Practice of Epidemiology

# Inconsistency in UK Biobank Event Definitions ~~from~~From Different Data Sources and ~~its~~Its Impact on Bias and Generalizability: A Case Study of Venous Thromboembolism

Emily Bassett, James Broadbent, Dipender Gill, Stephen Burgess, and Amy M. Mason*

* Correspondence ~~to Amy M. Mason,~~[BS1] to Dr. Amy M. Mason, MRC Biostatistics Unit, University of Cambridge, East Forvie Building, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, United Kingdom (e-mail: am2609@medschl.cam.ac.uk).

Initially submitted April 5, 2023; accepted for publication November 16, 2023.



The UK Biobank study contains several sources of diagnostic data, including hospital inpatient data and data on self-reported conditions for ~~–~~approximately 500,000 participants~~,~~ and primary~~ -~~care data for ~~–~~approximately 177,000 participants (35%). ~~Epidemiological~~Epidemiologic investigations require a primary disease definition, but whether to combine data sources to maximize statistical power or focus on ~~one~~only 1 source to ensure a consistent outcome is not clear. The consistency of disease definitions was

investigated for venous thromboembolism (VTE) by ~~looking at~~evaluating overlap when defining cases from 3 sources: hospital ~~in-patient~~inpatient data, primary~~-~~care reports, and self-reported questionnaires. VTE cases showed little overlap between data sources, with only 6% of reported events for ~~those~~persons with primary~~-~~care data being identified by all ~~three of~~ 3 sources (hospital, primary~~-~~care, and self-~~report,~~reports), while 71% appeared in only ~~in one~~1 source. Deep vein thrombosis~~-~~only events represented 68% of self-reported VTE cases and 36% of hospital-reported VTE cases, while pulmonary embolism~~-~~only events represented 20% of self-reported VTE cases and 50% of hospital-reported VTE cases. Additionally, different distributions of sociodemographic characteristics were observed; for example, patients in 46% of hospital~~-~~reported VTE cases were female, compared with 58% of self-reported VTE cases. These results illustrate how seemingly neutral decisions taken to improve data quality can affect the representativeness of a ~~dataset.~~data set.

bias; deep vein thrombosis; event definition; generalizability; pulmonary embolism; representativeness; sociodemographic characteristics; UK Biobank; venous thromboembolism

Abbreviations: ~~???~~DVT, deep vein thrombosis; PE, pulmonary embolism; VTE, venous thromboembolism.



Venous thromboembolism (VTE) is a condition that occurs when a blood clot forms inside the veins, preventing blood flow. Its incidence is roughly 100 events per 100,000 person-years (1, 2). Approximately two-thirds of cases are deep vein thrombosis (DVT) (3–5), where the blood clot forms in a deep vein, typically the pelvis, thigh, or lower leg. A third of cases are pulmonary embolism (PE), which occurs when the clot breaks loose and travels to the lungs (3–5). In rare cases, thrombosis may occur in other veins.

Factors associated with greater[SB2] risk of VTE include obesity (6), height (7), smoking ~~status~~[8]status (8), hypertension ~~(8)~~(9), social deprivation ~~((9, 10)~~, 11), education ~~(8)~~(9), immobilization ~~(11)~~(12), surgery ~~(12)~~(13), use of hormone replacement therapy ~~(HRT)~~ or oral contraceptives ~~((13, 14)~~, 15), and pregnancy ~~(15)~~(16). Risk of VTE also increases with age, as does the proportion of VTEs that are ~~PEs~~[17]PEs (17). There is little consistent evidence for overall differences in VTE risk by sex: ~~there~~There are reports of higher rates for men (2, ~~8, 16)~~9, 17), no significant difference ~~((17, 18)~~, 19), and higher rates for women (3, ~~19–21)~~20–22) when combining across all ages. However, there may be different patterns of risk across the lifetime, with risk ~~in women~~being higher ~~in~~during the reproductive years among women and ~~risk in men~~ higher in old age among men (2, 3, ~~19)~~20), and men being at higher risk of recurrent events ~~((22, 23)~~, 24). VTE risk is higher for individuals of African ancestry than for those of European ancestry ~~(24)~~(25)~~,~~ and higher for individuals of European ancestry than for those of Hispanic and Asian ancestry ~~((25, 26)~~, 27).

Previous studies on VTE in the UK Biobank have used a combination of self-reported physician diagnosis of DVT or PE, details from hospital inpatient records, and death certificates, or a subset of these sources ~~(27–29)~~(28–30). Details from primary ~~-~~care records in the UK Biobank are used less often ~~used~~, as they are not available for the entire cohort. Other studies have used different sources to determine VTE cases. A 23andMe study used self-reported VTEs alone ~~(30)~~(31), while a large Norwegian study used a combination of inpatient and outpatient hospital records (3). ~~As~~Because studies do not typically ~~breakdown~~break down results by source of diagnosis report, it is unclear how much different reporting sources ~~of report~~ could ~~impact~~affect the number and sociodemographic makeup of identified cases.

Electronic health records used for research are susceptible to “informed presence bias~~”~~.”~~.~~ Patients do not appear in health records at random, but are influenced by the

symptoms they have and how well they can communicate them to clinicians ~~((31, 32)~~, 33). Clinician suspicion determines who is scored for suspected DVT/PE. A probability assessment using the modified ~~deep vein thrombosis~~DVT and ~~pulmonary embolism~~PE Wells ~~Scores~~scores determines who then gets to access further tests such as D-dimer measurement, ~~ultrasound~~ultrasonography, and radiological imaging ~~(33)~~(34). Currently in the ~~UK's~~United Kingdom's National Health Service, individuals with DVT and low-risk PE can be treated as outpatients ~~(34)~~(35), while those with high-risk PE would be admitted to a hospital as inpatients—this could influence which data sources record a VTE and whose VTEs get recorded.

~~The~~Our aim ~~of~~in this investigation is to determine how using different sources of data may ~~impact~~affect VTE case populations within the UK Biobank. We will do this by considering how closely reports of VTE from different data sources correspond and whether the populations reported as cases are similar. We will not consider any specific reporting method as a "gold standard" of truth to determine the accuracy of other methods, nor will we attempt to estimate VTE incidence in the general UK population. Instead, we will compare how similar each definition is to the others within the UK Biobank.

## METHODS

### Study participants

The UK Biobank is a large prospective cohort study containing diagnostic data for 503,317 participants~~,~~ aged 37~~ to~~ 73 years~~,~~ who were recruited across England, Scotland, and Wales between 2006 and 2010.

### Data sources within the UK Biobank

*Self-reported outcomes.*

At ~~enrolment~~enrollment and resurvey, participants answered a touch-screen questionnaire, including specific questions about prior physician diagnoses of blood clots in the leg or lungs as well as more general questions about serious medical conditions. These were followed up with a verbal interview ~~(35)~~(36). Where participants were not certain about prior diagnoses, their responses were matched where possible to health conditions in a coding tree by a medical professional ~~(36)~~(37). Self-reported VTEs were coded as either DVT, PE, or other VTE.

*Hospital data.*

~~ICD-9 and -10~~*International Classification of Diseases, Ninth Revision*, and *International Classification of Diseases, Tenth Revision*, coded hospital inpatient episodes were obtained from the Hospital Episode Statistics provider for England, the Patient Episode Data for Wales, and the Scottish Morbidity Records for Scotland ~~(37)~~(38). These ~~datasets~~data sets contain information on admission and discharge, operations, diagnoses, maternity care, and psychiatric care. Main and secondary diagnoses throughout the ~~patient~~patient's admission are recorded. ~~This~~These data ~~is~~are only available within the UK Biobank for patients who are admitted to the hospital and occupy a bed.

*Death certificate data.*

~~ICD-10~~*International Classification of Diseases, Ninth Revision*, coded national death registry data were obtained from the Health and Social Care Information Centre (now NHS England) for England and Wales~~,~~ and the Information Services Department ~~(ISD)~~ for Scotland ~~(38)~~(39). This includes primary and secondary causes of death determined by a ~~doctor~~physician who attended the patient in their last illness or a coroner ~~(39)~~(40).

*Primary ~~-~~care records.*

Primary--care data were captured for 230,000 participants, covering records from selected general practice services in England, Scotland, and Wales (40)(41). We took a subset of 177,363 participants that ensured continuous coverage overlapping with their recruitment into the UK Biobank-- details. Details on choices made can be found in Web Appendix 1. (available at https://doi.org/10.1093/aje/kwad232).

## Event definitions

VTE cases were determined using the four4 data sources: death certificate data, hospital data, self-reported outcomes, and primary--care records (in the primary--care cohort only). We considered events reported by each data source in turn, andas well as a combined outcome including events reported by any data source.

VTE cases were broken down into PE and DVT via matching to any of the codes in Web Table 1. If a participant matched to VTE, but not to PE or DVT, they were classed as "Otherother VTE"."

## Medication use

One concern about the self-reported outcomes is that case definitions may be much less accurate. To investigate this concern, we considered whether patterns of relevant medication use were similar between the cases reported via different sources.

VTEs are often treated with anti-coagulants.anticoagulants. While warfarin is not recommended as the first line treatment in the current UK guidelines, the standard of care prior to 2020 was a low--molecular-weight heparin bridge followed by warfarin (33)(34).

There are two2 sources of general medication usage data within the UK Biobank. One is self-reported data, collecting lists of all regularly taken prescription medications during the touch--screen questionnaire and verbal interview (data- field 20,003) (35).20003) (36). The other source is the primary--care records prescription data, which is onlyare available only for

the cohort with primary--care data (data- field 42,03942039). Matching on drug names was undertaken to identify participants who had taken either any anticoagulant or warfarin at some point (details of matching are shown in Web Table 2).

## Statistical methods

We cross-tabulated the events in both the full UK Biobank sample and in the cohort with available primary--care data. In both groups, we compared anti-coagulantanticoagulant medication use and demographic data defined by the various data sources. The variables we compared were age at baseline, gendersex, smoking status, ethnicity, body mass index (BMIweight (kg)/height (m)$^2$), current employment status, highest level of education, history of manual or shift work, Townsend deprivation index (a greater meansscore reflects more depriveddeprivation), house ownership, and car ownership. For participants in England, we also looked at the Index of Multiple Deprivation (IMD) and the scores that determine the IMDIndex of Multiple Deprivation.

Proportional Venn diagrams were plotted to getobtain a visual understanding of the various overlaps of cases. Agreement between the methods was evaluated using the Fleiss κ value between all of the sets and CohanCohen's κ pairwise between each method. κKappa coefficients <less than 0.6 are taken asto indicate inadequate agreement, in line with recommendations for health-related studies,; those[SB3] of 0.6–0.87 are taken as moderate agreement, 0.8–0.9 as strong agreement, and >greater than 0.9 as almost perfect agreement (41)(42).

# RESULTS

## Study population

Table 1 contains a summary of the overall demographicsdemographic characteristics of the UK Biobank- participants. The defined primary--care cohort reproducesreproduced the

known biases within the UK Biobank ~~dataset~~data set—there ~~is~~was a "healthy volunteer bias~~",~~," with participants being more likely to be older, to be female, to have a lower ~~BMI,~~body mass index, to smoke less, to live in less socioeconomically deprived areas, and to have a greater rate of higher education[SB4] than the average person in the ~~UK~~[43]United Kingdom (43) (Web Table 3). The primary ~~-~~-care cohort ~~has~~had a similar ~~gender~~sex and age balance, a slightly greater proportion of White participants, higher rates of unemployment, and lower rates of higher education ~~compared with~~than the full ~~dataset~~data set.

**Event definitions in the full UK Biobank sample**

No single data source ~~captures~~captured all VTE cases, and the percentage captured by different methods ~~varies~~varied by case type. Taking a report from any source as a case and breaking cases down by ~~sub-diagnosis:~~subdiagnosis, 13% of participants had both PE and DVT, 54% had DVT only, 30% had PE only, and 3% had a VTE that fit into neither category.

There ~~is~~was little agreement between self-report data and inpatient hospital data ($\kappa$ = 0.32). Only 20.2% of VTE cases ~~are~~were reported by both sources (Figure 1), while 79.8% ~~appear~~appeared only in a single source (51.5% ~~appear~~appeared only in self-reports and 28.3% ~~appear~~appeared only in hospital data). There ~~is~~was a larger overlap of hospital events being self-reported when we ~~restrict~~restricted the data to prevalent events, but we ~~do~~did not see a matching hospital report for the majority of incident self-reported events (see Web Figure 1~~) (available at https://doi.org/10.1093/aje/kwad232~~).

The data from death certificates did not add any additional clarity (Web Tables 4 and 5, Web Figures 2 and 3). There were 741 cases of VTE identified from primary and secondary causes of death, of which 388 cases did not appear in another data source. Due to

the small proportion of cases identified through this method ~~(~~(approximately 4%), we did not analyze death certificate data further.

Considering the ~~two sub-diagnoses~~2 subdiagnoses (DVT and PE), there ~~is~~was a difference in the reporting source ~~of the report~~ by ~~sub-diagnosis: more~~subdiagnosis: More DVTs ~~are~~were only self-reported (67.6%) than ~~are~~were in hospital records only (16.3%) or both in the hospital records and self-reported (16.1%), whereas ~~PE are~~PEs were most likely to be in hospital records only (46.9~~%)~~%), although nearly a third ~~appear~~appeared only as self-reports (32.3%).

The proportion of DVT to PE events also ~~varies~~varied with the data source (Figure 2). If we ~~consider~~considered only hospital data, 50% of events ~~are~~were PE only, 36% were DVT only, and 9% were both~~;~~, whereas when taking self-reported outcomes as the data source, 20% of events ~~are~~were PE only, 68% were DVT only, and 11% ~~are~~were both.

We also ~~see~~saw variation in the ~~demographics~~demographic characteristics of the identified cases (Table 2). For example, using only hospital data, the case population ~~is~~was 45.7% female, while using the self-reported data the case population ~~is~~was 58.3% female.

**Event definitions in primary ~~-~~care cohort**

The primary ~~-~~care cohort within the UK Biobank ~~shows similar~~ showed patterns of case overlap ~~similar~~ to those of the full participant group (Figure 3). Adding the additional cases from the primary ~~-~~care records ~~does~~did not explain many of the undocumented self-reported VTE events and ~~adds~~added an additional set of otherwise uncaptured outcomes.

The highest agreement ~~is~~was between hospital data and self-reported data ($\kappa$ = 0.33), but this ~~is~~was still inadequate in terms of concordance. Primary ~~-~~care data ~~have~~had slightly more agreement with hospital data than self-reported data ($\kappa$ = 0.29 vs. 0.21). Only 5.5% of VTE cases ~~are~~were reported by all ~~three~~3 sources, while 71.3% ~~appear~~appeared only in a

single source: 43.9% ~~appear~~appeared only in self-reports, 21.8% ~~appear~~appeared only in hospital data, and 5.6% ~~appear~~appeared only in primary~~-~~-care reports. Splitting the data into prior ~~and post~~ registration and postregistration, there ~~is~~was a clear time-period effect due to the lack of self-reports ~~post registration~~postregistration for many participants and the sparsity of hospital reports prior to registration. In all cases, the primary~~-~~-care data and the hospital data ~~have~~had little overlap~~.~~ (Web ~~Figure~~Figures 4 and ~~Web Figure~~ 5~~;~~, Web Tables 6–8).

There ~~is~~was a difference in the source of the report for the ~~sub-diagnoses: most~~subdiagnoses: Most DVTs ~~are~~were only self-reported (60.4%), while more PEs ~~are~~were in hospital records only (36.1%) than in any other category. There ~~is~~was slightly better agreement between sources for PE (κ ~~between~~= 0.33–0.35 when comparing hospital data, primary~~-~~-care data, and self-~~report~~reports) than for DVT (κ ~~between~~= 0.14–0.27). (See Web Table 9~~.~~.)

## Anticoagulant usage in the primary~~-~~-care cohort

There ~~are~~were different patterns of reported anticoagulant use between the different case groups, but all ~~have~~had much higher rates than the control group (Web Table 10). ~~Cases of~~Patients whose VTE ~~was~~ identified using only hospital data ~~are~~were more likely to have a record of anticoagulant drug use at some point in their primary~~-~~-care records (64.7% used some sort of anticoagulant~~,~~; 50.4% were on warfarin~~))~~, whereas those identified via primary~~-~~-care records only had much lower use (37.9% and 26.8~~%~~%, respectively). Self-reported~~-~~-only cases fell between these ~~two~~2 groups (51.2% and 33.2~~%~~%, respectively). In contrast, anticoagulant drug use ~~amongst~~among controls (~~that is,~~i.e., individuals with no reported VTE event from any source) was much lower (18.9% and 2.5~~%~~%, respectively). This ~~provides~~provided an indication that there ~~are~~were likely to be true VTE events ~~amongst~~among the self-reported~~-~~-only cases. Self-reported rates of anticoagulant use were much lower~~,~~ but more consistent between definitions (Web Figure 6).

**Differences in ~~socio-demographics~~sociodemographic characteristics between cases from each data source**

The self-reported cases ~~are~~were younger and more likely to be female than the hospital data cases. They ~~are~~were more likely to have been assessed at the UK Biobank ~~centres~~centers in Wales~~,~~ and less likely to have been assessed in Scotland. There ~~are~~were also differences between these ~~two~~2 case groups in terms of mean ~~BMI~~body mass index, house ownership, and multiple car ownership. The cases identified by primary ~~-~~-care data ~~are~~were somewhere between the other ~~two~~2 case groups in terms of both ~~gender~~sex and age, with lower levels of deprivation~~,~~ and higher rates of house and multiple car ownership.

## DISCUSSION

Our investigation found that using different data sources in the UK Biobank results in substantial differences in the number, balance, and ~~socio-demographics~~sociodemographic characteristics of VTE cases considered. None of the data sources ~~have~~had good agreement with each other. The majority of DVT events ~~appear~~appeared only as self-reported outcomes. For PE, the largest group of events ~~is~~was reports from hospital data only. One likely reason for this is severity, with DVTs being more likely to be treated in outpatient settings ~~(34)~~(35) while PE is more often life-threatening, resulting in hospitalization. Hospital reports ~~constitute~~constituted the majority of ~~post-registration~~postregistration events in the study, while the majority of events prior to registration ~~are~~were self-reported. However, this ~~is~~was not a pure effect of time~~-~~ period, as there ~~are~~were self-reports after registration that ~~are~~were not seen in hospital records and hospital reports before registration than ~~are~~were not-self reported. For both diseases, only a small proportion of participants ~~are~~were detected by multiple data sources as having an event. This suggests a need to be attentive to how use of different data sources may influence case definition and composition.

Large studies ~~into~~of patient characteristics affect our perception of diseases: ~~for~~For example, studies claiming that VTE predominantly affects male or female patients ~~likely impact~~probably affect physicians' ~~perception~~perceptions of reported symptoms, as has been seen for cardiovascular disease ~~(42)~~(44) and depression ~~(43)~~(45). This can ~~impact~~affect how readily they diagnose future patients. As a result, decisions drawn from biased data can lead to greater health inequality, as has previously been observed for algorithmic decisions ~~(44, 45)~~(46, 47). Future studies can also be biased by these perceptions, with well-meaning and seemingly neutral decisions taken to improve data quality ~~impacting~~affecting the representativeness of subsequent research findings using the same case definitions.

## Accuracy of self-reported data for determining health outcomes

Self-reported outcome data are often viewed unfavorably compared ~~to~~with hospital-reported or physician-collected data. However, several studies considering the accuracy of self-reporting of VTEs ~~compared to~~in comparison with physician-~~-~~collected data have found little to substantiate this view. Heckbert et al. ~~looked at~~(48) evaluated the agreement between self-~~report~~reports and hospital discharge codes for 99,500 participant reports in the Women's Health initiative. The concordance between self-reported and hospital-reported events was good ($\kappa$ = 0.67 for PE~~,~~ and $\kappa$ = 0.71 for DVT). However, both self-reported and hospital-~~-~~reported events had higher concordance with physician-adjudicated events for PE ($\kappa$ = 0.83 and $\kappa$ = 0.84, respectively~~),~~) and for DVT ($\kappa$ = 0.72 and $\kappa$ = 0.80) ~~(46)~~(48)~~. This is~~. These are much higher levels of agreement than we saw in the UK Biobank, which may be because participants were asked specifically about PE and DVT, whereas the UK Biobank asked an open question about physician-~~-~~diagnosed conditions. Another possibility is that ~~that~~ the low overlap ~~is because~~reflects the fact that the self-reports referred mostly ~~refer~~ to events occurring prior to registration, while the bulk of the hospital data ~~is~~were collected after registration. Several much smaller studies have found similar concordances. Frezzato et al.

~~showed~~(49) demonstrated that the question, "Do you think you ever had venous thromboembolism?" had a sensitivity of 84% and a specificity of 88% compared ~~to~~with medical records ~~in~~among 267 Italian participants ~~(47)~~. Greenbaum et al. (50) found an 88.9% positive predictive value for PE and a 69.7% positive predictive value for DVT when comparing self-~~report~~reports with ~~surgeon assessment~~surgeons' assessments in a US cohort of 3,976 ~~post-surgery~~postsurgery patients ~~(48)~~. This leads us to conclude that there is no strong inherent reason to disregard the self-reported data on VTEs as less accurate than the medical reports.

There is also a considerable body of literature on potential sources of bias in externally validated data. One concern is informed presence bias, which is influenced by socioeconomic factors, such as ~~healthcare~~health-care costs ~~(49)~~(51)~~, levels of education~~, educational level, and distance ~~to travel to healthcare (50);~~from[SB5] health-care services (52). Perceptions about the ~~healthcare~~health-care system can ~~impact patient~~affect patients' willingness to self-report outcomes ~~(51)~~(53). Poor communication between patient and clinician could also be a factor in discordance, as more complicated conditions are both harder to diagnose (and thus underreported in medical records) and ~~hard to understand~~ harder for the patient to understand (and thus ~~mis-reported~~misreported or underreported in self-reported data). This might explain why patients are more likely to self-report DVTs, a more commonly understood illness than PE. These factors mean that ~~two~~2 patients with identical symptoms and underlying conditions may be represented[SB6] differently in different data sources.

**Biases ~~impacting~~affecting VTE reporting**

We found that changing the data source for defining a VTE outcome from hospital data to self-reported data altered the ~~socio-demographic~~sociodemographic characteristics of cases under consideration. There ~~is~~was also noticeable variation between VTE case

proportions by assessment ~~centre~~ center; this could ~~be~~have been due to underlying geographic variation[SB7] in NHS provision ~~(52)~~(54). Self-reported cases were younger and ~~majority~~mostly female, while hospital-defined cases were ~~majority~~mostly male. This is particularly salient given conflicting evidence on whether VTEs are more prevalent in male or female patients and the impact this perception might have on subsequent diagnoses.

Several different factors might explain why women are more likely to self-report a disease without an equivalent medical record. Previous investigations have suggested age as a potential reason for differences in case prevalence, and self-reported events are mostly captured prior to registration (2, 3, ~~19)~~20). Prevalent events are subject to survivorship bias, but it is unclear why this would induce a gendered difference. We observed a difference in the mean age of cases between self-reported and hospital data~~,~~; the absolute difference was 0.6 years. However, this difference is unlikely to explain such a large discrepancy in ~~gender~~sex rates. It is also possible that this discrepancy is a result of ~~gender~~sex bias in diagnosis. Diagnostic and treatment bias ~~on~~according to sociodemographic factors is well-documented for cardiovascular diseases. Worldwide, women are less likely to undergo a detailed risk factor assessment for cardiovascular disease even when ~~doctors~~physicians are presented with identical symptoms ~~(53, 54)~~(55, 56) and are more likely to be misdiagnosed ~~(55, 56)~~(57, 58) or to have their symptoms dismissed as psychogenic ~~(57)~~(59). Women with cardiovascular symptoms are less likely than men to be referred to a specialist ~~(58, 59)~~(60, 61), and to receive advanced diagnostics ~~(60)~~(62), coronary procedures ~~(61, 62)~~(63, 64), and appropriate drug treatment ((~~63–~~65~~)~~ 67). It is unclear to what extent this generalizes specifically to VTEs: ~~a~~One study of DVT events found more women than men were sent for a diagnostic workup for DVT, but the actual diagnosis of DVT was higher in men with more severe thrombotic events ~~(66)~~(68). However, women have poorer quality-of-life outcomes 1 year after diagnosis ~~(67)~~(69), worse bleeding outcomes, and more VTE mortality in long-

term follow-up[70]-up (70). Given this, the magnitude of the impact of gendersex bias on VTE reporting is uncertain.

## Strengths and weaknesses of study

The strengths of our study include the large sample size. The previouspreviously largest study comparing self-reported VTEs towith hospital reports usedwas the Women's Health Initiative study. This study was one-fifth the size of the UK Biobank study, and only considered outcomes in women (46). Otherwere considered (48). In other comparisons, study populations have either been much smaller (21, 47, 68)(22, 49, 71), or looked atinvestigators have evaluated a more general cardiovascular outcome ((69–72)–75). Our results are also strengthened by the robust data linkage between the self-reported and hospital data. We know that the same participants appearappeared in both datasets,data sets; thus, differences in prevalence arewere not due to the biases of different samples but due to how well the different sources capturecaptured case numbers. Use of the linked primary--care data allowed us to investigate whether primary-care diagnosis could explain why so many participants only self-report VTE events. A further strength is the comparison of DVT and PE events. AsBecause these events are caused by related biological mechanisms, differences in diagnosis and reporting patterns between the diseases will more strongly represent differences arising from social, behavioral, and clinical factors.

Weaknesses inof our study include the fact that the UK Biobank is not representative of the UK population. This non-representativenessnonrepresentativeness limits our ability to extend conclusions beyond this datasetdata set, and none of the figures given here should be used as accurate estimates of the prevalence of VTEsVTE in the UK population. Nevertheless, the UK Biobank has a large influence on health research and perceptions about medical conditions worldwide. As such, it is vital to identify potential biases that may be introduced in considering particular data sources within the UK Biobank.

Another weakness is that the choices made in defining our primary ~~-~~-care cohort may ~~introduce~~have introduced additional bias. The primary ~~-~~-care cohort appears to be reasonably representative of the UK Biobank population in most characteristics but is distributed differently geographically. We acknowledge both the reproduction of the original biases of the UK Biobank and the possible intensification of them.

It is unclear whether these patterns found for VTEs in the UK Biobank would be similar for other conditions. Studies have found that more widely recognized and easily diagnosed illnesses tend to have greater agreement between self-~~report~~reports and official records ~~(71)~~(74, ~~(72)~~ 75), that community ~~-~~-managed conditions are less likely to be reflected in hospital ~~records[78]~~records (73), and that more serious diseases have higher agreement between ~~sources[81].~~sources (76). However, there is not much consistency in how accuracy is reported between these studies, and it is difficult to ~~conclude~~form a conclusion about whether a specific disease will have a strong overlap between hospital records and self-report. We would expect, in line with the previous literature, that these differences will be less marked for more common and more well-known diseases, and for diseases ~~where~~for which there are empirical tests for diagnosis; this is reflected in our findings for DVT and PE.

**Recommendations**

For studying VTEs~~,~~ and in general, we recommend that researchers look at the reports coming from all the possible sources within the UK Biobank, how the reports overlap, and whether there are any clues in the medication or demographic data that may help them identify the most appropriate definition to use. ~~Using self~~Self-reported data ~~is~~are particularly useful for identifying cases before baseline, while hospital data will capture more incident events. ~~Including~~Inclusion of primary ~~-~~-care data may ~~create a lower~~result in an analysis with less powered analysis due to the smaller number of participants with available data, but it has the potential to capture events rarely seen in hospital, such as depression. While primary ~~-~~-care

data were not useful in validating self-reported events in the case of VTEs, there may be conditions where there is a much larger overlap between sources of report, in which case the self-reported data could be used as a proxy for the missing primary ~~-~~-care data in the full cohort. We would recommend that self-reported data be included for case definitions of VTE, either as sensitivity analysis alongside a more parsimonious main definition or as the primary analysis together with a sensitivity analysis that excludes the self-report data. This gives the researcher the greatest flexibility for understanding the impact this decision ~~has~~might have on their analysis.

In conclusion, there are large differences between the VTE case populations defined based on routinely collected hospital data and those defined based on self-reported data in the UK Biobank, ~~both~~ in terms of both the number of events reported and the ~~demographics~~demographic characteristics of the case populations. Such differences are likely to affect our perception of the typical VTE patient. As such, our findings suggest that in future studies, researchers need to ~~take~~ be aware of potential demographic differences underlying seemingly neutral event definitions in order to avoid entrenching further inequalities in ~~healthcare~~health care.



## ACKNOWLEDGMENTS

Author affiliations[BS8]: MRC Biostatistics Unit, University of Cambridge, Cambridge, ~~UK~~United Kingdom (Emily Bassett, James Broadbent, Stephen Burgess, Amy M. Mason); York University[BS9], Toronto, Ontario, Canada (James Broadbent); Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, ~~UK~~London, United Kingdom (Dipender Gill); Chief Scientific Advisor Office, Research and Early Development, Novo Nordisk, Copenhagen, Denmark (Dipender Gill); British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, ~~University~~School of Clinical Medicine, University of Cambridge, Cambridge,

~~UK~~United Kingdom (Stephen Burgess, Amy M. Mason~~));~~ and Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge ~~UK~~, United Kingdom (Stephen Burgess, Amy M. Mason).

This work was supported[BS10] by core funding from the~~:~~ British Heart Foundation (grant RG/18/13/33946~~),~~ and Chair Award CH/12/2/29428), the British Heart Foundation Cambridge ~~CRE (~~Centre for Cardiovascular Research Excellence (grant RE/18/1/34212), ~~British Heart Foundation Chair Award (CH/12/2/29428), United Kingdom~~the UK Research and Innovation Medical Research Council (grant MC_UU_00002/7), and the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (grants BRC-1215-20014~~;~~ and NIHR203312~~) (*).).~~ This work was also supported by Health Data Research UK, which is funded by the ~~UK~~ Medical Research Council, the Engineering and Physical Sciences Research Council, the Economic and Social Research Council, the Department of Health and Social Care (England), the Chief Scientist Office of the Scottish Government Health and Social Care Directorates, the Health and Social Care Research and Development Division (Welsh Government), the Public Health Agency (Northern Ireland), the British Heart Foundation ~~and Wellcome. A.M.M. is funded by the NIHR Blood and Transplant Research Unit (BTRU) in Donor Health and Behavior (NIHR203337)(*),~~ and ~~was funded by the EU/EFPIA~~the Wellcome Trust. S.B. was supported by the Wellcome Trust (grant 225790/Z/22/Z). A.M.M. was supported by the NIHR Blood and Transplant Research Unit in Donor Health and Behaviour (grant NIHR203337) and by the European Union/European Federation of Pharmaceutical Industries and Associations Innovative Medicines Initiative Joint Undertaking BigData@Heart (grant ~~(~~116074).

D.G. ~~is~~was supported by the British Heart Foundation Centre of Research Excellence at Imperial College London (grant RE/18/4/34215).

~~S.B. is supported by the Wellcome Trust (Grant No. 225790/Z/22/Z).~~

~~*The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.~~

~~Data Availability Statement:~~ This research was conducted using the UK Biobank Resource under ~~Application Number~~application 20480. Individual~~--~~level data from the UK Biobank cannot be shared publicly for ethical ~~&~~and privacy reasons. The data will be shared ~~on~~upon reasonable request to the corresponding author, with the permission of the UK Biobank.

~~Thanks: N/A.~~

~~Conference presentation: N/A.~~

~~Preprint Information: N/A.~~

~~Disclaimer: N/A.~~

The views expressed in this article are those of the author(s) and are not necessarily those of the NIHR or the UK Department of Health and Social Care.

Conflict of interest[BS11]: none declared.

**REFERENCES**


1. Oger E. Incidence of venous thromboembolism: a community-based study in ~~Western~~western France. EPI-GETBP Study Group. Groupe ~~d'Etude~~d'Etude de la Thrombose de Bretagne Occidentale. *Thromb Haemost*. 2000;83(5):657–660.

2. Heit JA. Epidemiology of venous thromboembolism. *Nat Rev Cardiol*. 2015;12(8):464–474.

3. Næss IA, Christiansen S, Romundstad P, et al. Incidence and mortality of venous thrombosis: a population-based study. *J Thromb Haemost*. 2007;5(4):692–699.

4. Allaert F-A, Benzenine E, Quantin C. Hospital incidence and annual rates of hospitalization for venous thromboembolic disease in France and the USA. *Phlebology*. 2017;32(7):443–447.

5. White RH. The epidemiology of venous thromboembolism. *Circulation*. 2003;107(23– suppl– 1):I4–I8.

6. Hagan KA, Harrington LB, Kim J, et al. Adiposity throughout the life course and risk of venous thromboembolism. *Thromb Res*. 2018;172:67–73.

7. Zöller B, Ji J, Sundquist J, et al. Body height and incident risk of venous thromboembolism: a cosibling design. *Circ Cardiovasc Genet*. 2017;10(5):e001651.

~~8.~~8. Cheng Y-J, Liu Z-H, Yao F-J, et al. Current and former smoking and risk for venous thromboembolism: a systematic review and meta-analysis. *PLoS Med*. 2013;10(9):e1001515.

9. Lind MM, Johansson M, Själander A, et al. Incidence and risk factors of venous thromboembolism in men and women. *Thromb Res*. 2022;214:82–86.

~~9.~~10. ~~H~~oward TA, Judd CS, Snowden GT, et al. ~~Clement ND~~ Incidence and risk factors associated with venous thromboembolism following primary total HIP arthroplasty in low-risk patients when using aspirin for prophylaxis. *Hip Int*. 2022;32(5):562–567.

~~10.~~11. Kort D, van Rein N, van der Meer F, et al. Relationship between neighborhood socioeconomic status and venous thromboembolism: results from a population-based study. *J Thromb Haemost*. 2017;15(12):2352–2360.

~~11.~~12. Horner D, Goodacre S, Pandor A, et al. Thromboprophylaxis in lower limb immobilisation after injury (TiLLI). *Emerg Med J*. 2020;37(1):36–41.

~~12.~~13. Tadesse TA, Kedir HM, Fentie AM, et al. Venous thromboembolism risk and thromboprophylaxis assessment in surgical patients based on ~~caprini~~Caprini risk assessment model. *Risk Manag Healthc Policy*. 2020;13:2545–2552.

~~13.~~14. Lutsey PL, Zakai NA. Epidemiology and prevention of venous thromboembolism. *Nat Rev Cardiol*. 2023;20(4):248–262.

~~14.~~15. Anderson FA, Spencer FA. Risk factors for venous thromboembolism. *Circulation*. 2003;107(23~~–~~ suppl~~–~~ 1):I9–I16.

~~15.~~16. Barco S, Nijkeuter M, Middeldorp S. Pregnancy and venous thromboembolism. *Semin Thromb Hemost*. 2013;39(5):549–558.

~~16.~~17. Silverstein MD, ~~Heit~~ JA, Mohr DN, ~~Petterson TM~~et al. ~~, O'Fallon WM, Melton LJ~~ Trends in the incidence of deep vein thrombosis and pulmonary embolism: a 25-year population-based study. *Arch Intern Med*. 1998;158(6):585–593.

~~17.~~18. Arshad N, Isaksen T, Hansen J-B, et al. Time trends in incidence rates of venous thromboembolism in a large cohort recruited from the general population. *Eur J Epidemiol*. 2017;32(4):299–305.

~~18.~~19. Alotaibi GS, Wu C, Senthilselvan A, et al. Secular trends in incidence and mortality of acute venous thromboembolism: the AB-VTE population-based study. *Am J Med*. 2016;129(8):879.e19–879.e25.

~~19.~~20. Arnesen CAL, Veres K, Horváth-Puhó E, et al. Estimated lifetime risk of venous thromboembolism in men and women in a Danish nationwide cohort: impact of competing risk of death. *Eur J Epidemiol*. 2022;37(2):195–203.

~~20.~~21. Melgaard L, Nielsen PB, Overvad TF, et al. ~~Larsen TB~~ Sex differences in risk of incident venous thromboembolism in heart failure patients. *Clin Res Cardiol*. 2019;108(1):101–109.

~~21.~~22. Huerta C, Johansson S, Wallander M-A, et al. ~~García Rodríguez LA~~ Risk factors and short-term mortality of venous thromboembolism diagnosed in the primary care setting in the United Kingdom. *Arch Intern Med*. 2007;167(9):935–943.

~~22.~~23. Cushman M, Glynn R, Goldhaber S, et al. Hormonal factors and risk of recurrent venous thrombosis: the prevention of recurrent venous thromboembolism trial. *J Thromb Haemost*. 2006;4(10):2199–2203.

~~23.~~24. Eichinger S, Heinze G, Jandeck LM, et al. Risk assessment of recurrence in patients with unprovoked deep vein thrombosis or pulmonary embolism: the Vienna prediction model. *Circulation*. 2010;121(14):1630–1636.

~~24.~~25. Bell EJ, Lutsey PL, Basu S, et al. Lifetime risk of venous thromboembolism in two cohort studies. *Am J Med*. 2016;129(3):339.e19–339.e26.

~~25.~~26. Goldhaber SZ. Risk factors for venous thromboembolism. *J Am Coll Cardiol*. 2010;56(1):1–7.

~~26.~~27. Neeman E, Liu V, Mishra P, et al. Trends and risk factors for venous thromboembolism among hospitalized medical patients. *JAMA Netw Open*. 2022;5(11):e2240373.

~~27.~~28. Anderson JJ, Ho FK, Niedzwiedz CL, et al. Remote history of VTE is associated with severe COVID-19 in middle and older age: UK Biobank cohort study. *J Thromb Haemost*. 2021;19(10):2533–2538.

~~28.~~29. Kolin DA, Kulm S, Elemento O. Prediction of primary venous thromboembolism based on clinical and genetic factors within the UK Biobank. *Sci Rep* 2021;11(1):21340.

~~29.~~30. Klarin D, Emdin CA, Natarajan P, et al. Genetic analysis of venous thromboembolism in UK Biobank identifies the ZFPM2 locus and implicates obesity as a causal risk factor. *Circ Cardiovasc Genet*. 2017;10(2):e001643.

~~30.~~31. Hinds DA, Buil A, Ziemek D, et al. Genome-wide association analysis of self-reported events in 6135 individuals and 252 827 controls identifies 8 loci associated with thrombosis. *Hum Mol Genet*. 2016;25(9):1867–1874.

~~31.~~32. Goldstein BA, Bhavsar NA, Phelan M, et al. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *Am J Epidemiol*. 2016;184(11):847–855.

~~32.~~33. Jeanselme V, Martin G, Peek N, et al. Deepjoint: robust survival modelling under clinical presence shift [preprint]. *arXiv ~~preprint.~~* 2022. (https://doi.org/10.48550/arXiv.2205.13481~~.~~). Accessed June 21, 2023.

~~33.~~34. National Institute for Health and Care ~~I~~Excellence ~~NIfHaC~~. ~~.~~ *~~NICE Guidelines No. 158: Venous Thromboembolic Diseases: Diagnosis, Management and Thrombophilia Testing~~*. (NICE guideline NG158). London, United Kingdom: National Institute for Health and Care Excellence;~~.~~ 2020. ~~NG158.~~ https://www.nice.org.uk/guidance/ng158. Accessed June, 21 2023.

~~34.~~35. Condliffe R. Pathways for outpatient management of venous thromboembolism in a UK centre. ~~*Thrombosis Journal.*~~*Thromb J.* 2016;14(1):1–8.

~~35.~~36. UK Biobank. *Verbal Interview Stage. Version 1.1.* https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/Interview.pdf. Published ~~.~~ 2012. ~~. https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/Interview.pdf~~ Accessed ~~21st~~ June 21, 2023.

~~36.~~37. UK Biobank. Data-~~Field~~field 20002. https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20002. Published ~~.~~ 2023. ~~. https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20002~~ Accessed ~~21st~~ June 21, 2023.

~~37.~~38. UK Biobank. *Mapping Inpatient Hospital Data Across England, Scotland and Wales. Version 1.5.* https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/inpatient_mapping.pdf. Published ~~.~~ 2019. ~~. https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/inpatient_mapping.pdf~~ Accessed ~~21st~~ June 21, 2023.

~~38.~~39. UK Biobank. *Mortality Data: Linkage to Death Registries. Version 2.0.* https://biobank.ctsu.ox.ac.uk/crystal/ukb/docs/DeathLinkage.pdf. Published ~~.~~ 2020. ~~. https://biobank.ctsu.ox.ac.uk/crystal/ukb/docs/DeathLinkage.pdf~~ Accessed ~~21st~~ June 21, 2023.

~~39.~~40. Office for National Statistics. ~~Guidance U.~~ *Guidance for Doctors Completing Medical Certificates of Cause of Death in England and Wales*. ~~Office for National Statistics~~London, United Kingdom: HM Passport Office; 2020.

~~40.~~41. UK Biobank. ~~*UK Biobank*~~ *Primary Care Linked Data. Version 1.0*. 2019. https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/primary_care_data.pdf . Accessed June ~~21st~~21, 2023.

~~41.~~42. Newell SA, Girgis A, Sanson-Fisher RW, et al. The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population: a critical review. *Am J Prev Med*. 1999;17(3):211–229.

~~42.~~43. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol*. 2017;186(9):1026–1034.

44. Mosca L, Linfante AH, Benjamin EJ, et al. National study of physician awareness and adherence to cardiovascular disease prevention guidelines. *Circulation*. 2005;111(4):499–510.

~~43.~~45. Hamberg K. Gender bias in medicine. *Womens Health*. 2008;4(3):237–243.

~~44.~~46. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447–453.

~~45.~~47. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. *J Am Med Inform Assoc*. 2018;25(8):1080–1088.

~~46.~~48. Heckbert SR, Kooperberg C, Safford MM, et al. Comparison of self-report, hospital discharge codes, and adjudication of cardiovascular events in the Women's ~~health initiative~~Health Initiative.. *Am J Epidemiol*. 2004;160(12):1152–1158.

~~47.~~49. Frezzato M, Tosetto A, Rodeghiero F. Validated questionnaire for the identification of previous personal or familial venous thromboembolism. *Am J Epidemiol*. 1996;143(12):1257–1265.

~~48.~~50. Greenbaum JN, Bornstein LJ, Lyman S, et al. The validity of self-report as a technique for measuring short-term complications after total hip arthroplasty in a joint replacement registry. *J Arthroplasty*. 2012;27(7):1310–1315.

~~49.~~51. Smith KT, Monti D, Mir N, et al. Access is necessary but not sufficient: factors influencing delay and avoidance of health care services. *MDM Policy ~~&~~ ~~Practice.~~Pract.* 2018;3(1):2381468318760298.

~~50.~~52. Barik D, Thorat A. Issues of unequal access to public health in India. *Front Public Health*. 2015;3:245.

~~51.~~53. Bower JK, Patel S, Rudy JE, et al. Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. *Curr Epidemiol Rep*. 2017;4(4):346–352.

~~52.~~54. Appleby J, Raleigh V, Frosini F, et al. *Variations in* ~~health care~~*Health Care: the Good, the Bad and the Inexplicable*. ~~. The Good, The Bad and The Inexplicable~~ London, United Kingdom: ~~:~~ The ~~*King's*~~King's Fund;~~. England.~~ 2011.

~~53.~~55. Crilly M, Bundred P, Hu X, et al. Gender differences in the clinical management of patients with angina pectoris: a cross-sectional survey in primary care. *BMC Health Serv Res*. 2007;7(1):1–9.

~~54.~~56. Bartys S, Baker D, Lewis P, et al. Inequity in recording of risk in a local population-based screening programme for cardiovascular disease~~. *Eur*~~. *Eur* ~~J~~ *Prev Cardiol*. 2005;12(1):63–67.

55.57. Wu J, Gale CP, Hall M, et al. Editor's choice—impact of initial hospital diagnosis on mortality for acute myocardial infarction: a national cohort study. *Eur Heart J Acute Cardiovasc Care*. 2018;7(2):139–148.

56.58. Arslanian-Engoren C. Gender and age differences in ~~nurses'~~nurses' triage decisions using vignette patients. *Nurs Res*. 2001;50(1):61–66.

57.59. Chiaramonte GR, Friend R. Medical ~~students'~~students' and ~~residents'~~residents' gender bias in the diagnosis, treatment, and interpretation of coronary heart disease symptoms. *Health Psychol*. 2006;25(3):255–266.

58.60. Clerc Liaudat C, Vaucher P, De Francesco T, et al. Sex/gender bias in the management of chest pain in ambulatory care. *Womens Health (Lond)*. 2018;14:~~1745506518805641~~1745506518805641.

59.61. Bach DS, Radeva JI, Birnbaum HG, et al. Prevalence, referral patterns, testing, and surgery in aortic valve disease: leaving women and elderly patients behind? *J Heart Valve Dis*. 2007;16(4):362–369.

60.62. Chang AM, Mumma B, Sease KL, et al. Gender bias in cardiovascular testing persists after adjustment for presenting characteristics and cardiac risk. *Acad Emerg Med*. 2007;14(7):599–605.

61.63. Fogg AJ, Welsh J, Banks E, et al. Variation in cardiovascular disease care: an Australian cohort study on sex differences in receipt of coronary procedures. *BMJ Open*. 2019;9(7):e026507.

62.64. Shah AS, Griffiths M, Lee KK, et al. High sensitivity cardiac troponin and the under-diagnosis of myocardial infarction in women: prospective cohort study. *BMJ*. 2015;350:h626.

~~63.~~65. Murphy N, Simpson CR, McAlister F, et al. National survey of the prevalence, incidence, primary care burden, and treatment of heart failure in Scotland. *Heart*. 2004;90(10):1129–1136.

~~64.~~66. Williams D, Bennett K, Feely J. Evidence for an age and gender bias in the secondary prevention of ischaemic heart disease in primary care. *Br J Clin Pharmacol*. 2003;55(6):604–608.

~~65.~~67. Jarvie JL, Foody JM. Recognizing and improving health care disparities in the prevention of cardiovascular disease in women. *Curr Cardiol Rep*. 2010;12(6):488–496.

~~66.~~68. Bauersachs RM, Riess H, Hach-Wunderle V, et al. Impact of gender on the clinical presentation and diagnosis of deep-vein thrombosis. *Thromb Haemost*. 2010;103(04):710–717.

~~67.~~69. Giustozzi M, Valerio L, Agnelli G, et al. Sex-specific differences in the presentation, clinical course, and quality of life of patients with acute venous thromboembolism according to baseline risk factors. Insights from the PREFER in VTE. *Eur J Intern Med*. 2021;88:43–51.

~~68.~~70. Chan SM, Brahmandam A, Valcarce-Aspegren M, et al. Sex differences in long-term outcomes of patients with deep vein thrombosis. *Vascular*. 2023;31(5):994–1002.

71. Fritz BA, Escallier KE, Ben Abdallah A, et al. Convergent validity of three methods for measuring postoperative complications. *Anesthesiology*. 2016;124(6):1265–1276.

69.72. Stoye G, Zaranko B. *How Accurate Are Self-Reported Diagnoses? Comparing Self-Reported Health Events in the English Longitudinal Study of Ageing with Administrative Hospital Records*. 2020London, United Kingdom.: Institute for Fiscal Studies; 2020.

70.73. Saunders CL, Gkousis E. *Impact of telephone triageTelephone Triage on accessAccess to primary carePrimary Care for people living with multiple long-term health conditions: rapid evaluationPeople Living With Multiple Long-Term Health Conditions: Rapid Evaluation*. Southampton, United Kingdom:. National Institute for Health and Care Research: Southampton (UK).; 2022.

71.74. Ho PJ, Tan CS, Shawon SR, et al. Comparison of self-reported and register-based hospital medical data on comorbidities in women. *Sci Rep*. 2019;9(1):1–9.

72.75. Haapanen N, Miilunpalo S, Pasanen M, et al. Agreement between questionnaire data and medical records of chronic diseases in middle-aged and elderly Finnish men and women. *Am J Epidemiol*. 1997;145(8):762–769.

73. Cheng Y-J, Liu Z-H, Yao F-J, et al. Current and former smoking and risk for venous thromboembolism: a systematic review and meta-analysis. *PLoS Med*. 2013;10(8):e1001515.

74. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am J Epidemiol*. 2017;186(9):1026–1034.

75. Chan SM, Brahmandam A, Valcarce-Aspegren M, et al. Sex differences in long-term outcomes of patients with deep vein thrombosis. *Vascular*. 2022;31(5):994–1002.

76. Bergmann MM, Byers T, Freedman DS, et al. Validity of self-reported diagnoses leading to hospitalization: a comparison of self-reports with hospital records in a prospective study of American adults. *Am J Epidemiol*. 1998;147(10):969–977.

**Figure 1.** ~~Venn diagram of the proportional~~Proportional[SB12] overlap in ~~VTE~~ venous thromboembolism cases between different data sources in all UK Biobank participants~~.~~ ~~These~~, 2006–2010. In Web Figure 1, these are expanded further into events occurring prior to and ~~post~~after registration in the UK Biobank ~~in web Figure 1~~.

**Figure 2.** Proportions of deep vein thrombosis (DVT) ~~to~~and pulmonary embolism (PE) ~~in~~cases among identified venous thromboembolism ~~(VTE)~~ cases when ~~ascertaining~~ cases were ascertained from different sources, UK Biobank, 2006–2010.

**Figure 3.** ~~Venn diagram of the proportional~~Proportional overlap in ~~VTE~~venous thromboembolism cases between different data sources in ~~the~~ UK Biobank participants with primary ~~-~~care data~~. These~~, 2006–2010. In Web Figures 4 and 5, these are expanded further into events occurring prior to and ~~post~~after registration in the UK Biobank ~~in web figures 4 & 5~~.



[Tables follow]

**Table 1.** ~~Demographics~~Demographic[SB13] Characteristics of UK Biobank Participants at Recruitment, 2006–2010

| Characteristic | All Participants (*n* = 502,520) | | Primary Care Cohort (*n* = 177,358) | |
|---|---|---|---|---|
| | % | Mean (SD) | % | Mean (SD) |
| Female sex | 54.4 | | 54.5 | |
| Age ~~in,~~ years | | 57.0 (8.1) | | 57.2 (8.0) |
| White ~~Ethnicity~~race/ethnicity | 94.6 | | 95.7 | |
| Assessment center | | | | |
| Wales ~~Assessment Center~~ | 4.1 | | 10.3 | |
| Scotland ~~Assessment Centre~~ | 7.1 | | 12.6 | |
| London ~~Assessment Centre~~ | 13.7 | | 6.8 | |
| Unemployment | 43.1 | | 44.1 | |
| Manual work[a] | 7.7 | | 7.9 | |
| Higher ~~Education~~education | 60.2 | | 59.8 | |
| Townsend deprivation index | | −1.29 (3.1) | | −1.4 (3.0) |
| ~~Own~~Owning one's house outright | 51.5 | | 52.9 | |
| ~~Two or more~~≥2 cars in household | 48.8 | | 48.3 | |
| Shift ~~working~~ work[b] | 5.6 | | 5.6 | |
| Body mass index[c] | | | | |
| ~~Mean female BMI~~[e] Women | | 27.1 (5.2) | | 27.2 (5.2) |
| ~~Mean male BMI~~ Men | | 27.8 (4.2) | | 27.9 (4.3) |
| Current smoking | | | | |
| ~~Current smokers (~~Women~~)~~ | 8.9 | | 8.7 | |
| ~~Current smokers (men)~~ Men | 12.5 | | 12.1 | |

~~Abbreviations:???~~Abbreviation: SD, standard deviation.

[a] Answered "~~Usually~~usually" or "~~Always~~always" to the question, "Does your work involve heavy manual or physical work~~?" ?~~".

[b] Answered "~~Usually~~usually" or "~~Always~~always" to the question, "Does your work involve shift work~~?"~~?".

[c] Weight (kg)/height $(m)^2$.

**Table 2.** Demographic Comparison ~~between~~Between Case Populations Defined ~~Via the~~via Different Data Sources, UK Biobank, 2006–2010[a]

| Variable | Data Source | | | | | | *P* for Difference Between Self-Reported and Hospital Data |
|---|---|---|---|---|---|---|---|
| ~~Data Source~~ | Hospital Data (*n* = 9,272) | | Primary Care Data (*n* = 1,441) | | Self-Reported (*n* = 13,727) | | ~~*P* value for Difference Between Self-Reported and Hospital Data~~ |
| | % | Mean (SD) | % | Mean (SD) | % | Mean (SD) | |
| Female sex | 45.7 | | 54.0 | | 58.3 | | <0.0001 |
| Age ~~in,~~ years | | 60.3 (7.2) | | 60.1 (7.2) | | 59.7 (7.4) | <0.0001 |
| White race/ethnicity | 96.5 | | 97.3 | | 96.0 | | 0.051 |
| Assessment center | | | | | | | |
| Wales ~~Assessment Centre~~ | 3.9 | | 14.1 | | 4.9 | | 0.0003 |
| Scotland ~~Assessment Centre~~ | 7.0 | | 11.7 | | 5.8 | | 0.0002 |
| London ~~Assessment Centre~~ | 11.4 | | 5.3 | | 11.7 | | 0.486 |
| Not working (retired or unemployed) | 60.4 | | 61.1 | | 59.9 | | 0.448 |
| Higher ~~Education~~education | 52.7 | | 52.4 | | 53 | | 0.655 |
| Current ~~smokers~~smoker | 12.2 | | 12.3 | | 12.4 | | 0.652 |
| Work history | | | | | | | |
| Heavy manual work ~~history~~[b] | 6.5 | | 6.4 | | 5.9 | | 0.063 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| __Shift work ~~history~~[c] | 4.6 | | 4.4 | | 4.5 | | 0.721 |
| Townsend deprivation index | | −0.89 (3.31) | | −1.07 (3.17) | | −0.93 (3.27) | 0.365 |
| ~~Own~~Owning one's house outright | 56.8 | | 57.5 | | 55.0 | | 0.007 |
| ~~More than~~ ≥1 car in household | 40.7 | | 41.2 | | 42.4 | | 0.010 |
| ~~BMI~~Body mass index[d] | | 29.4 (5.6) | | 29.2 (5.6) | | 29.2 (5.7) | 0.009 |

~~Abbreviations:???~~Abbreviation: SD, standard deviation.

[a] A larger version of this table can be seen as Web Table 11. *P* values are reported for independent sample *t*~~ ~~tests for continuous variables and $\chi^2$ tests of proportions for binary variables.

[b] Answered "~~Usually~~usually" or "~~Always~~always" to the question, "Does your work involve heavy manual or physical work~~?" ~~?".

[c] Answered "~~Usually~~usually" or "~~Always~~always" to the question, "Does your work involve shift work~~?"~~?".

[d] Weight (kg)/height $(m)^2$.