

SHIFT: Federated Feature-Level Transfer Learning for High-Dimensional Survival Analysis under Case–Cohort Sampling

Tingyin Wang^{1,2}, Yudong Wang^{2,3}, Jie Hu^{2,3}, Yong Chen^{1-3, 4-6*†}

¹ The Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA

² The Center for Health AI and Synthesis of Evidence (CHASE), University of Pennsylvania, PA, USA

³ Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

⁴ Penn Medicine Center for Evidence-based Practice (CEP), Philadelphia, PA, USA

⁵ Penn Institute for Biomedical Informatics (IBI), Philadelphia, PA, USA

⁶ Leonard Davis Institute of Health Economics, Philadelphia, PA, USA

Background

Multi-cohort observational studies that pair longitudinal follow-up with high-dimensional molecular profiling, whether proteomic, metabolomic, or transcriptomic, are increasingly central to biomarker discovery for incident disease. Assaying hundreds to thousands of analytes on every participant of a large cohort is, however, prohibitively expensive. Case–cohort designs address this by restricting measurement to all incident cases and a random subcohort, but the resulting outcome-dependent sampling invalidates the classical Cox partial likelihood and limits effective sample size to the observed failure count^{1,2}. For rare endpoints, this constraint is severe: in the multi-country proteomic data analyzed below, an ovarian-cancer target cohort yields as few as one LASSO-selected protein and a target-only concordance index of 0.535, effectively chance.

A natural remedy is to borrow strength from related cohorts collected under harmonized protocols, an approach directly aligned with the OHDSI mission of federated, multi-site evidence generation. However, naive pooling across cohorts induces severe bias when molecular signatures differ across populations due to lifestyle exposures, assay batch effects, or disease etiology³. Existing high-dimensional transfer learning methods operate at the source level, retaining or discarding entire cohorts based on global similarity, a binary strategy that is particularly wasteful when the number of source cohorts is small, and that does not accommodate the outcome-dependent sampling structure of case–cohort designs. We propose SHIFT (Shared Informative Feature Transfer), a transfer learning framework for high-dimensional Cox regression that performs transfer at the feature level and provides a formal guarantee against negative transfer.

Methods

SHIFT consists of three steps. First, the informative support S of the target model is identified via a weighted Lasso–Cox estimator on the target case–cohort data, using Barlow inverse-probability weights to correct for the sampling design. Second, each source cohort is fit separately with an unweighted Cox–Lasso, and a debiasing step on S estimates the cohort-specific discrepancy $\delta(k)$ restricted to the support. Third, source information is aggregated through adaptive weights $\alpha_k \propto \left(1 + \|\delta^{(k)}\|_2\right)^{-1}$, which up-weight large, on-support-compatible sources and down-weight discrepant ones.

We classify each source into three types: *transferable* (small discrepancy on both S and S^c), *partial-transferable* (small on S , large on S^c), and *non-transferable* (large on S). By restricting transfer to S , SHIFT exploits partial-transferable sources, which all-or-nothing methods discard, while suppressing non-transferable ones. We prove a non-asymptotic error bound showing SHIFT attains a strictly faster rate than the target-only weighted Lasso–Cox

when sources are compatible on S , and recovers the target-only rate otherwise.

We evaluate SHIFT on multi-cohort proteomic case-cohort data from European prospective cohorts coordinated by IARC, spanning Italy, Spain, the Netherlands, and the United Kingdom, with three cancer endpoints (lung, breast, ovarian) selected to span a gradient of etiological concentration and target sample size. Each scenario rotates which country serves as the target; the remaining three act as auxiliary sources. Predictions are evaluated by the concordance index on held-out target data under 3-fold stratified cross-validation, averaged over 30 independent subcohort draws (20% sampling fraction). Source code and reproducibility materials will be made publicly available on GitHub.

Results

Across six real-data scenarios spanning positive, mixed, and negative transfer regimes, SHIFT achieves the highest mean C-index in every scenario (Table 1). Naive pooling falls below the target-only baseline in five of six scenarios, with its largest deficit in the most extreme negative-transfer scenario (3b, ovarian/Netherlands), where Pool collapses to 0.519 (essentially random discrimination) while SHIFT preserves predictive utility at 0.695. The magnitude of SHIFT's advantage scales with regime difficulty: modest under positive transfer (where target-only is near a ceiling), moderate under mixed transfer, and substantial under negative transfer, matching the theoretical guarantee that SHIFT improves on target-only when sources are compatible and matches it when they are not. Figure 1 visualizes this pattern across all six scenarios.

Table 1. Mean C-index across six real-data transfer scenarios, averaged over 30 subcohort draws. Best result per row in bold.

Regime	Scenario (Target)	Target-only	Pool	SHIFT
Positive	1a Lung (Italy)	0.713	0.707	0.741
Positive	1b Lung (Spain)	0.789	0.789	0.795
Mixed	2a Breast (Italy)	0.657	0.645	0.662
Mixed	2b Breast (Netherlands)	0.596	0.579	0.623
Negative	3a Ovarian (UK)	0.663	0.698	0.796
Negative	3b Ovarian (Netherlands)	0.535	0.519	0.695

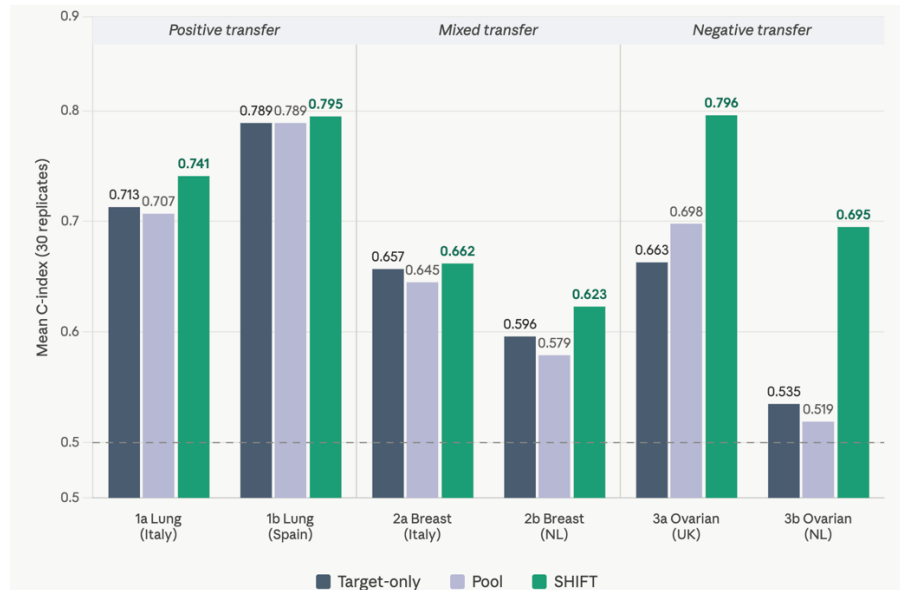


Figure 1. Mean concordance index (C-index) across the six real-data transfer scenarios for three estimators: target-only, naive pooling (Pool), and SHIFT. Scenarios are grouped by transfer regime (positive, mixed, and negative). Bars show the mean over 30 independent subcohort draws (20% sampling fraction), and the dashed reference line at 0.5 marks chance-level discrimination. SHIFT attains the highest C-index in every scenario, with its largest gains under negative transfer.

Conclusion

SHIFT addresses a problem of direct relevance to the OHDSI community: how to safely borrow predictive signal across heterogeneous observational cohorts when the target alone is underpowered. By performing transfer at the feature level rather than the source level, and by explicitly correcting for case-cohort sampling weights, SHIFT functions as a "free option" on auxiliary cohorts: it captures efficiency gains where sources are compatible and absorbs no measurable loss where they are not. The framework is particularly suited to rare-endpoint settings, federated multi-site analyses, and any context in which uncritical cohort concatenation risks importing population-specific biases. A federated implementation that operates on summary statistics only, suitable for the OHDSI network-study paradigm, is a natural extension and is the focus of ongoing work.

References

1. Barlow WE. Robust variance estimation for the case-cohort design. *Biometrics*. 1994;50(4):1064-72.
2. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986;73(1):1-11.
3. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med*. 2023;21:70.
4. Li S, Cai TT, Li H. Transfer learning for high-dimensional linear regression: prediction, estimation and minimax optimality. *J R Stat Soc Ser B*. 2022;84(1):149-73.
5. Tian Y, Feng Y. Transfer learning under high-dimensional generalized linear models. *J Am Stat Assoc*. 2023;118(544):2684-97.
6. Riboli E, Hunt KJ, Slimani N, et al. European prospective investigation into cancer and nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002;5(6B):1113-24.