

Characterization: Does it matter how you pick your non-cases in risk factor analyses?

Jenna M Reps & Patrick B Ryan
Johnson & Johnson, Titusville, NJ, USA

Background

The OHDSI community have developed an open-source R package for descriptive analyses called 'Characterization' [1]. The Characterization R package performs various descriptive analyses directly on OMOP data including risk factor, target baseline, case series, dechallenge-rechallenge and time-to-event analyses. The risk factor analysis finds covariates (aka features or attributes) that are associated with experiencing an outcome during a time at risk within a target population. That is, it can be used to find covariates that when present may suggest an increased or decreased chance of having the outcome in some future time-period compared to not having the covariate. It works by finding the patients in the target cohort who have the outcome during the time-at-risk (cases) and the patients in the target cohort who do not have the outcome during the time-at-risk (non-cases) and calculating the standardized mean difference (SMD) [2] using the proportion of cases and non-cases with the covariate. However, there are different ways that the non-cases can be defined, some more efficient than others. In this study we aim to quantify how much impact the choice of defining non-cases has on the covariates that are identified as being associated with having the outcome.

Methods

We compare three different ways to define non-cases:

1. Efficient: non-cases are the target population less the cases.
2. CohortIncidence: non-cases are the target population less the cases that occurred during time-at-risk and with at least 1 day of time-at-risk. In addition, patients are excluded if they left the database before time-at-risk starts, their outcome washout covers the whole time-at-risk, or their time-at-risk is 0 days.
3. PatientLevelPrediction: non-cases are the target population less the cases that occurred during the time-at-risk or within washout days prior. In addition, patients are excluded if they left the database before time-at-risk starts, they have the outcome during the outcome washout prior to time-at-risk start or their time-at-risk is 0 days.

We ran the risk factor analysis for ten target cohorts, sixteen outcomes (seven acute that can reoccur and nine chronic that cannot reoccur) and two time-at-risks (while exposed and fixed 1 year after index). We included candidate covariates: age/age in 5-year groups, Charlson comorbidity index, sex, any condition or drug that occurred in the 365 days prior to index and the constrained predictor cohort covariates [3]. Merative CCAE mapped to the OMOP CDM was used to run the analysis. We only included covariates that 1% or more of either the cases or non-cases had.

We compared agreement of identified covariates between the three models by calculating the spearman's rank correlation coefficient of the covariates' SMDs between each pair of modes, if a covariate was not identified by one of the modes it was assigned a SMD of 0. We also compared overlap percentage as the number of 'associated' covariates identified in common between two modes at different min SMD values (0.05, 0.1, 0.2, 0.5 and 0.8).

Results

It took 44 minutes to execute mode 'Efficient', 59 minutes to execute mode 'CohortIncidence' and 56 minutes to execute mode 'PatientLevelPrediction'. There was a total of 781,074 covariates included in the analysis. The rank correlation is displayed in Table 1.

Table 1- the Spearman's rank coefficient comparing the SMD values across all covariates for all prediction tasks (All outcomes), prediction tasks for acute outcomes (Acute) and prediction tasks for Chronic outcomes (Chronic).

Outcomes	Efficient vs CohortIncidence	Efficient vs PatientLevelPrediction	CohortIncidence vs PatientLevelPrediction
All	0.980	0.979	0.998
Acute	0.990	0.987	0.996
Chronic	0.973	0.973	1

The overlap percent per SMD threshold is displayed in Table 2.

Table 2- overlap across each pair of modes in covariates identified at different min SMD threshold. There was 781,074 covariates in total.

Threshold	Efficient vs CohortIncidence			Efficient vs PatientLevelPrediction			CohortIncidence vs PatientLevelPrediction		
	Both	Efficient only	CI only	Both	Efficient only	PLP only	Both	CI only	PLP only
0.05	95.7%	0.1%	4.2%	95.4%	0.2%	4.4%	99.7%	0.1%	0.3%
0.1	94.4%	0.2%	5.5%	94.0%	0.2%	5.8%	99.5%	0.0%	0.4%
0.2	91.7%	0.1%	8.2%	90.8%	0.1%	9.1%	99.0%	0.0%	0.9%
0.5	89.2%	0.2%	10.6%	87.1%	0.2%	12.8%	97.5%	0.0%	2.5%
0.8	90.2%	1.8%	8.0%	87.9%	1.8%	10.3%	97.4%	0.0%	2.6%

The CohortIncidence and PatientLevelPrediction mode generate very correlated SMD values for covariates when the outcomes are acute and can reoccur. When the outcome is chronic the CohortIncidence and PatientLevelPrediction mode exclude the same patients and are equivalent. The PatientLevelPrediction mode appears to identify more associated covariates at higher SMD thresholds.

The efficient mode was slightly faster (44 mins vs 59/56 minutes) and there was high correlation between the efficient and other modes' SMDs when the outcomes where acute. This was slightly lower for chronic outcomes. Some covariates associated with the outcome were not identified with the efficient mode but would be with other modes, making this mode appear less sensitive.

Conclusions

The results show that the SMD calculated using different modes to identify the non-cases are highly correlated and most modes identify the same covariates as being associated with the outcome. The 'Efficient' mode is much quicker to run when implementing a large-scale analysis (hundreds or targets and hundreds of outcomes) and appears to perform similarly to 'PatientLevelPrediction' and 'CohortIncidence' mode. The 'CohortIncidence' and 'PatientLevelPrediction' modes perform almost identically. The difference between these modes is that 'CohortIncidence' includes a patient as long as they have 1 day or more of time-at-risk, whereas 'PatientLevelPrediction' excludes anyone who has the outcome before time-at-risk that causes them to lose 1 or more day or time-at-risk. The results suggest that how you handle patients who have the outcome within the outcome washout period prior to time-at-risk makes little difference in the risk factors identified.

If you are going to run a large-scale risk factor analysis then use the 'Efficient' mode as this will save a significant amount of time and computational energy. The 'Efficient' mode is a good approximation of the results obtained with more accurate modes. If you are not performing a large-scale analysis, then both the 'CohortIncidence' and 'PatientLevelPrediction' modes are suitable, and it makes little difference which is picked. Pick CohortIncidence' when you are also primarily interested in

incidence rates or 'PatientLevelPrediction' when you are primarily interested in patient-level prediction models.

References

- [1] Reps, J., 2026. *Characterization Package Specification*. [Online] Available at: <https://ohdsi.github.io/Characterization/articles/Specification.html> [Accessed 20 April 2026].
- [2] Andrade, C., 2020. Mean difference, standardized mean difference (SMD), and their use in meta-analysis: as simple as it gets. *The Journal of clinical psychiatry*, 81(5), pp.20f13681-20f13681.
- [3] Reps, J.M., Wong, J., Fridgeirsson, E.A., Kim, C., John, L.H., Williams, R.D., Fisher, R.R. and Ryan, P.B., 2025. Finding a constrained number of predictor phenotypes for multiple outcome prediction. *BMJ Health & Care Informatics*, 32(1), p.e101227.

Supplement

Efficient vs CohortIncidence covariates only identified by CohortIncidence mode using while exposed TAR and threshold 0.5 - **10 example variables**

target_name	outcome_name	covariate_name	e_smd	ci_smd	plp_smd
warfarin	earlier arrhythmia	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Heart disease	-0.05	0.51	0.51
warfarin	all events DVT	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Venous thrombosis	0.07	0.53	0.64
warfarin	all events DVT	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Thrombosis	0.08	0.52	0.63
warfarin	all events DVT	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Thrombosis of blood vessel	0.08	0.53	0.63
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	earliest heart failure	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Structural disorder of heart	0.27	0.51	0.51
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	earliest heart failure	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Heart disease	0.30	0.56	0.56
statins	earlier arrhythmia	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Heart disease	0.33	0.53	0.53
warfarin	all events DVT	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Vascular disorder	0.35	0.57	0.60
warfarin	all events DVT	drug_era group (DrugGroupEraLongTerm) during day -365 through 0 days relative to index: Heparin group	0.36	0.56	0.60
warfarin	all events DVT	drug_era group (DrugGroupEraLongTerm) during day -365 through 0 days relative to index: enoxaparin	0.37	0.59	0.62

CohortIncidence vs PatientLevelPrediction covariates only identified by PatientLevelPrediction mode using while exposed TAR and threshold 0.5 – **10 example variables**

target_name	outcome_name	covariate_name	e_smd	ci_smd	plp_smd
-------------	--------------	----------------	-------	--------	---------

Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Hepatic failure	-0.38	-0.09	0.73
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Hepatic coma due to viral hepatitis	0.17	0.11	1.05
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Hepatic coma due to hepatitis	0.17	0.11	1.05
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Coma	0.15	0.11	1.10
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Hepatic coma	0.16	0.11	1.08
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Toxic metabolic encephalopathy	0.08	0.18	1.42
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Hepatic encephalopathy	0.13	0.18	1.41
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Metabolic encephalopathy	0.10	0.18	1.42
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Central nervous system complication	0.13	0.20	1.32
Earliest event of Acute Hepatic Failure, NO prior liver cirrhosis	All events of Acute Hepatic Failure, with washout period of 365 days	condition_era group (ConditionGroupEraLongTerm) during day -365 through 0 days relative to index: Viral disease	0.20	0.21	0.59