Aggregating and harmonizing registry databases for comparative analyses – lessons learnt

Eva-Maria Didden[1], James Weaver[2,5], Dmytro Dymshyts[2,5], Amelie Beaudet[3], Audrey Muller[1], Andrius Kavaliunas[4]

[1] Global Epidemiology Organization, Actelion Pharmaceuticals Ltd., a Johnson & Johnson company, Allschwil, Switzerland
[2] Global Epidemiology Organization, Janssen Research and Development, Raritan, NJ, USA
[3] Global Market Access, Actelion Pharmaceuticals Ltd., a Johnson & Johnson company, Allschwil, Switzerland
[4] Global Epidemiology Organization, Janssen-Cilag AB, a Johnson & Johnson company, Solna, Sweden
[5] Observational Health Data Sciences and Informatics, New York, NY, USA

Background

Pulmonary Arterial Hypertension [PAH] is a rare group of diseases characterized by progressive increase of pulmonary vascular resistance leading to right ventricular failure and premature death (*1*). Several risk assessment algorithms have been published for stratification of PAH patients into 1-year mortality risk (e.g., low, intermediate-low, intermediate-high or high risk), based on their disease status and severity (*2*).
PAH is incurable, but treatable with the goal of achieving low risk status. Endothelial receptor antagonists [ERA] targeting the endothelin pathway, phosphodiesterase type 5 inhibitors [PDE5i] and soluble guanylate cyclase stimulators and [sGCS] targeting the nitride oxide pathway, and prostacyclin analogue and prostacyclin receptor agonist [PA and PRA] targeting the prostacyclin pathway are used as monotherapy or in combination to treat patients with PAH. According to the 2022 ERS/ECS guidelines, patients should receive PRA as add-on therapy if they are at intermediate-low risk at follow-up.

While the benefits of these agents and new treatment modalities have been intensely studied in the randomized controlled clinical trials, evidence from real-world clinical practice is lacking. The objectives of this study were: 1) describe the demographic and clinical characteristics of patients treated for PAH in real-world settings and 2) assess the feasibility of a comparative effectiveness analysis between patients treated with PRA add-on combination (target cohort) and patients treated with other PAH-specific treatment regimens (PDE5i+ERA or sGCS+ERA combination therapy without PRA (comparator cohort)). The study intended to assess whether target cohort therapy resulted in greater benefit (e.g., time to hospitalization, death, parenteral therapy, disease worsening) than comparator cohort therapy in real-world settings. Such evidence would add knowledge and understanding of PAH therapy use in routine care where patients have multiple comorbidities and variable individual treatment regimens.

Methods

This non-interventional study was based on the secondary use of clinical data. Because PAH is rare (*3*), data from four individual disease-specific observational cohort studies were pooled into one database. Data from the four studies were harmonized to the OMOP CDM (v. 5.3.1 at the time of the analysis) and custom vocabulary mapping was developed and applied per a previous PAH data

extract-transform-load procedure (*4*,*5*). The analysis only included variables that were consistently collected and reported across the four studies.

The study population included adult patients diagnosed with PAH who initiated PAH-specific drug treatment between October 2013 and November 2021. Index date specification required in-depth medical consideration, especially for the comparator cohort. For the target cohort, index was defined as the PRA add-on date. For the comparator cohort, index was defined as the date of reaching intermediate-low/intermediate-high/high risk status during 9 months after PDE5i+ERA or sGCS+ERA combination therapy initiation, i.e., when PRA treatment should have been considered as add-on therapy per treatment guidelines. Values closer to 6 months were prioritized over values that are taken further away. Comparator cohort index is the hypothesized counterfactual to PGA initiation.

For the descriptive analysis, i.e., the characterization of the study cohorts, we summarized continuous variables using mean, median, standard deviation, minimum, maximum, upper and lower quartiles, and categorical variables using counts and percentages. We computed Kaplan-Meier estimates to describe the time to occurrence of the event of interest.
For fitting the propensity-score (PS) model, we used a logistic regression and included the following variables: age (years), sex, PAH classification, time from PAH diagnosis to index (months), mortality risk, comorbidities (0, 1, ≥2).
We implemented several 1:1 PS matching strategies (optimal matching (*6*) and two greedy nearest neighbor matching approaches (*7*)) and evaluated these by assessing standardized mean difference (SMD) of observed covariates between the target and comparator cohorts before and after matching. After-matching covariate SMDs assessed exchangeability and reported if any potentially confounding variables were imbalanced at *abs*(SMD)>0.1.

Results

Before and after matching covariate prevalence and SMDs are reported in Table 1 where the greedy nearest neighbor matching strategy was used (*8*). Both cohorts had mostly similar demographic and clinical characteristics. However, the target cohort had a substantially longer time since diagnosis. In addition, geographic distribution differed by cohort.

Table 1. Characteristics of the study population

(Bolded covariates: included in PS model, Highlighted covariates: after matching abs(SMD)>0.1

Covariate	BM T	BM C	BM SMD	AM T	AM C	AM SMD
riskCat: High	0.101	0.083	0.045	0.094	0.097	-0.006
riskCat: IntHigh	0.236	0.357	-0.189	0.304	0.281	0.036
riskCar: IntLow	0.470	0.414	0.079	0.440	0.429	0.016
riskCat: Low	0.194	0.146	0.089	0.161	0.194	-0.060
indexYear: 2017	0.113	0.261	-0.273	0.157	0.150	0.014
indexYear: 2018	0.229	0.210	0.033	0.221	0.235	-0.023
indexYear: 2019	0.355	0.323	0.048	0.369	0.329	0.058
indexYear: 2020	0.178	0.125	0.106	0.161	0.166	-0.009
indexYear: 2021	0.124	0.081	0.100	0.092	0.120	-0.064
age	58.417	61.299	-0.138	60.037	59.507	0.026
timeSincePahDx (months)	60.706	27.134	0.340	38.894	33.279	0.066
Sex: Female	0.716	0.736	-0.032	0.740	0.740	0.000
Sex: Male	0.284	0.264	0.032	0.260	0.260	0.000
Area: Europe	0.512	0.297	0.317	0.401	0.433	-0.046
Area: NorthAmerica	0.488	0.703	-0.317	0.599	0.567	0.046

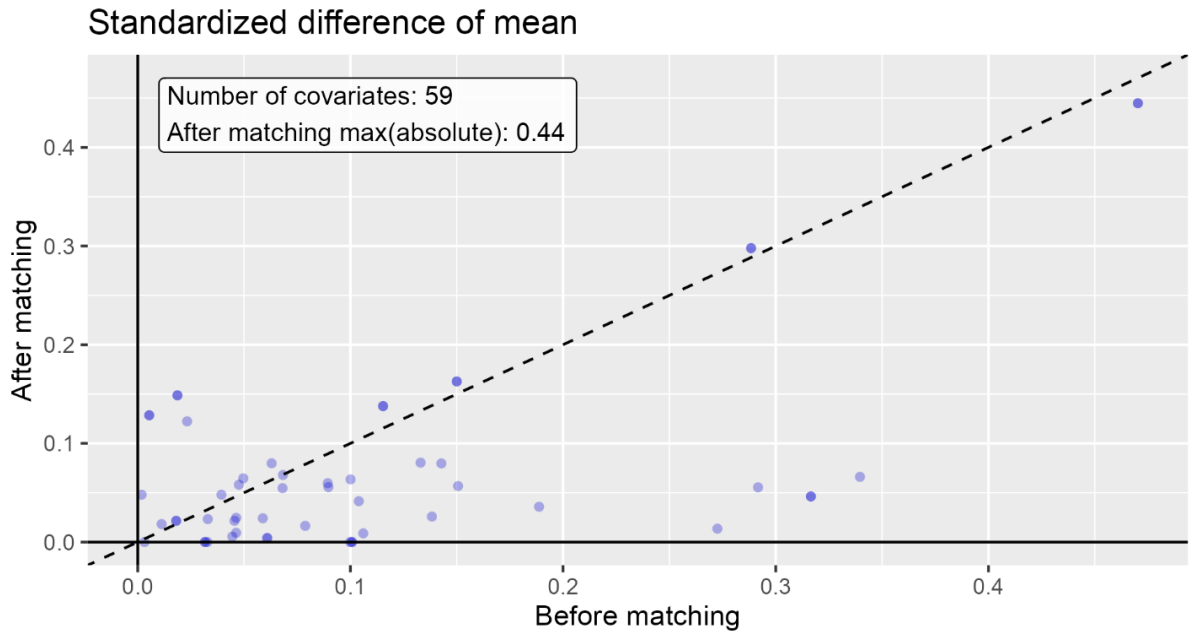
Country: AUT	0.005	0.006	-0.011	0.007	0.009	-0.018
Country: BEL	0.003	0.000	0.056	0.000	0.000	NA
Country: CAN	0.039	0.052	-0.046	0.051	0.058	-0.022
Country: CHE	0.002	0.001	0.002	0.000	0.002	-0.048
Country: CZE	0.005	0.000	0.068	0.005	0.000	0.068
Country: DEU	0.203	0.125	0.151	0.152	0.182	-0.057
Country: DNK	0.002	0.006	-0.050	0.002	0.009	-0.065
Country: ESP	0.093	0.059	0.090	0.065	0.085	-0.056
Country: FIN	0.005	0.004	0.003	0.007	0.007	0.000
Country: FRA	0.002	0.000	0.039	0.002	0.000	0.048
Country: GBR	0.062	0.022	0.143	0.055	0.032	0.080
Country: GRC	0.034	0.023	0.046	0.032	0.030	0.009
Country: ITA	0.036	0.013	0.104	0.030	0.021	0.041
Country: NLD	0.011	0.003	0.068	0.012	0.005	0.055
Country: SVK	0.025	0.003	0.133	0.016	0.005	0.080
Country: SWE	0.026	0.032	-0.023	0.016	0.046	-0.122
Country: USA	0.450	0.651	-0.292	0.548	0.509	0.055
classPah: Idio	0.488	0.465	0.033	0.468	0.468	0.000
classPah: Ctd	0.000	0.000	NA	0.000	0.000	NA
classPah: Chd	0.000	0.000	NA	0.000	0.000	NA
classPah: Other	0.104	0.078	0.063	0.106	0.074	0.080
Htn: NoNa	0.535	0.464	0.101	0.488	0.488	0.000
Htn: Yes	0.465	0.536	-0.101	0.512	0.512	0.000
Diabetes: NoNa	0.792	0.826	-0.061	0.800	0.802	-0.004
Diabetes: Yes	0.208	0.174	0.061	0.200	0.198	0.004
cvDx: NoNa	0.822	0.825	-0.005	0.843	0.772	0.129
cvDx: Yes	0.178	0.175	0.005	0.157	0.228	-0.129
cbDx: NoNa	0.964	0.928	0.115	0.963	0.917	0.138
cbDx: Yes	0.036	0.072	-0.115	0.037	0.083	-0.138
liverDx: NoNa	0.971	0.923	0.150	0.965	0.910	0.163
liverDx: Yes	0.029	0.077	-0.150	0.035	0.090	-0.163
renalDx: NoNa	0.789	0.778	0.019	0.818	0.730	0.149
renalDx: Yes	0.211	0.222	-0.019	0.182	0.270	-0.149
metabolicDx: NoNa	0.583	0.865	-0.470	0.567	0.841	-0.445
metabolicDx: Yes	0.417	0.135	0.470	0.433	0.159	0.445
ctAiDx: NoNa	0.988	0.893	0.288	0.986	0.885	0.298
ctAiDx: Yes	0.012	0.107	-0.288	0.014	0.115	-0.298
gynDx: NoNa	0.994	0.996	-0.018	0.993	0.995	-0.022
gynDx: Yes	0.006	0.004	0.018	0.007	0.005	0.022
Comorbidities: 0	0.349	0.417	-0.100	0.353	0.353	0.000
Comorbidities: 1	0.335	0.304	0.046	0.320	0.304	0.025
Comorbidities: ≥2	0.316	0.278	0.059	0.327	0.343	-0.024

Key – BM T: before matching target covariate baseline prevalence, BM C: before matching comparator covariate baseline prevalence, BM SMD: before matching standardized mean difference, AM T: after matching target covariate baseline prevalence, AM C: after matching comparator covariate baseline prevalence, AM SMD: after matching standardized mean difference, riskCat: mortality risk category, IntHigh: intermediate-high, IntLow: intermediate-low, timeSincePahDx: time between PAH diagnosis and index date (months), classPah: PAH class combined, Idio: idiopathic, Ctd: with connective tissue diseases, Chd: with congenital heart diseases, Htn: hypertension, cvDx: cardiovascular disorders, cbDx: cerebrovascular disorders, liverDx: liver disorders, renalDx: renal disorders, metabolicDx: metabolic disorders excluding diabetes, ctAiDx: connective tissue disorders and autoimmune conditions, NoNa: No/not recorded

After matching by all strategies, residual imbalance persisted, including time from PAH diagnosis to index, which is understood to be associated with study outcomes making it a plausible confounder. Further, baseline conditions remained imbalanced after matching, making confounding by initial health status a plausible threat to validity. Hence, the comparative effectiveness analysis was not carried forward due to high risk of bias as empirically demonstrated. Figure 1 illustrates target and comparator covariate prevalence before and after matching using the greedy nearest neighbor approach (*7*). This includes prevalence for covariates used in fitting the PS model and other covariates

from the PS model analysis set. The figure shows that after matching (y-axis) several covariates remained imbalanced (abs(SMD)>0.1). The covariates are identifiable as highlighted rows in Table 1.

Figure 1. Target and comparator covariate prevalence before and after PS matching.



This study, based on deriving long term therapeutic strategy outcomes from different data sources, when key variables to assess disease severity are not regularly and consistently documented in long term studies, had several challenges and limitations:

- Data pooling across four studies with heterogeneous designs (incl. eligibility criteria), content, and temporal coverage
- Residual confounding (e.g., inability to match the patients on time from diagnosis to index, difficulty to specify an appropriate comparator cohort index date)
- Unmeasured confounding (e.g., lack of consistently and regularly collected clinical data over time)
- Real world clinical practice where triple therapy (i.e., PRA add-on) is delayed; patients may meet high-risk criteria before PRA is prescribed
- Lack of risk progression and disease severity data before PRA initiation and beyond initial assessment
- Short patient observation time
- Few outcome events.

Conclusion

Although the clinically rich pooled PAH database provided insights into the characteristics of PAH patients treated with different regimens, it appeared infeasible for PS matching to create adequately exchangeable exposure cohorts for valid comparative analyses. Post-hoc assessment suggested that our main limitations were 1) difficulty with designing an appropriate index date for the comparator group, and 2) differences between patient populations, especially in terms of time between diagnosis and treatment initiation. This is an example where – despite in-depth study design comparisons, database evaluations, and feasibility assessments – comparative effectiveness analysis was deemed infeasible.

This study also highlighted the importance of the PS matching diagnostics evaluation in the comparative effectiveness research, in particular exemplifying an informed decision making to prevent biased results.

References

1. Humbert M, et al., 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. European Respiratory Journal. 2023. DOI: 10.1183/13993003.00879-2022
2. Hoeper MM, et al., COMPERA 2.0: a refined four-stratum risk assessment model for pulmonary arterial hypertension. Eur Respir J. 2022 Jul 7;60(1):2102311. doi: 10.1183/13993003.02311-2021. PMID: 34737226; PMCID: PMC9260123.
3. Leber L, Beaudet A, Muller A. Epidemiology of pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension: identification of the most accurate estimates from a systematic literature review. Pulm Circ. 2021 Jan 7;11(1):2045894020977300. doi: 10.1177/2045894020977300. PMID: 33456755; PMCID: PMC7797595.
4. Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, Wetherill G, Brand M, Didden EM. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol. 2021 Nov 2;21(1):238. doi: 10.1186/s12874-021-01434-3. PMID: 34727871; PMCID: PMC8565035.
5. Handbook for PH registries to OMOP CDM conversion: https://github.com/OHDSI/ETL--PulmonaryHypertensionRegistries
6. MatchIt package (Ho, Imai, King, & Stuart, 2011) in R. https://doi.org/10.18637/jss.v042.i08.
7. Schuemie M, Suchard M, Ryan P (2024). CohortMethod: New-User Cohort Method with Large Scale Propensity and Outcome Models. https://ohdsi.github.io/CohortMethod, https://github.com/OHDSI/CohortMethod.
8. Schuemie M, Reps J, Black A, Defalco F, Evans L, Fridgeirsson E, Gilbert JP, Knoll C, Lavallee M, Rao GA, Rijnbeek P, Sadowski K, Sena A, Swerdel J, Williams RD, Suchard M. Health-Analytics Data to Evidence Suite (HADES): Open-Source Software for Observational Research. Stud Health Technol Inform. 2024 Jan 25;310:966-970. doi: 10.3233/SHTI231108. PMID: 38269952; PMCID: PMC10868467.