Supplemental Material

Data source

CPRD is based on information documented in the primary care setting and includes demographics, medical history, symptoms and diagnoses recorded with Read medical codes, unstructured medical notes, letters to and from secondary care and prescriptions issued by the general practitioner (GP). HES include dates of hospital admission and discharge, primary and other main reasons for treatment recorded with ICD-10 codes, and surgical operations and procedures performed during hospital stay recorded with OPCS-4 codes. ONS data contain the date and cause of death recorded as ICD-10 in death certificates.

CTEPH algorithm

The CTEPH ascertainment algorithm included information from three sources: (1) hospital-based information, (2) GP-based information and (3) anonymized clinical notes:
(1) Hospital-based information consisted of discharge diagnoses of PH (coded with 4-digit ICD-10 code: I27.0) and specific treatment for pulmonary arterial hypertension recorded with OPCS-version 4.6 codes, i.e. (i) administration of specific drugs: X82.1 (ambrisentan, sildenafil, tadalafil, vardenafil), X82.2 (bosentan, sitaxentan), X82.3 (iloprost) and X82.4 (epoprostenol), (ii) lung transplant and (iii) a pulmonary endarterectomy, (2) GP-based information consisted of GP-based diagnoses of arterial PH and PH-specific treatment, i.e. (i) prostacyclins (epoprostenol and iloprost), (ii) endothelin receptor antagonists (ambrisentan, sitaxsentan and bosentan), (iii) phosphodiesterase-5 inhibitors (sildenafil, tadalafil and vardenafil) in the absence of a recording for erectile dysfunction, or (iv) a lung transplantation or a pulmonary endarterectomy, and (3) anonymized clinical notes. Clinical notes were obtained by an electronic search of all medical notes of patients in the VTE cohort for any word strings containing "LUNG" or "PULM" followed by "HYPERTE" within 15/16 characters, "CHRONIC THROMB" followed by "PH" within 11 characters, "CPH", "CTPH" or "CTEPH", or “PAP” for estimated pulmonary artery pressure (PAP) measurements. Systolic PAP values were estimated from the mean PAP values (mean PAP = 0.61 x systolic PAP + 2). Values without information on systolic or mean PAP where considered to be systolic values. Mean PAP values > 100 mmHG were excluded.
The CTEPH ascertainment was based on evidence of pulmonary arterial hypertension recorded at least 90 days after the initial DVT or PE. Potential cases of chronic pulmonary arterial hypertension were selected if there was a (i) hospital code for PH, (ii) a mean pulmonary arterial pressure of $\geq$ 25 mmHg identified in the anonymized electronic medical notes, or (iii) a diagnosis of PH recorded with a Read medical code or mentioned in the medical notes but in association with PH-specific treatment.

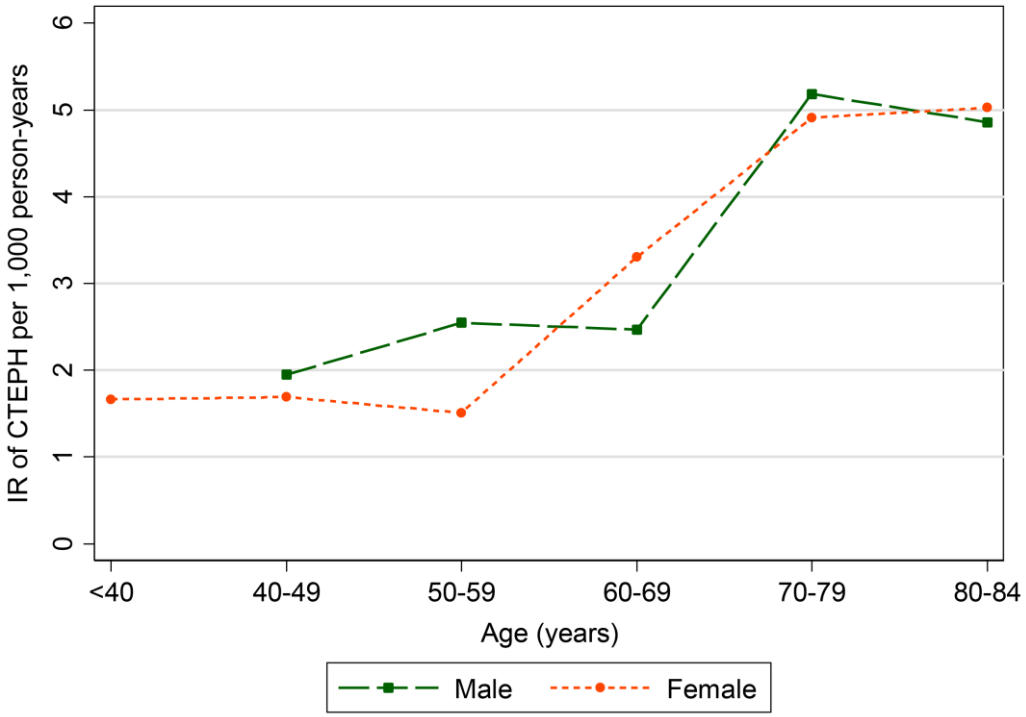
Sensitivity and specificity of the CTEPH ascertainment algorithm

We estimated sensitivity and specificity of the CTEPH ascertainment algorithm (without the manual review step) using clinical notes as the gold standard to determine if a patient had CTEPH. Therefore, we restricted the pool to patients with VTE, at risk for CTEPH and to those with potential recordings of clinical notes, and split the pool into two groups, i.e. those with CTEPH according to our algorithm (positives) and those without CTEPH (negatives). The

result of the manual review of the positives, as described in the manuscript, identified the "true positives" and the "false positives". The negatives that had at least one clinical note for PH were also manually review with the same approach as the positives. Negatives, where a CTEPH was located after manual review, formed the "false negatives" and all other negatives formed the group of "true negatives".

The sensitivity of our CTEPH ascertainment algorithm was estimated as number of "true positives" divided by the sum of number of "true positives" and "false negatives" and the specificity was estimated as number of "true negatives" divided by the sum of number of "true negatives" and "false positives".

Supplementary Figure 1: Incidence rate of CTEPH after acute VTE stratified by age and gender



CTEPH: Chronic thromboembolic pulmonary hypertension; IR: Incidence rate; VTE: Venous thromboembolism.

Incidence rates for ages <40 in males not presented as based on < 5 cases of CTEPH.

Supplementary Table 1: Risk factors for CTEPH - Sensitivity analysis using controls matched on year of birth and gender

Risk factor	Cases n (%)	Controls n (%)	Crude OR (95% CI)	Adjusted OR (95% CI)
Total	296	2918		
First VTE				
DVT	73 (24.7)	1484 (50.9)	1	1
PE ± DVT	223 (75.3)	1434 (49.1)	3.17 (2.41 - 4.18)	2.95 (2.18 - 4.00)
Recurrent VTE				
None	222 (75.0)	2613 (89.5)	1	1
DVT	21 (7.1)	147 (5.0)	1.77 (1.09 - 2.88)	2.79 (1.63 - 4.78)
PE ± DVT	53 (17.9)	158 (5.4)	4.13 (2.90 - 5.88)	4.57 (3.05 - 6.86)
Other comorbidities				
Inflammatory bowel disease	8 (2.7)	67 (2.3)	1.20 (0.57 - 2.51)	0.98 (0.44 - 2.22)
Systemic lupus erythematosus	5 (1.7)	14 (0.5)	3.57 (1.29 - 9.92)	4.69 (1.46 - 15.09)
Systemic sclerosis	<5	<5	-	-
Asthma	57 (19.3)	570 (19.5)	0.97 (0.71 - 1.31)	0.58 (0.40 - 0.84)
COPD	87 (29.4)	347 (11.9)	3.33 (2.50 - 4.43)	3.12 (2.15 - 4.51)
Atrial fibrillation	118 (39.9)	489 (16.8)	3.74 (2.85 - 4.91)	2.71 (1.99 - 3.70)
Heart failure	102 (34.5)	311 (10.7)	4.81 (3.64 - 6.37)	2.95 (2.13 - 4.08)
Hypothyroidism	34 (11.5)	276 (9.5)	1.27 (0.86 - 1.88)	1.01 (0.65 - 1.58)
Splenectomy	6 (2.0)	13 (0.4)	4.62 (1.73 - 12.39)	3.05 (1.02 - 9.09)
Cancer	33 (11.1)	578 (19.8)	0.50 (0.35 - 0.73)	0.46 (0.30 - 0.69)
Diabetes	57 (19.3)	404 (13.8)	1.50 (1.10 - 2.05)	1.30 (0.90 - 1.89)

CI: Confidence interval; COPD: Chronic obstructive pulmonary disease; DVT: Deep vein thrombosis; OR: Odds ratio; PE: Pulmonary embolism; VTE: Venous thromboembolism.

[1] Derived from multivariate conditional logistic regression model adjusting for all presented covariates, and body mass index, smoking status and source of first VTE diagnosis; [2]: At index day; [3]: Any history before index day.