

Supplementary to “Conformalized Multiview Learning”

Saptarshi Roy^{1†}, Sreya Sarkar^{2†}, Chitrak Banerjee³, Erina Paul⁴, Piyali Basak⁴, Himel Mallick^{5,6,7*}

¹*Department of Statistics, Texas A&M University, College Station, TX, USA*

²*Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA, USA*

³*Analytical AI, Bristol Myers Squibb, Princeton, NJ, USA*

⁴*Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, NJ, USA*

⁵*Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York, NY*

⁶*Division of Gastroenterology and Hepatology, Weill Cornell Medicine, Cornell University, New York, NY*

⁷*Department of Statistics and Data Science, Cornell University, Ithaca, NY*

[†] These authors contributed equally

* Corresponding Author Email: him4004@med.cornell.edu

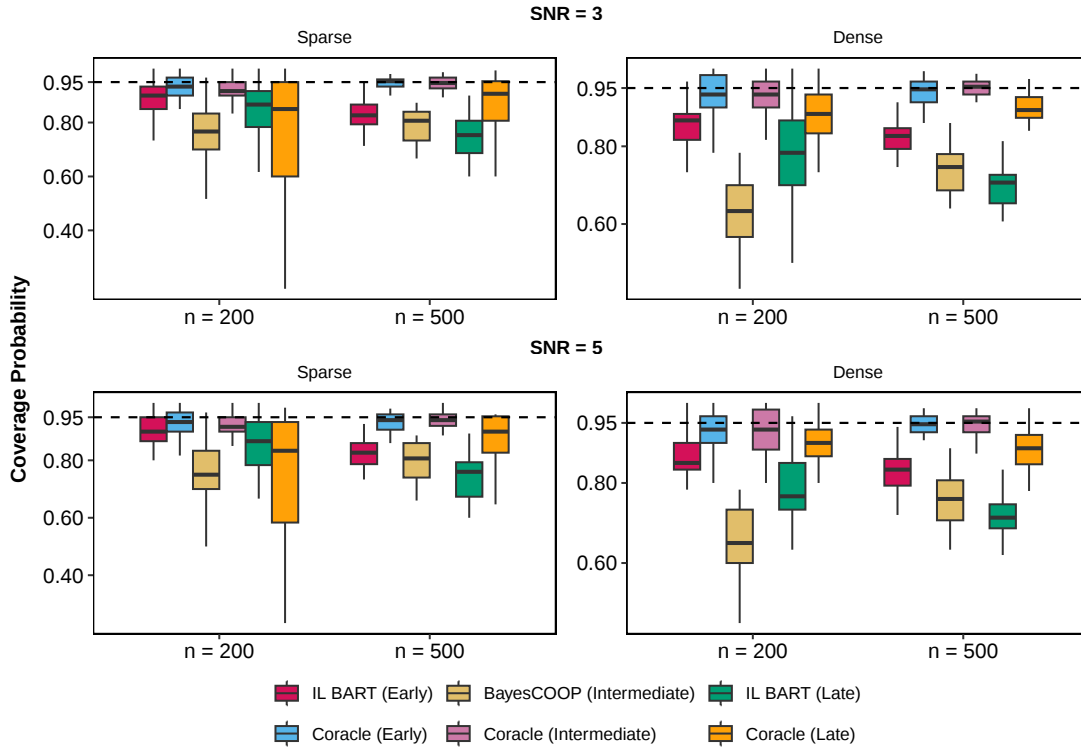


Figure S1: **Data-driven, non-factor model-based simulations with BART as base learner for late fusion in moderate SNR settings.** Coverage performance for Coracle and published non-conformal baselines is evaluated under two signal-to-noise ratios (SNR = 3 and 5), two sample sizes ($n = 200, 500$), and both sparse and dense coefficient regimes. Here, *IL BART* denotes *IntegratedLearner* (Mallick et al., 2024) with BART as the base learner (Chipman et al., 2010), and *BayesCOOP* (Roy et al., 2025) represents the intermediate fusion baseline. For each configuration, median coverage probabilities are computed using completely held-out test data and summarized over 25 replications. Coracle maintains high and stable coverage across early, intermediate, and late fusion architectures, whereas non-conformal fusion methods exhibit noticeable deviations from nominal coverage.

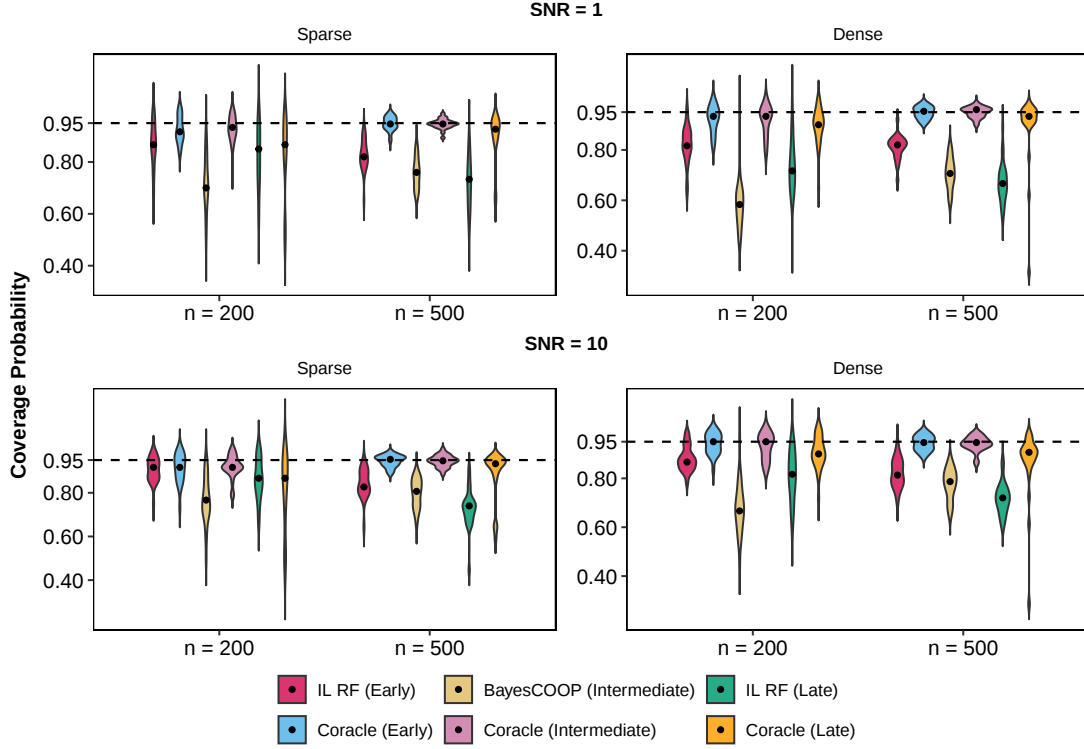


Figure S2: **Data-driven, non-factor model-based simulations with random forest as base learner for late fusion in high- and low-SNR settings.** Coverage performance for Coracle and published non-conformal baselines is evaluated under two signal-to-noise ratios ($\text{SNR} = 1$ and 10), two sample sizes ($n = 200, 500$), and both sparse and dense coefficient regimes. Here, *IL RF* denotes *IntegratedLearner* (Mallick et al., 2024) with random forest as the base learner (Breiman, 2001; Liaw and Wiener, 2002), and *BayesCOOP* (Roy et al., 2025) represents the intermediate fusion baseline. For each configuration, median coverage probabilities are computed using completely held-out test data and summarized over 25 replications. Coracle maintains high and stable coverage across early, intermediate, and late fusion architectures, whereas non-conformal fusion methods exhibit noticeable deviations from nominal coverage.

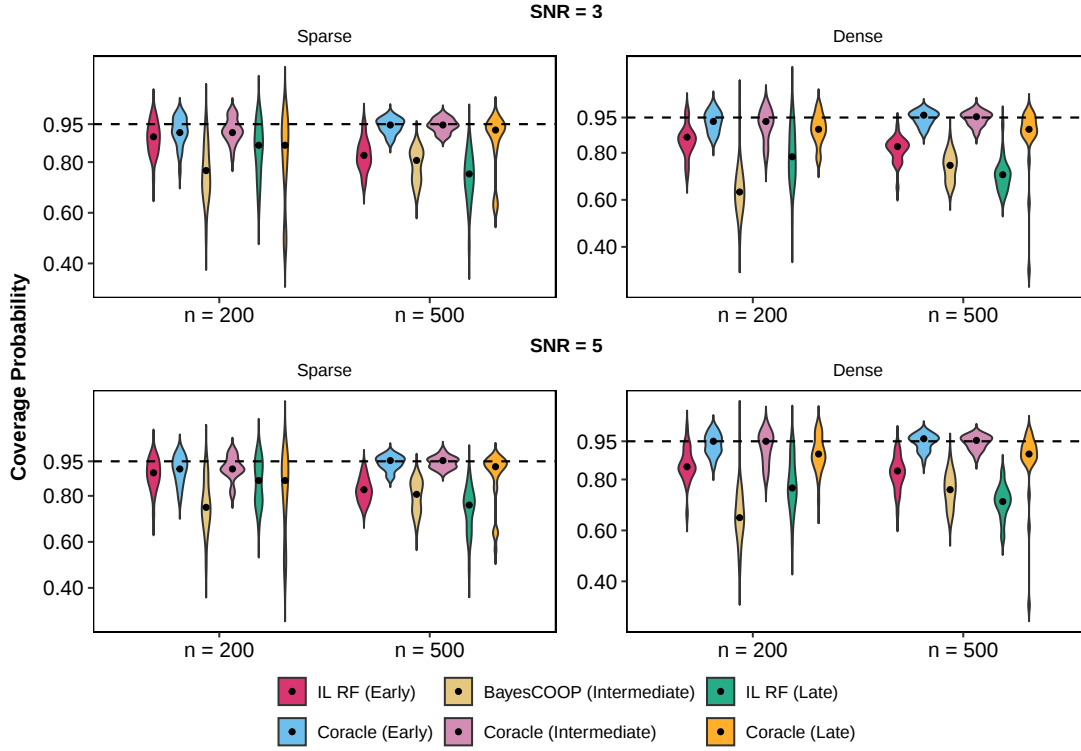


Figure S3: **Data-driven, non-factor model-based simulations with random forest as base learner for late fusion in moderate SNR settings.** Coverage performance for Coracle and published non-conformal baselines is evaluated under two signal-to-noise ratios ($\text{SNR} = 3$ and 5), two sample sizes ($n = 200, 500$), and both sparse and dense coefficient regimes. Here, *IL RF* denotes *IntegratedLearner* (Mallick et al., 2024) with random forest as the base learner (Breiman, 2001; Liaw and Wiener, 2002), and *BayesCOOP* (Roy et al., 2025) represents the intermediate fusion baseline. For each configuration, median coverage probabilities are computed using completely held-out test data and summarized over 25 replications. Coracle maintains high and stable coverage across early, intermediate, and late fusion architectures, whereas non-conformal fusion methods exhibit noticeable deviations from nominal coverage.

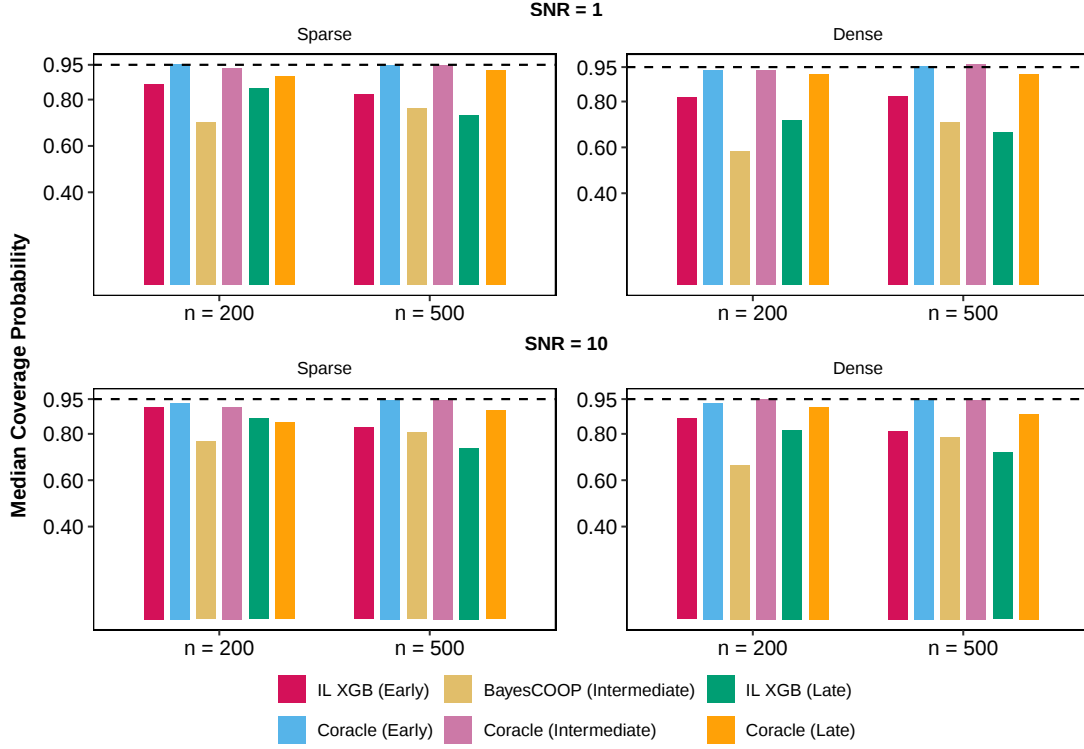


Figure S4: **Data-driven, non-factor model-based simulations with XGBoost as base learner for late fusion in high- and low-SNR settings.** Coverage performance for Coracle and published non-conformal baselines is evaluated under two signal-to-noise ratios (SNR = 1 and 10), two sample sizes ($n = 200, 500$), and both sparse and dense coefficient regimes. Here, *IL XGB* denotes *IntegratedLearner* (Mallick et al., 2024) with XGBoost as the base learner (Chen and Guestrin, 2016), and *BayesCOOP* (Roy et al., 2025) represents the intermediate fusion baseline. For each configuration, median coverage probabilities are computed using completely held-out test data and summarized over 25 replications. Coracle maintains stable near-nominal coverage across fusion strategies, while non-conformal methods remain noticeably miscalibrated.

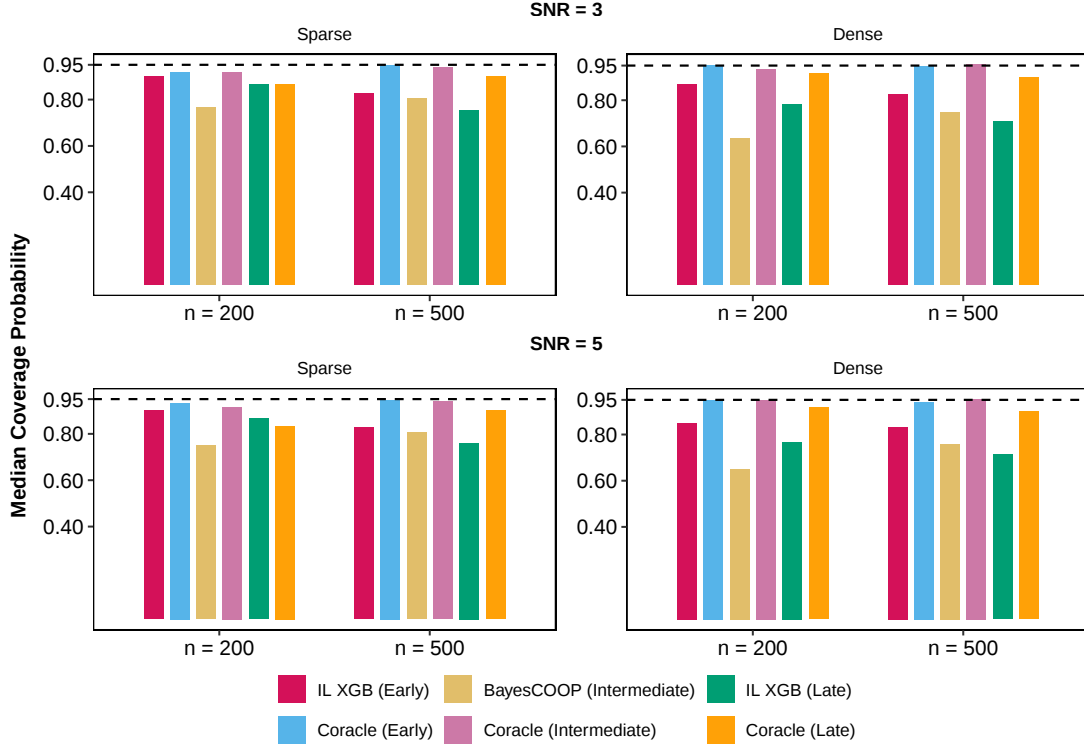


Figure S5: **Data-driven, non-factor model-based simulations with XGBoost as base learner for late fusion in moderate SNR settings.** Coverage performance for Coracle and published non-conformal baselines is evaluated under two signal-to-noise ratios ($\text{SNR} = 3$ and 5), two sample sizes ($n = 200, 500$), and both sparse and dense coefficient regimes. Here, *IL XGB* denotes *IntegratedLearner* (Mallick et al., 2024) with XGBoost as the base learner (Chen and Guestrin, 2016), and *BayesCOOP* (Roy et al., 2025) represents the intermediate fusion baseline. For each configuration, median coverage probabilities are computed using completely held-out test data and summarized over 25 replications. Coracle maintains stable near-nominal coverage across fusion strategies, while non-conformal methods remain noticeably miscalibrated.

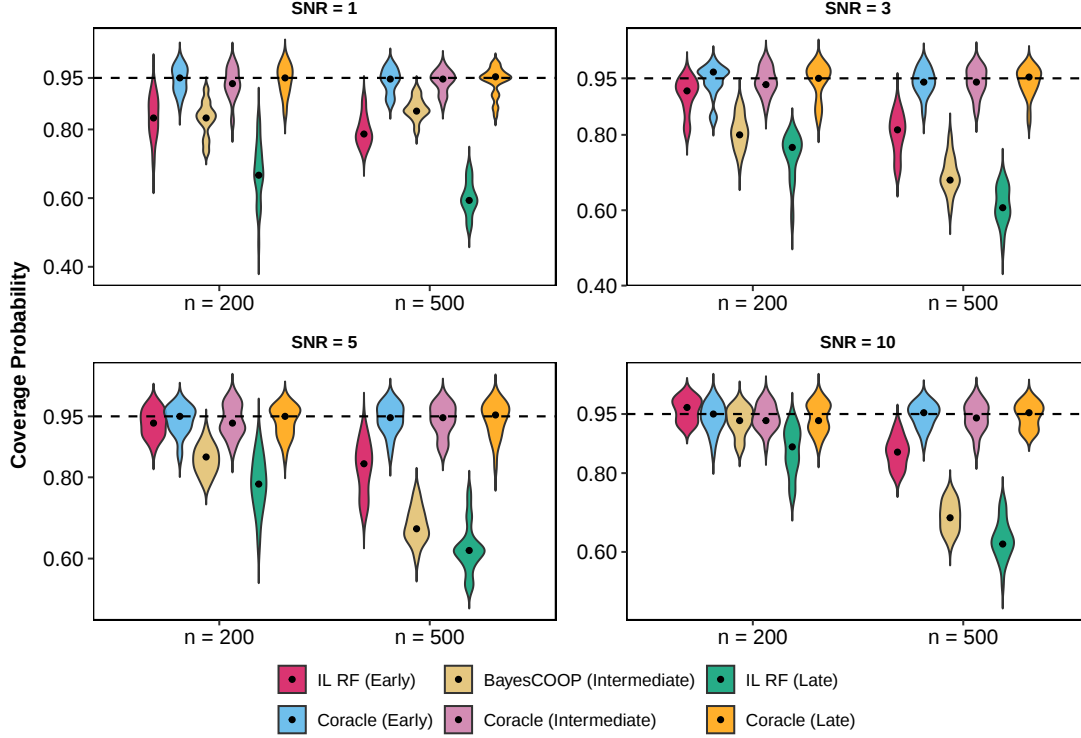


Figure S6: **Factor model-based simulations with random forest as base learner for late fusion.** Coverage performance for Coracle and published non-conformal baselines is evaluated under four signal-to-noise ratios ($\text{SNR} = 1, 3, 5, 10$) and two sample sizes ($n = 200, 500$). Here, *IL RF* denotes *IntegratedLearner* (Mallick et al., 2024) with random forest (Breiman, 2001; Liaw and Wiener, 2002) as the base learner, and *BayesCOOP* (Roy et al., 2025) represents the intermediate fusion baseline. For each configuration, median coverage probabilities are computed using completely held-out test data and summarized over 25 replications. Coracle consistently achieves near-nominal coverage across all fusion architectures, whereas non-conformal alternatives exhibit substantial miscalibration.

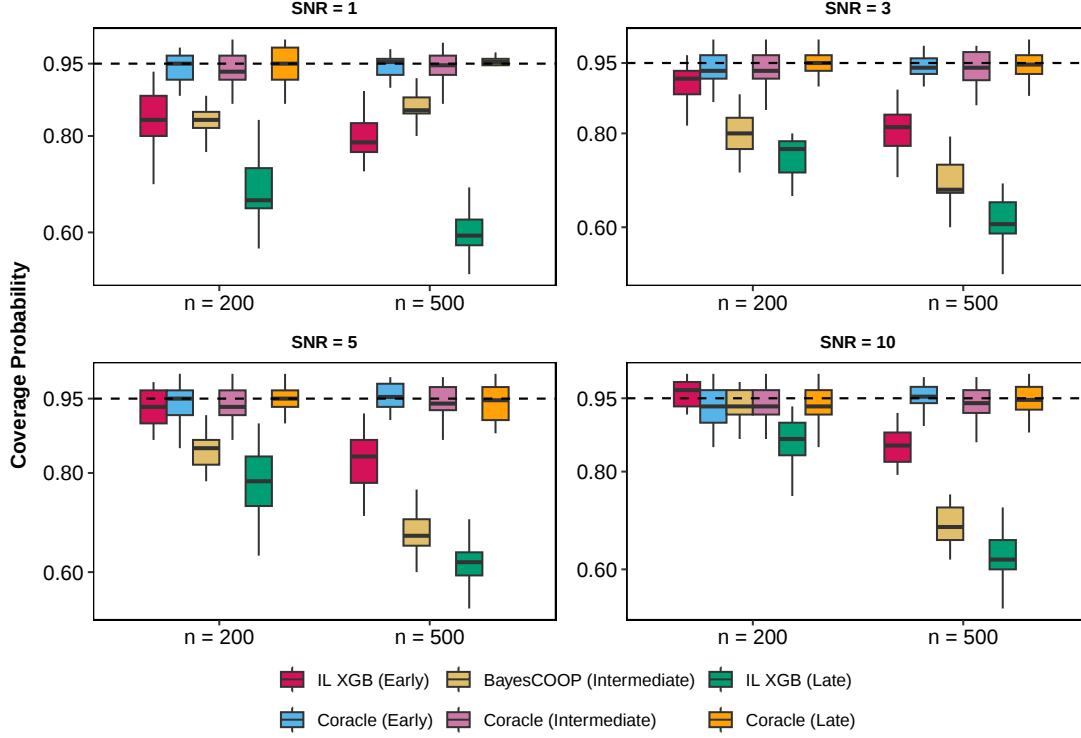


Figure S7: **Factor model–based simulations with XGBoost as base learner for late fusion.** Coverage performance for Coracle and published non-conformal baselines is evaluated under four signal-to-noise ratios ($\text{SNR} = 1, 3, 5, 10$) and two sample sizes ($n = 200, 500$). Here, *IL XGB* denotes *IntegratedLearner* (Mallick et al., 2024) with XGBoost (Chen and Guestrin, 2016) as the base learner, and *BayesCOOP* (Roy et al., 2025) represents the intermediate fusion baseline. For each configuration, median coverage probabilities are computed using completely held-out test data and summarized over 25 replications. Coracle consistently achieves near-nominal coverage across all fusion architectures, whereas non-conformal alternatives exhibit substantial miscalibration.

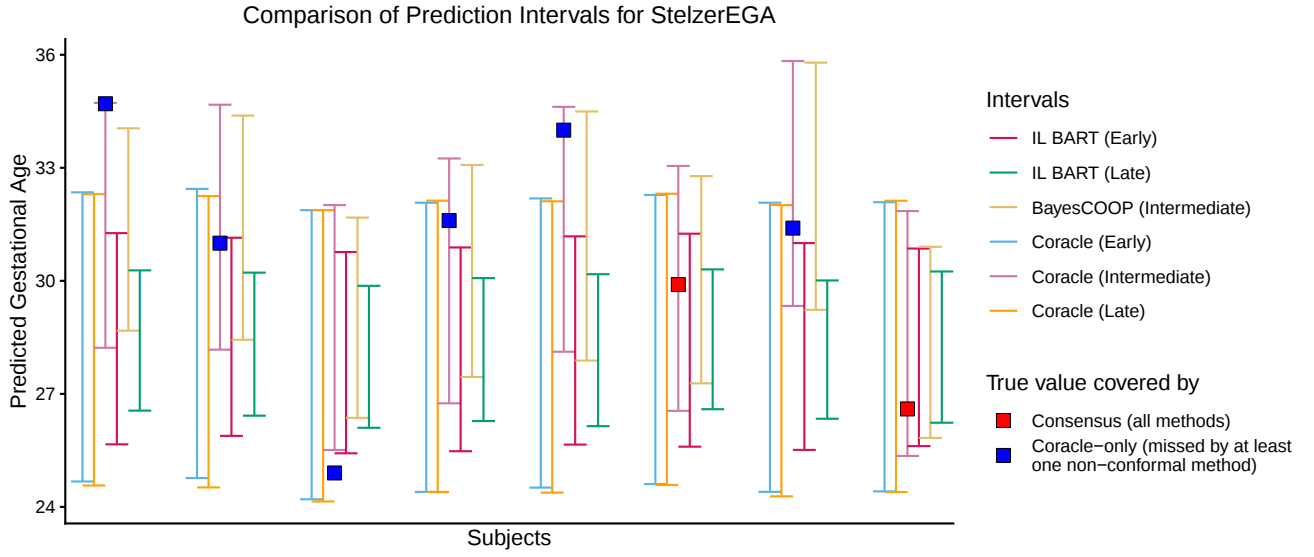


Figure S8: **Prediction interval comparison for the StelzerEGA dataset** (Stelzer et al., 2021). Subject-level prediction intervals for gestational age are shown for representative early, intermediate, and late fusion base models and their corresponding Coracle-conformalized versions, evaluated on completely independent validation data. Here, *IL BART* denotes *IntegratedLearner* (Mallick et al., 2024) with BART (Chipman et al., 2010) as the base learner (shown under early and late fusion), and *BayesCOOP* (Roy et al., 2025) denotes the Bayesian cooperative learning model used as the intermediate fusion baseline. For each subject, the true value is indicated, along with markers distinguishing cases in which all methods cover the true value from those in which Coracle achieves valid coverage but at least one non-conformal approach fails to do so.

References

- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.
- Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.
- Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.
- Himel Mallick, Anupreet Porwal, Satabdi Saha, Piyali Basak, Vladimir Svetnik, and Erina Paul. An integrated bayesian framework for multi-omics prediction and classification. *Statistics in medicine*, 43(5):983–1002, 2024.
- Saptarshi Roy, Sreya Sarkar, Erina Paul, Piyali Basak, Nengjun Yi, and Himel Mallick. Bayesian cooperative learning for multimodal integration. *bioRxiv*, 2025. doi: 10.1101/2025.10.23.684056. URL <https://www.biorxiv.org/content/early/2025/10/23/2025.10.23.684056>.
- Ina A Stelzer, Mohammad S Ghaemi, Xiaoyuan Han, Kazuo Ando, Julien J Hédou, Dorien Feytaerts, Laura S Peterson, Kristen K Rumer, Eileen S Tsai, Edward A Ganio, et al. Integrated trajectories of the maternal metabolome, proteome, and immunome predict labor onset. *Science Translational Medicine*, 13(592), 2021.