevation myocardial infarction." The length of each bar represents how common one type of thrombosis was compared to the other (for example, myocardial infarction and cerebral infarction are more common than abdominal thrombosis). The diversity of the SNOMED CT concepts (representing diagnosis codes) occurring in each thrombosis type is represented by the variation of colors in each bar. For example, in IPCI and IQVIA_Germany, most thrombosis subtypes are driven by one or two SNOMED CT concept, while in IQVIA_Australia and CUMC, each thrombosis subtype is composed of a diversity of SNOMED CT concepts. *CUMC* Columbia University Irving Medical Center, *IPCI* Integrated Primary Care Information, *IQVIA_Australia* IQVIA® Australia Longitudinal Patient Data, *IQVIA_Germany* IQVIA® Disease Analyser Germany, *Optum_Extended_DoD* Optum® De-Identified Clinformatics® Extended Data Mart Database—Date of death, *SNOMED CT* Systematized Nomenclature of Medicine-Clinical Terms

thrombocytopenia is defined using a diagnosis code or a platelet measurement of ≤ 150,000 per microliter (Supplementary Fig. 4, see the electronic supplementary material). Comparable covariate distribution was also observed when thrombosis is defined using a broad version of diagnosis codes compared to when thrombosis is defined using a narrow version of diagnosis codes.

3.4 Makeup of Thrombosis Events Among TWT Patients

Figure 4 illustrates the distribution of thrombotic events among TWT patients. The thrombotic events are represented by a single SNOMED CT concept and grouped by thrombosis subtypes.

Despite variation in coding practices and granularity of medical terms used across different data sources the most common thrombotic events among TWT patients were concepts related to MI, ischemic and hemorrhagic strokes, DVT, and PE. For example, Optum_EHR MI-related diagnoses accounted for over 30% of TWT patients, while in APHM, cerebral infarction diagnoses accounted for over 20% of TWT cases and MI accounted for around 5%. In MDCR, MI- and PE-related diagnoses accounted for around 20%, followed by cerebral infarction, DVT, and hemorrhagic stroke (around 15%).

3.5 Background Incidence Proportion of Platelet Measures and Low Platelet Values

Table 3 illustrates the incidence proportion of platelet measures among patients with thrombotic events and the frequency of low platelet value among those with a platelet measure. In hospital EHR data sources, 26–88% of patients with thrombotic events had a platelet measurement within 7 days and 11–23% of those had a value that was 150,000 or less. In GP data sources, only 4–19% were recorded to have a platelet measurement within 7 days and 2–13% of those had a value of 150,000 or less.

4 Discussion

VITT has been identified as a rare but serious adverse event associated with some COVID-19 vaccines, and further research is required to better characterize and understand this new phenomenon. Observational healthcare data offer the opportunity for such research, but this research is predicated on reliable phenotyping of the outcome. In this study, we explored the historical trend of the co-occurrence of TWT using 17 observational health data sources across the world. We applied multiple phenotypes of TWT definitions, estimated the background rate of TWT, characterized TWT patients, and explored the makeup of thrombosis types among over 75 million TWT patients. Our findings highlight important limitations on the use of TWT definitions in real world data as a proxy for VITT.

While TWT was overall a rare event, considerable heterogeneity in background rates was observed across data sources. The observed magnitude of heterogeneity across sources within age and sex subgroups suggests that residual differences are present. The remaining heterogeneity may be related to differences in healthcare systems, setting, data capture processes, or true differences in subpopulations or individual patients. These differences may also be due to systematic error, selection bias, or differential outcome measurement error between data sources [15].

Using different TWT case ascertainment definitions led to different background rate estimates. Most notably, a lower incidence proportion is estimated when using a lower platelet measure of 120,000 per microliter of blood to identify thrombocytopenia. Also, our results suggest that TWT

Table 3 Incidence proportion of platelet measures among patients with thrombotic events and the frequency of a platelet measurement ≤ 150,000 per microliter among those with a platelet measure

Data source	Thrombosis events N	Any platelet measure N	Proportion of platelet measures (%)	Platelet < 150,000/µl N	Proportion of platelet < 150,000/ among any platelet measure (%)
FIMIM-IMASIS	27,841	24,545	88.16	4787	19.50
HIC	64,231	49,851	77.61	5464	10.96
CUMC	234,584	134,499	57.34	31,439	23.37
Optum_EHR	3,684,392	1,781,677	48.36	322,018	18.07
CC_Serbia	19,236	5079	26.40	1179	23.21
IQVIA_Australia	18,272	3523	19.28	133	3.78
SIDIAP	414,458	79,106	19.09	10,625	13.43
IQVIA_Germany	395,295	28,072	7.10	1553	5.53
CPRD	412,313	27,621	6.70	1401	5.07
Biobank_UK	39,305	2182	5.55	53	2.43

Biobank_UK UK Biobank, *CC_Serbia* University Clinical Center of Serbia, *CPRD* Clinical Practice Research Datalink, *CUMC* Columbia University Irving Medical Center, *FIMIM-IMASIS* Information System of Parc Salut Mar Barcelona, *HIC* Health Informatics Centre from University of Dundee, *IQVIA_Australia* IQVIA® Australia Longitudinal Patient Data, *IQVIA_Germany* IQVIA® Disease Analyser Germany, *Optum_EHR* Optum® de-identified Electronic Health Record Dataset, *SIDIAP* Information System for Research in Primary Care

phenotypes that strictly rely on observed measurements of platelet count without considering the clinical diagnosis of thrombocytopenia are not feasible for many available observational data sources. Due to the lack of completeness of lab measurements in most real world data sources, relying on platelet count to possibly improve the specificity of the definition is likely to lead to very low sensitivity. As we observed notable differences in IRs by phenotype and/or data source, caution is needed when IRs are compared or interpreted across time or population.

Few studies have reported on the IR of TWT across different populations. Bhuyan et al. reported a TWT background rate of 3.75 (3.51–4.00) per 1 million persons per 14 days in a US data source, which is within the range reported in this study [2]. In a similar analysis across six European countries, Burn et al. reported that TWT background rates varied by data source and ranged from 0.5 to 4.4 per 100,000 person-years across different types of thrombosis [13]. Consistent with our findings, the authors found that the incidence of TWT was higher among men of older age, with those affected typically having more comorbidities and greater medication use than the general population.

VITT cases reported after the adenovirus COVID-19 vaccines were likely to be healthy women of reproductive age [5, 16, 17]. Our data suggest that the patient profiles of the captured TWT cases (using any of the proposed phenotypes and across all data sources) are not consistent with such a profile. In contrast, TWT patients in this study were likely to be men of older age with a high prevalence of comorbidities. In addition, the majority of the TWT patients identified in this study had one of the common thrombosis events such as MI, strokes, PE or DVT. This is not consistent with the type of thrombosis observed in VITT cases after the adenovirus COVID-19 vaccines, where rare events such as CVST and splenic and hepatic thrombosis were also observed [16, 17].

A consistent outcome definition that can be applied in observational health data is necessary for conducting research studies as well as clinical case detection activities. Our data suggest that the current definitions for TWT are likely to capture cases that are not a true representation of the new emerging clinical phenomena of VITT, but are merely a coincidental co-occurrence of two common clinical events (thrombosis and thrombocytopenia). As such, the co-occurrence of thrombosis and thrombocytopenia among some patients can be explained by the presence of comorbidities such as malignancies and liver disease. Consequently, applying any of the current definitions of TWT to identify VITT cases is likely to lead to considerable levels of false positives. Further research is needed to quantify the associated misclassification error. Observational studies that rely on the co-occurrence of diagnosis codes of thrombosis and thrombocytopenia to investigate the association between exposures and VITT [18, 19] need to be carefully assessed

for the effect of measurement bias associated with capturing VITT cases. Statistical approaches may be utilized to account and check for balance on possible confounders and factors that may indicate differential measurement error such as malignancies, liver disease, and others.

Another challenge that may face future observational research on VITT is surveillance bias. Given the emerging awareness of VITT as a phenomenon, clinicians are more likely to request a platelet count (or other related diagnostic workup) for patients presenting with thrombosis, especially among patients exposed to the vaccines. We found a considerable heterogeneity in the observed incidence of platelet measures among patients with thrombosis and in the proportion of measurements that are low. Given this heterogeneity, surveillance bias may be difficult to avoid or control for in related observational studies. Quantification of VITT risks may be extremely difficult in situations when surveillance bias is likely, and scientists and readers of scientific texts should be aware of such problems in observational studies [20].

The primary limitation of this study is that all TWT definition outcomes are subject to measurement error. All definitions were based on the presence of diagnosis codes and measurement values and were not validated further. While we utilized multiple case definitions, all our analysis relied on data from 2017 to 2019 using a target population of all people in each data source with more than 365 days of observation indexed on 1 January, 365 days of time at risk, and 365 days of outcome-specific clean windows for thrombosis and 90 days for thrombocytopenia to allow for recurrent events.

Some limitations relate to the use of each data source. None of the data sources used have full coverage of all diagnosis and measurement data occurring in both inpatient and outpatient settings. For example, information on hospital admission was not available in the primary care datasets used (CPRD GOLD in the UK, IQVIA in Germany and Australia, and IPCI in the Netherlands) and events that happened during inpatient visits were not included. On the other hand, the EHR data sources were subject to incomplete capture of medical events and measurements recorded in other healthcare institutions. The bias of incomplete information was partially mitigated by including only those patients who had at least 1 year of continuous observation. The administrative claims data sources offered reliable data capture but lacked measurement lab data.

In conclusion, our research across 17 data sources suggests that identifying VITT in observational data presents a substantial challenge, as implementing VITT case definitions based on the co-occurrence of TWT results in large and heterogeneous IRs in a cohort composed of patients with baseline characteristics that are inconsistent with the VITT cases reported to date. Thrombosis and thrombocytopenia

each independently are relatively commonly occurring conditions, and as such, the temporal co-occurrence of the two is not uncommon and cannot be assumed to be negligible. Our characterization of TWT highlights that thrombosis, when defined to include common venous and arterial events, is largely driven by background rates of deep vein thrombosis, MI, and ischemic stroke, and underscores that thrombocytopenia, when defined by platelet measurements, can be highly variable based on the source data capture process. Considering these findings, we caution against using any of the current TWT phenotypes in observational data as a basis for estimating background rates for VITT safety surveillance and advise that further refinement of the case definition is needed before observational data can be reliably used to generate unbiased population-level effect estimates. Individual case reviews may potentially provide insights that can suggest further phenotype definition refinements. Finally, additional research is needed to fully assess the potential of using the co-occurrence of thrombosis with thrombocytopenia in observational data to capture VITT patients for safety and epidemiological studies. Most importantly, the associated measurement error and its variance need to be accurately estimated and incorporated in relevant studies and findings.

Declarations

References

1. Shimabukuro TT. Update: thrombosis with thrombocytopenia syndrome (tTS) following COVID-19 vaccination. 2021.
2. Bhuyan P, et al. Very rare thrombosis with thrombocytopenia after second AZD1222 dose: a global safety database analysis. Lancet. 2021;398(10300):577–8.
3. Long B, Bridwell R, Gottlieb M. Thrombosis with thrombocytopenia syndrome associated with COVID-19 vaccines. Am J Emerg Med. 2021;49:58–61.
4. Shay DK. Safety Monitoring of the Janssen (Johnson & Johnson) COVID-19 Vaccine—United States, March–April 2021. MMWR. Morbidity and Mortality Weekly Report, 2021;70.
5. See I, et al. Case series of thrombosis with thrombocytopenia syndrome following COVID-19 vaccination—United States, December 2020–August 2021. medRxiv. 2021;384:2124.
6. Chen RT, Black S. Proposed Brighton Collaboration process for developing a standard case definition for study of new clinical

syndrome X, as applied to Thrombosis with Thrombocytopenia Syndrome (TTS). 2021. https://brightoncollaboration.us/wp-content/uploads/2021/04/TTS-Case-Finding-and-Definition-Process.v1.0-1-1.pdf. Accessed 11 Apr 2022.

7. Chen RT, Black S. Updated Proposed Brighton Collaboration process for developing a standard case definition for study of new clinical syndrome X, as applied to Thrombosis with Thrombocytopenia Syndrome (TTS).

8. Overhage JM, et al. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54–60.

9. OHDSI CDM Documentation. GitHub IO. https://github.com/OHDSI/CommonDataModel. Accessed 11 Apr 2022.

10. Vlug A, et al. Postmarketing surveillance based on electronic patient records: the IPCI project. Methods Inf Med. 1999;38(04/05):339–44.

11. Biobank data. https://www.ukbiobank.ac.uk/. Accessed 11 Apr 2022.

12. JMCD data. https://www.jmdc.co.jp/en/jmdc-claims-database. Accessed 11 Apr 2022.

13. Burn E, et al. Background rates of five thrombosis with thrombocytopenia syndromes of special interest for COVID-19 vaccine safety surveillance: incidence between 2017 and 2019 and patient profiles from 20.6 million people in six European countries. medRxiv. 2021;384:403.

14. Rao G. CohortDiagnostics. 2020; https://ohdsi.github.io/CohortDiagnostics/. Accessed 11 Apr 2022.

15. Li X, et al. Characterising the background incidence rates of adverse events of special interest for covid-19 vaccines in eight countries: multinational network cohort study. BMJ. 2021;373:n1435.

16. Shimabukuro TT. Thrombosis with thrombocytopenia syndrome (TTS) following Janssen COVID-19 vaccine. 2021.

17. MacIntyre CR, et al. Thrombosis with Thrombocytopenia Syndrome (TTS) following AstraZeneca ChAdOx1 nCoV-19 (AZD1222) COVID-19 vaccination–A risk–benefit analysis for people< 60 years in Australia. Vaccine. 2021;39(34):4784–7.

18. Pottegård A, et al. Arterial events, venous thromboembolism, thrombocytopenia, and bleeding after vaccination with Oxford-AstraZeneca ChAdOx1-S in Denmark and Norway: population based cohort study. BMJ. 2021;373:n1114.

19. Andrews NJ, et al. Risk of venous thrombotic events and thrombocytopenia in sequential time periods after ChAdOx1 and BNT162b2 COVID-19 vaccines: a national cohort study in England. Lancet Regional Health-Europe. 2022;13: 100260.

20. Hemminki K, et al. Surveillance bias in cancer risk after unrelated medical conditions: example urolithiasis. Sci Rep. 2017;7(1):1–4.

Authors and Affiliations

Azza Shoaibi[1,2] · **Gowtham A. Rao**[1,2] · **Erica A. Voss**[1,2] · **Anna Ostropolets**[2,3] · **Miguel Angel Mayer**[4] · **Juan Manuel Ramírez-Anguita**[4] · **Filip Maljković**[5] · **Biljana Carević**[6] · **Scott Horban**[7] · **Daniel R. Morales**[7] · **Talita Duarte-Salles**[8] · **Clement Fraboulet**[9] · **Tanguy Le Carrour**[10] · **Spiros Denaxas**[11] · **Vaclav Papez**[11] · **Luis H. John**[12] · **Peter R. Rijneek**[12] · **Evan Minty**[13] · **Thamir M. Alshammari**[2,14] · **Rupa Makadia**[1,2] · **Clair Blacketer**[1,2] · **Frank DeFalco**[1,2] · **Anthony G. Sena**[1,2] · **Marc A. Suchard**[2,15] · **Daniel Prieto-Alhambra**[16] · **Patrick B. Ryan**[1,2]

1 Janssen Research & Development, LLC, 1125 Trenton-Harbourton Road, Titusville, NJ 08560, USA

2 Observational Health Data Sciences and Informatics (OHDSI), OHDSI Collaborators, New York, NY, USA

3 Department of Biomedical Informatics, Columbia University Medical Center, New York, NY, USA

4 Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain

5 Heliant Ltd, Belgrade, Serbia

6 Department of Hospital Epidemiology, University Clinical Center of Serbia, Belgrade, Serbia

7 Health Informatics Centre (HIC), University of Dundee, Dundee, Scotland

8 Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain

9 Assistance Publique - Hôpitaux de Marseille, Marseille, Provence-Alpes-Cote d'Azur, France

10 Easter-eggs, Paris, France

11 University College London, Institute of Health Informatics, London, UK

12 Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands

13 Departmnet of Biomedical Data Sciences, Stanford University, California, USA

14 Collage of pharmacy, Riyadh Elm University, Riyadh, Saudi Arabia

15 Departments of Biomathematics and Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

16 Centre for Statistics in Medicine (CSM), Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDROMS), University of Oxford, Oxford, UK