



A cautionary note on the naive use of general-population biobanks to study pulmonary arterial hypertension, with a focus on Mendelian randomisation

Copyright ©The authors 2025

This version is distributed under the terms of the Creative Commons Attribution Licence 4.0.

Received: 6 March 2025

Accepted: 29 Aug 2025

To the Editor:

Pulmonary hypertension (PH) is defined by a mean pulmonary artery pressure >20 mmHg [1]. Patients with PH are assigned to one of five internationally recognised groups. Pulmonary arterial hypertension (PAH), or group 1 PH, is a heterogeneous collection of conditions characterised by increased precapillary pulmonary vascular resistance. Groups 2 to 5 PH comprise PH caused, in turn, by left heart disease, lung diseases (e.g. COPD), chronic thromboembolism, and miscellaneous causes such as haematological diseases.

PAH is increasingly being investigated using data from biobanks with a general population sampling frame (figure 1a), as opposed to disease-specific cohorts. General population biobanks tend to define PAH with a single medical record code. Many studies use these data to perform Mendelian randomisation (MR; figure 1b). MR is a study design that uses genetic variants specifically associated with an exposure of interest to test causal claims [2, 3]. We demonstrate two issues with existing general population biobank PAH data: low power and non-random misclassification. These result in a failure to replicate findings from gold standard PAH datasets in general population biobanks, and spurious findings from general population biobanks that fail to replicate in gold standard datasets.

Because PAH is rare, with a prevalence below 50 cases per million [1], population-based biobanks have few cases. Power in case-control studies is not substantially improved by increasing the case-to-control ratio beyond 1:4 [4]. Consequently, biobanks are typically less well powered than PAH-specific cohorts, despite having tens or hundreds of times more participants.

Low power has two implications: true-positive associations are more likely to be missed, and detected associations have lower odds of being true. Indeed, a reported association between variants proxying *IL6R* signalling and PAH risk observed in an early release of FinnGen failed to replicate in much larger cohorts [5].

Meta-analyses can address low power. Researchers therefore meta-analysed three general-population biobanks (UK Biobank (UKB), FinnGen release 12, and Million Veteran Program (MVP)) with 3302 apparent PAH cases and 1 205 457 controls (https://mvp-ukbb.finnngen.fi/pheno/19_HYPTEPSPUL) [6–8]. The prevalence in this meta-analysis is much higher than expected based on general population surveys. This might imply the presence of misclassification.

Misclassification (when people categorised as cases do not have the condition, or people categorised as controls do) reduces power [9]. The RHODES *et al.* [10] genome-wide association study (GWAS) contained 2085 PAH cases with gold standard expert centre diagnoses, and 9659 controls. The FinnGen-UKB-MVP meta-analysis is theoretically better powered because it has more “cases” and controls. Failure to detect an association observed in RHODES *et al.* [10] in the FinnGen-UKB-MVP meta-analysis would suggest misclassification of cases in the general population biobanks.

RHODES *et al.* [10] detected associations with three independent variants at genome-wide significance ($p < 5 \times 10^{-8}$): rs2856830 in the *HLA-DPA1/DPB1* gene cluster, and rs13266183 and rs10103692 near the



Shareable abstract (@ERSpublications)

There are two issues with using existing data from general-population biobanks to study PAH: low power and non-random misclassification. Relative to gold standard data, this results in false-positive and false-negative findings. <https://bit.ly/4nksm6c>

Cite this article as: Woolf B, De Bie E, McLaughlin V, *et al.* A cautionary note on the naive use of general-population biobanks to study pulmonary arterial hypertension, with a focus on Mendelian randomisation. *Eur Respir J* 2025; 66: 2500436 [DOI: 10.1183/13993003.00436-2025].

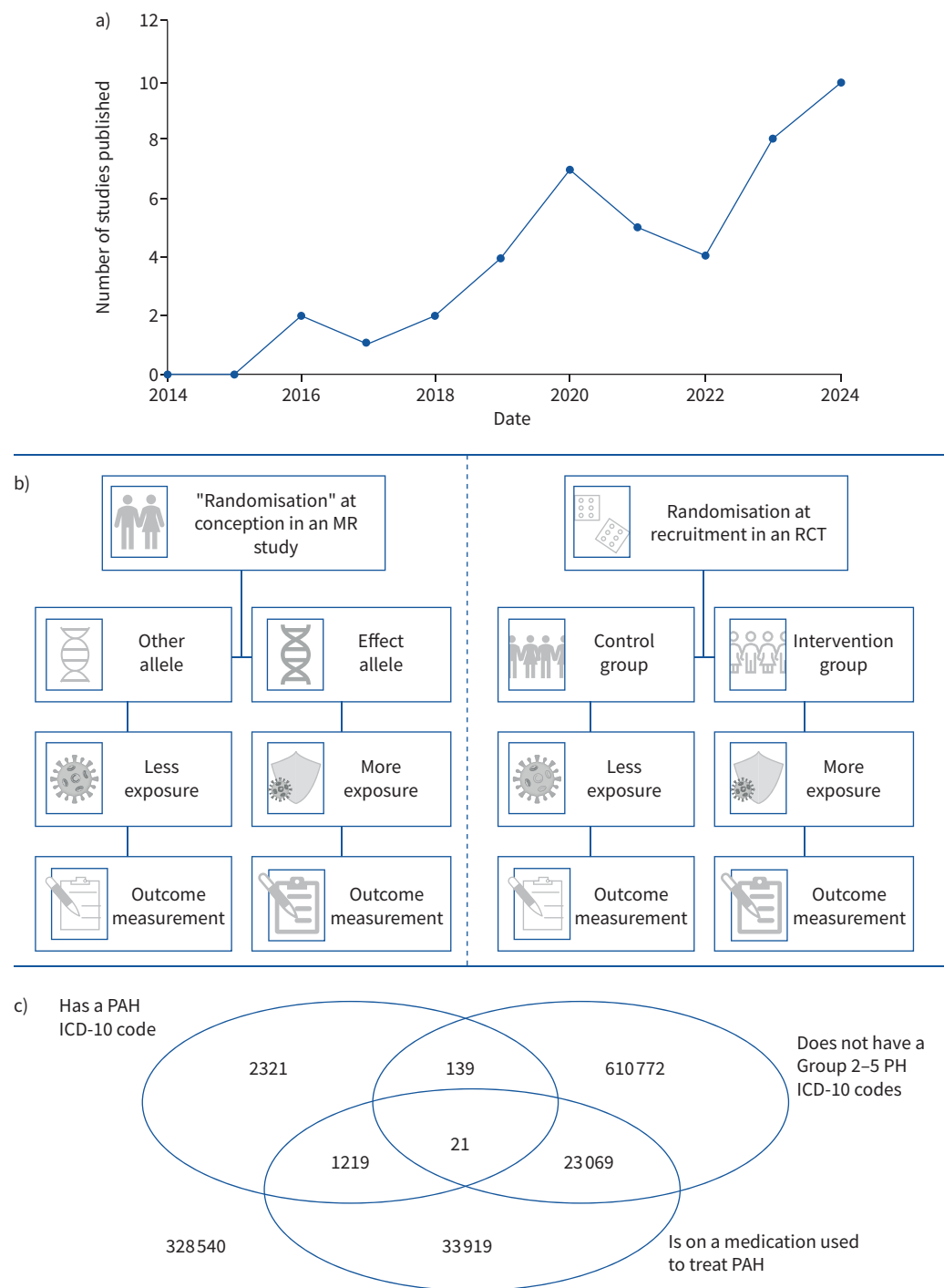


FIGURE 1 a) Results of a search for: “pulmonary arterial hypertension” AND (“biobank” OR “UK Biobank” OR “UKB” OR “UKBB” OR “FinnGen”) between 1 January 2014 and 31 December 2024 in PubMed. b) Illustration of the analogy leveraged in Mendelian randomisation (MR) investigations between genetic inheritance and randomisation in a randomised controlled trial (RCT): in an RCT, participants are randomised to receive an intervention which changes the exposure or a placebo and then followed up over time. Analogously, in a MR investigation, participants randomly inherit at conception an effect allele which impacts the exposure of interest, or a non-effect allele which does not. We would encourage readers interested in learning more about MR to read the primers cited [2, 3]. c) Venn diagram with the frequency per million of the pulmonary arterial hypertension (PAH) ICD-10 code I27.0, prescriptions of medication used to treat PAH (see below; sildenafil, tadalafil, selexipag, ambrisentan, bosentan, macitentan, epoprostenol, iloprost, treprostinil and sotatercept), and not having ICD-10 codes related to groups 2 to 5 pulmonary hypertension (PH; C96, D18.1, D57, D69.0,

D73, D86, E01, E02, E03, E04, E05, E06, E07, E74.0, E75.2, G47.3, I11, I26, I27.8, I34, I35, I50, J40, J41, J42, J43, J80, J81, J82, J84, J98.5, L95, L94, Q85.0, R06.8, T70, W94 and Y84.1) among individuals with electronic health record data in the All of Us biobank (version 7). The category “on a medication to treat PAH” includes individuals taking at least one of the listed drugs. While calcium channel blockers are used to treat some PAH patients, they are rarely used as a monotherapy treatment and are highly non-specific to PAH, so were not included in this list. Some of the other listed medications are also used for other conditions (for example, sildenafil for erectile dysfunction), and so many of the individuals listed as “on a medication to treat PAH” may not have PAH.

SOX17 gene. The *SOX17* gene was then followed up in cell and animal models, which demonstrated the functional relevance of the variants. However, none of these variants are associated with PAH in the FinnGen-UKB-MVP meta-analysis ($p=0.506$ for rs2856830, $p=0.241$ for rs13266183, and $p=0.318$ for rs10103692). This strongly suggests that the general population biobanks do not accurately identify PAH cases.

Although not ideal, when misclassification is random, information can still be gained from large numbers of noisy observations [9]. However, non-random misclassification can introduce bias. Rare conditions like PAH are susceptible to misclassification bias in general population biobanks because a tiny percentage of controls that are non-randomly misclassified can overwhelm signal from the limited number of true cases. (Correctly classifying 99.99% of non-PAH individuals in a general population sample will result in 100 false cases per million, twice the prevalence of true cases.)

One possible source of non-random misclassification is individuals with non-group 1 PH who incorrectly receive a PAH medical record code. We examined this by testing if causes of groups 2 to 5 PH associate with PAH diagnosis in MR analyses. We used GWASs on pulmonary emboli (32 876 cases, 1 508 902 controls) and left heart failure (10 857 cases, 1 463 784 controls) from the FinnGen-UKB-MVP meta-analysis, and COPD (58 559 cases, 937 358 controls) from the Global Biobank Meta-Analysis [11]. Cases were defined using medical record codes. We selected genetic proxies for each phenotype using independent (clumping $r^2=0.001$ and distance=10 Mb) genome-wide significant variants, and meta-analysed MR Wald ratios using an inverse variance weighting. The MR analysis using the FinnGen-UKB-MVP meta-analysis found evidence supporting genetically predicted left heart failure (OR 1.916, 95% CI 1.660 to 2.210), COPD (OR 1.376, 95% CI 1.092 to 1.734) and pulmonary emboli (OR 1.110, 95% CI 1.016 to 1.212) as risk factors for PAH diagnoses. This was not replicated by RHODES *et al.* [10] (OR 1.209, 95% CI 0.967 to 1.512; OR 1.144, 95% CI 0.772 to 1.697; and OR 0.997, 95% CI 0.874 to 1.137, respectively).

These differences can be explained by case misclassification. Alternative explanations are not convincing. Failure to replicate may be due to case misclassification. An alternative explanation is lower power; however, this would only lead to wider confidence intervals, whereas we also see substantial attenuation in estimates. While there are demographic differences between RHODES *et al.* [10] and the FinnGen-UKB-MVP meta-analysis, the differences between the biobanks are at least as large. Thus, if demographic differences are important, there should also be heterogeneity in estimates between the biobanks. However, substantial heterogeneity in PAH associations was not observed (minimum heterogeneity $p_{FDR}=0.089$).

Non-random misclassification can bias downstream analyses. To illustrate, atrial fibrillation (AF) is associated with left heart failure but does not cause PAH. Misclassification of group 2 PH might create a false-positive AF–PAH association. Selecting variants (using the same parameters described above) from the FinnGen-UKB-MVP meta-analysis of AF (170 643 cases and 1 163 021 controls), we observe an association of genetically predicted AF with PAH in the FinnGen-UKB-MVP PAH meta-analysis ($p<0.001$) but not in RHODES *et al.* [10] ($p=0.457$).

Non-PH individuals can also be misclassified with PAH, for example due to a related cardiorespiratory disease. In the All of Us biobank there are 21 individuals per million who had a PAH ICD-10 code, were on a medication used to treat PAH, and had no group 2 to 5 related ICD-10 codes [12]. However, 129 individuals per million have a PAH ICD-10 code but are not on any PAH medication and do not have a group 2 to 5 PH medical record code (figure 1c); these likely largely represent non-PH individuals with an incorrect PAH diagnosis.

Researchers wishing to use general population biobanks to study rare diseases such as PAH need to ensure they carefully address misclassification. Because PAH cases are unlikely to be unmedicated, the literature

suggests supplementing medical records with PAH-related prescriptions and/or requiring elevated pulmonary pressures [13]. However, since non-group 1 PH individuals are prescribed PAH medication, requiring cases to use PAH medication may not address misclassification between PH groups. The blanket exclusion of people with conditions related to group 2 to 5 PH will also exclude true PAH cases due to high co-occurrence of some of these conditions in gold standard-diagnosed PAH patients [14].

To advance the field, a call-to-action was initiated by the Genetics and Genomics Task Force 3 at the 7th World Symposium on Pulmonary Hypertension to assemble a global, diverse, condition-specific genetic registry [15]. We would encourage researchers or clinicians with genotyped PH patients to support the registry.

To conclude, general population biobanks defining PAH with a single medical record code are likely to produce unreliable results due to non-random misclassification and low power. When misclassification is truly random, it can be addressed by increasing sample sizes. Non-random misclassification cannot be addressed so simply. The spurious MR findings presented here highlight that even “robust” designs require good quality data. Inaccurate PAH definitions are thus likely to bias conventional observational analyses. Currently available general population biobank GWASs of PAH should therefore be avoided in downstream analyses such as MR.

Benjamin Woolf ^{1,2,3}, **Eckart De Bie** ^{4,5}, **Vallerie McLaughlin**⁶, **Stefan Gräf** ⁴, **Mark Toshner** ^{4,5}, **Martin R. Wilkins** ⁷, **Christopher J. Rhodes**⁷ and **Stephen Burgess**^{3,8}

¹The MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. ²School of Psychological Science, University of Bristol, Bristol, UK. ³The MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. ⁴Victor Phillip Dahdaleh Heart and Lung Research Institute, Department of Medicine, University of Cambridge, Cambridge, UK. ⁵Royal Papworth Hospital NHS Foundation Trust, Cambridge, UK. ⁶Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. ⁷National Heart and Lung Institute, Imperial College London, London, UK. ⁸British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.

Corresponding author: Benjamin Woolf (benjamin.woolf@bristol.ac.uk)

Acknowledgements: We want to acknowledge the participants and investigators of the FinnGen study and other participating cohorts in the GWASs used here. Scientists and clinicians wishing to contribute data to the PH genetic registry can register their interest at <https://bit.ly/BuildingFuturePHCohorts>, which will be followed-up within the International Consortium for Genetic Studies in Pulmonary (Arterial) Hypertension (P(A)H-ICON, <https://pahicon.com>). We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study.

Data availability: GWAS data from the publication by C.J. Rhodes and co-workers are available from <https://www.ebi.ac.uk/gwas/publications/30527956>, Global Biobank Meta-analysis summary statistics are available from <https://www.globalbiobankmeta.org/>, and FinnGen meta-analysis summary statistics are available from https://www.finnngen.fi/en/access_results. The R code used in this study is available from <https://github.com/bar-woolf/applied-MR-code/blob/main/PAH%20biobank%20research%20letter%20code%20shareable.R>.

Author contributions: B. Woolf implemented the analysis. B. Woolf and E. De Bie wrote the first draft report. All authors contributed to the conceptualisation and writing the article.

Conflicts of interests: The authors have no potential conflicts of interest to disclose.

Support statement: B. Woolf is funded by an Economic and Social Research Council (ESRC) South West Doctoral Training Partnership (SWDTP) 1+3 PhD Studentship Award (ES/P000630/1) and the Wellcome Trust (225790/Z/22/Z). The research was supported by the UK Research and Innovation Medical Research Council (MC_UU_00032/7 and MC_UU_00040/1). Funding information for this article has been deposited with the Open Funder Registry.

References

- 1 Kovacs G, Bartolome S, Denton CP, *et al.* Definition, classification and diagnosis of pulmonary hypertension. *Eur Respir J* 2024; 64: 2401324.

- 2 Sanderson E, Glymour MM, Holmes MV, *et al.* Mendelian randomization. *Nat Rev Methods Primers* 2022; 2: 6.
- 3 Davies NM, Holmes MV, Smith GD. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* 2018; 362: k601.
- 4 Katki HA, Berndt SI, Machiela MJ, *et al.* Increase in power by obtaining 10 or more controls per case when type-1 error is small in large-scale association studies. *BMC Medical Research Methodology* 2023; 23: 153.
- 5 Woolf B, Perry JA, Hong CC, *et al.* Multi-biobank summary data Mendelian randomisation does not support a causal effect of IL-6 signalling on risk of pulmonary arterial hypertension. *Eur Respir J* 2024; 63: 2302031.
- 6 Kurki MI, Karjalainen J, Palta P, *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* 2023; 613: 508–518.
- 7 Verma A, Huffman JE, Rodriguez A, *et al.* Diversity and scale: genetic architecture of 2068 traits in the VA Million Veteran Program. *Science* 2024; 385: eadj1182.
- 8 Neale B. Neale Lab. Neale Lab UK Biobank GWASs. Date last accessed: 30 May 2022. www.nealelab.is/uk-biobank
- 9 Woolf B, Pedder H, Rodriguez-Broadbent H, *et al.* Silence is golden, but my measures still see – why cheaper-but-noisier outcome measures in large simple trials can be more cost-effective than gold standards. *Trials* 2024; 25: 532.
- 10 Rhodes CJ, Batai K, Bleda M, *et al.* Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. *Lancet Respir Med* 2019; 7: 227–238.
- 11 Zhou W, Kanai M, Wu KHH, *et al.* Global Biobank Meta-analysis Initiative: powering genetic discovery across human disease. *Cell Genom* 2022; 2: 100192.
- 12 The All of Us Research Program Investigators. The “All of Us” Research Program. *N Engl J Med* 2019; 381: 668–676.
- 13 Mathai SC, Hemnes AR, Manaker S, *et al.* Identifying patients with pulmonary arterial hypertension using administrative claims algorithms. *Ann Am Thorac Soc* 2019; 16: 797–806.
- 14 Lang IM, Palazzini M. The burden of comorbidities in pulmonary arterial hypertension. *Eur Heart J Suppl* 2019; 21: K21–K28.
- 15 Austin ED, Aldred MA, Alotaibi M, *et al.* Genetics and precision genomics approaches to pulmonary hypertension. *Eur Respir J* 2024; 64: 2401370.