Our strategy can be applied whether or not the training sample is naturally divisible into different studies or sub-populations. We investigated this case as a limiting case when heterogeneity is close to 0. The findings from our baseline analysis (see Figure 1) suggest that for larger datasets, it may be advantageous to train SSL's on subsections and ensemble using stacking weights that reward prediction across different sections of the data, rather than simply training a learner on the whole dataset. Focusing SSL's on fewer observations may allow the learners to capture more of the specific features of the dataset, which can then be combined in a way that promotes cross-study generalizability. Simply training one learner on the entire dataset may ignore potential heterogeneity in feature distributions between subsets of observations. Further exploration of the single study setting can be found in the supplement.

Acknowledgements

We thank Matt Ploenzke and Lorenzo Trippa for useful suggestions. Maya Ramchandran and Prasad Patil were supported by NIH-NCI Training Grant T32CA009337. Giovanni Parmigiani was supported by grants NSF-DMS 1810829 and NIH-NCI 4P30CA006516-51.

References

1. C. Bernau, M. Riester, A.-L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron and L. Trippa, Cross-study validation for the assessment of prediction algorithms, *Bioinformatics* **30**, i105 (2014).
2. P. Patil and G. Parmigiani, Training replicable predictors in multiple studies, *Proceedings of the National Academy of Sciences* **115**, 2578 (2018).
3. L. Breiman, Random forests, *Machine Learning* **45**, 5 (October 2001).
4. L. Breiman, Bagging predictors, *Machine Learning* **24**, 123 (August 1996).
5. J. Maudes, J. J. Rodríguez, C. García-Osorio and N. García-Pedrajas, Random feature weights for decision tree ensemble construction, *Inf. Fusion* **13**, 20 (January 2012).
6. S. J. Winham, R. R. Freimuth and J. M. Biernacka, A weighted random forests approach to improve predictive performance, *Statistical Analysis and Data Mining: The ASA Data Science Journal* **6** (2013).
7. R. Rahman, S. Haider, S. Ghosh and R. Pal, Design of probabilistic random forests with applications to anticancer drug sensitivity prediction, in *Cancer informatics*, 2015.
8. H. Kim, H. Kim, H. Moon and H. Ahn, A weight-adjusted voting algorithm for ensembles of classifiers, *Journal of the Korean Statistical Society* **40**, 437 (12 2011).
9. S. Basu, K. Kumbier, J. B. Brown and B. Yu, Iterative random forests to discover predictive and stable high-order interactions, *Proceedings of the National Academy of Sciences* **115** (2018).
10. B. F. Ganzfried, M. Riester, B. Haibe-Kains, T. Risch, S. Tyekucheva, I. Jazic, X. V. Wang, M. Ahmadifar, M. J. Birrer, G. Parmigiani, C. Huttenhower and L. Waldron, curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome., *Database (Oxford)* **2013**, p. bat013 (2013), PMCID: PMC3625954.
11. L. Breiman, Stacked regressions, *Machine Learning* **24**, 49 (July 1996).
12. J. Friedman, T. Hastie and R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**, 1 (2010).
13. Y. Zhang, C. Bernau, G. Parmigiani and L. Waldron, The impact of different sources of heterogeneity on loss of accuracy from genomic prediction models, *Biostatistics (Oxford, England)* **6**, p. 701 (September 2018).
14. C. R. Planey and O. Gevaert, Coincide: A framework for discovery of patient subtypes across multiple datasets, *Genome Medicine* **8** (March 2016).