Rank Inversion in PSG Model Selection under Leave-One-Site-Out Evaluation

Hongi Nagaputra, Weijie Sheng, Yuantao Qi, Zhijun Xiao, Yunyi Jiang, Caiyun Ma

School of Software Engineering, Yangzhou University, Yangzhou, China

Abstract

We examined model selection for polysomnographic prediction of future coded cognitive-impairment diagnosis using three systems scored on hidden validation site I0004 after development on three public training sites. Pooled leave-one-site-out (LOSO) age-conditioned area under the receiver operating characteristic curve (AUROC) ranked an N2-epoch system first (0.662), a four-member mixture second (0.601), and a 186-feature tabular system third (0.553). After refitting on all public training data, I0004 ranked the mixture first (0.647), the tabular system second (0.567), and the LOSO-leading N2 system third (0.563). The I0004 AUROC of the LOSO leader was 0.084 lower than that of the mixture. Because only three submitted systems were compared, LOSO folds differed in training size, and I0004 was available for model selection, this is a post hoc rank-reversal of point estimates, not evidence that LOSO generally fails or that the mixture removes site effects.

1. Introduction

Overnight polysomnography (PSG) records neural activity, sleep organization, respiration, oxygenation, and autonomic function. Sleep EEG has been used to predict later cognitive impairment [1], yet a five-cohort analysis found no consistent association between conventional sleep-stage proportions and incident dementia [2]. The endpoint here is therefore clinically plausible but is not a single established sleep biomarker: it is a future *coded diagnosis*, not directly measured cognitive decline or neuropathology [3].

Clinical PSG is also heterogeneous: channel availability, montage, sampling, equipment, and representation vary across recording sites, which has motivated multi-dataset and leave-one-dataset-out evaluation in automated sleep analysis [4]. For prediction models trained on clustered sites, internal-external validation can expose performance heterogeneity that random splits hide [5].

We study a narrower model-selection question. Three candidate systems were compared by pooled LOSO age-conditioned AUROC on the public training sites and were subsequently scored on a hidden validation site. We ask whether the LOSO *point-estimate* ranking agreed with that hidden-site ranking after full-data refitting. We do not use the comparison to identify a mechanism of site robustness.

2. Data and metric

Positive labels require at least two qualifying diagnosis codes, with the first 1–6 years after PSG; negatives require no qualifying diagnosis and sufficient follow-up [3]. We used the public small training set: S0001 (857 records), I0002 (54), and I0006 (192), totaling 1,103 records, of which 84 are positive-labeled. Of 1,103 records, 20 did not meet the positive or negative scoring definition and were excluded from the local pair metric, leaving 1,083 records. Site identity is not an explicit feature; missing channels are zero-filled, so acquisition patterns may still be implicit.

Age-conditioned AUROC is the probability of ranking a positive record above an age-matched negative, with eligible pairs restricted to an age difference of at most two years [3]. For each LOSO fold a model was fitted to two sites and predicted the excluded site. We concatenated the three out-of-site prediction sets and computed one pooled age-conditioned AUROC. The corresponding LOSO training sizes were 246 (hold-out S0001), 1,049 (I0002), and 911 (I0006). Official I0004 scoring refits each submitted entry on all 1,103 public records. I0004 is organizer-held validation, not test; I0007 is unused here. Exact I0004 labels and predictions are unavailable; the organizers state that validation prevalence lies between 5% and 15%. Figure 1 diagrams those roles: each LOSO row trains on two public sites and scores the third; official scoring trains on all three and scores hidden I0004.

3. Systems

System 2125 is an XGBoost model on 186 overnight features (10 demographic, 137 physiological including absolute and relative spectral summaries, 39 sleep-architecture summaries).

System 2334 trains epoch-level models on 30-s win-

dows labeled N2 by the same automated staging used at inference, then averages those probabilities over N2 epochs. N2 is a relatively abundant NREM state that contains spindle activity previously linked to cognitive aging; that motivation does not establish N2 specificity here, and a prospective EEG study reported its largest effects in N3/gamma rather than N2 [1].

System 2353 is the submitted mixture: XGBoost on the 186-vector, a frozen SleepFM representation trained for large-scale PSG disease prediction [6], LightGBM, and ExtraTrees. Nonnegative mixture weights on a 0.1 simplex were taken from inner out-of-fold scores, applied per held-out training site, and not changed after I0004 scores. The members span overnight summaries, a stage-restricted model family, and a pretrained PSG representation; without member ablation we cannot attribute the I0004 result to complementary information. A later packaged variant scored 0.639 on I0004 and did not replace 2353.

4. Results

Table 1 and Figure 2 show the same three systems. LOSO point estimates ranked 2334 > 2353 > 2125. I0004 ranked 2353 > 2125 > 2334: the LOSO leader fell to third and the LOSO runner-up ranked first. The

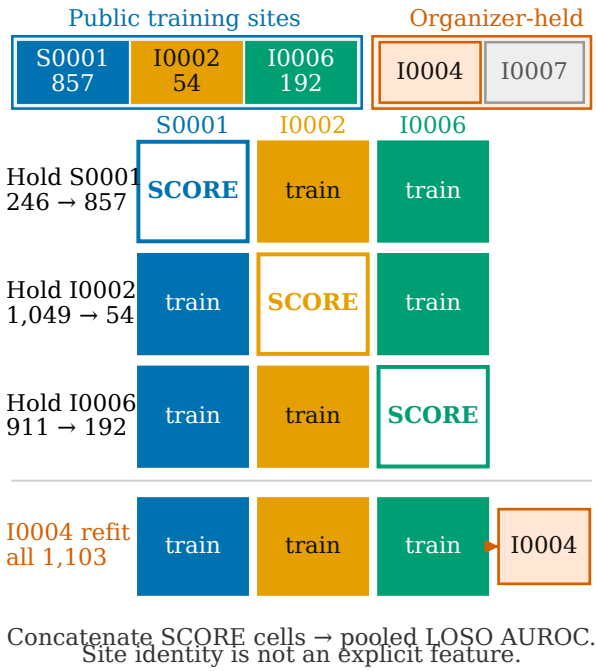

Figure 1. Leave-one-site-out protocol. Filled cells are used in training; hollow SCORE cells are the held-out site. Pooled LOSO concatenates the three SCORE sets. Official scoring refits on all 1,103 public records and scores hidden I0004; I0007 is unused.

change was model-specific: I0004 minus pooled LOSO was +0.014 for 2125, +0.046 for 2353, and −0.099 for 2334. These differences are descriptive and are not estimates of a pure site effect, because LOSO and I0004 also differ in training-set composition and size (Fig. 3).

Table 1. Pooled LOSO versus official I0004 age-conditioned AUROC. Δ is the arithmetic difference of the two reported scores, not an isolated site-transfer effect.

ID	Role	LOSO	I0004	L-rk	I-rk	Δ
2125	186-feature tabular	0.553	0.567	3	2	+0.014
2353	four-member mixture	0.601	0.647	2	1	+0.046
2334	N2-epoch mean	0.662	0.563	1	3	−0.099

On the 2125-versus-2334 comparison, within-site and cross-site leads were similar; dropping cross-site pairs removed about 1.4% of that local AUROC gap. That audit does not include 2353 and does not isolate site transfer after full-data refit. No per-record I0004 predictions are available, so I0004 uncertainty cannot be computed.

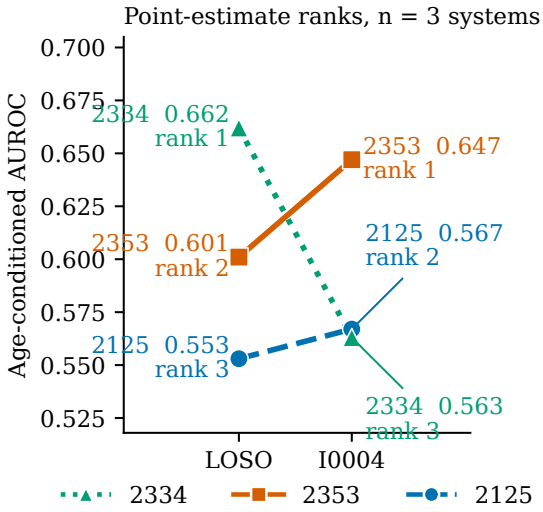

Figure 2. Pooled LOSO versus official I0004 age-conditioned AUROC for three systems. Ranks are of these point estimates only ($n = 3$). The I0004 AUROC of the LOSO leader (2334) was 0.084 lower than that of the mixture (2353).

5. Discussion

The observed rank reversal of point estimates should be read as model-selection instability, not as evidence that LOSO is intrinsically unreliable. These three systems were not designed a priori to test LOSO ranking; the reversal was identified after official I0004 scoring. The three LOSO

folds are not interchangeable transfer tasks (Fig. 3). Holding out S0001 trains on 246 files from I0002 and I0006 and holds out 857 files; holding out I0002 or I0006 trains on 1,049 or 911 files but holds out only 54 or 192. Concatenating those hold-out files places about 78% of the 1,103-file axis on the smallest-train fold. That 78/5/17 split is a file share, not a pair share. Age-conditioned AUROC is then computed on 1,083 eligible records; the 20 records that do not meet the scoring definition are not shown by site. For system 2125, the S0001-out fold had the lowest site-wise age-conditioned AUROC (0.522, versus 0.562 on I0002 and 0.664 on I0006). Comparable fold scores for 2334 and 2353 are not reported here, so the 2125 pattern is not a ranking of the other systems. Official I0004 scoring then refits on all 1,103 public files, including the 857 S0001 files that the S0001-out fold never saw. Site compatibility, learning-curve differences, cross-fold score scaling, and sampling variation therefore remain alternative explanations. Pooled age-conditioned AUROC also ranks some pairs whose scores came from different fitted models, a known caveat of pooled cross-validation AUC [7].

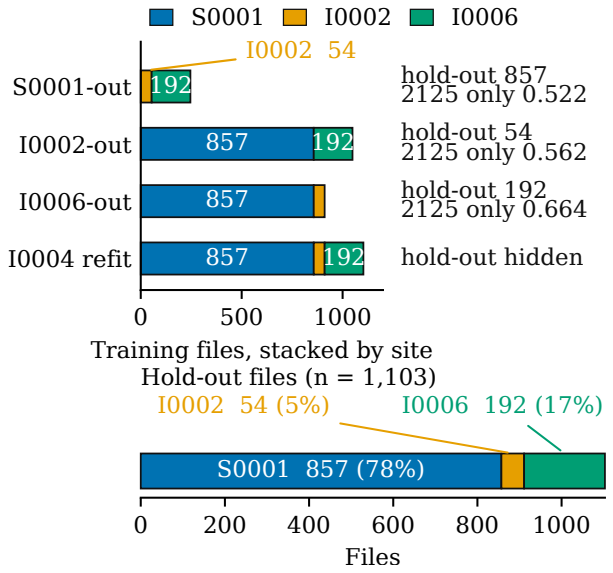

Figure 3. Training composition of each leave-one-site-out fold and of the I0004 refit, stacked by site. Top: records available at fit time (S0001-out trains on 246 files; I0004 trains on all 1,103). Numbers at right are 2125 fold AUROCs only; 2334 and 2353 fold scores are not shown. Bottom: site share of the 1,103 concatenated hold-out files. The pooled age-conditioned AUROC is computed on 1,083 eligible records; the 20 records that do not meet the scoring definition are not shown by site.

For the submission decision we selected 2353 because it had the highest organizer-reported I0004 score among these candidates. Because I0004 scores were observed during iterative submission, we treat I0004 as an organizer-held selection set rather than an untouched estimate of transportability [8]. I0007 remains the independent external check. LOSO provided site-structured internal evidence, but its single pooled point estimate did not identify the highest-scoring I0004 candidate in this three-site sample.

The I0004 result of 2353 does not identify why that score was higher. Future analyses on the public sites can test whether mixture members give complementary predictions, whether missing-channel patterns encode acquisition site, and whether alternative missing-value treatments change site-wise performance; they cannot identify the I0004 mechanism. Site-specific LOSO scores and paired uncertainty on local differences would make the internal ranking more interpretable than a single pooled number [5].

6. Conclusion

Among three submitted PSG risk-ranking systems, pooled LOSO point estimates did not preserve the ordering observed on hidden validation site I0004: the LOSO leader became the lowest-scoring candidate, while the LOSO runner-up ranked first. This observation does not identify a site-robust representation, and the LOSO-to-I0004 differences also include full-data refitting and sampling effects. Model-selection uncertainty across heterogeneous acquisition sites is better examined through site-specific performance and sensitivity analyses than compressed into one pooled internal rank. The selected mixture awaits evaluation on the unused I0007 test site.

Acknowledgments

The authors thank the organizers for official I0004 scoring.

References

[1] Haghayegh S, Herzog R, Bennett DA, Redline S, Yaffe K, Stone KL, Ibáñez A, Hu K. Predicting future risk of developing cognitive impairment using ambulatory sleep EEG: integrating univariate analysis and multivariate information theory approach. Journal of Alzheimers Disease 2025; 108(1_suppl):S105–S117.

[2] Yiallourou S, Baril AA, Wiedner C, Misialek JR, Kline CE, Harrison S, Cannon E, Yang Q, Bernal R, Bisson A, Himali D, Cavuoto M, Weihs A, Beiser A, Gottesman RF, Leng Y, Lopez O, Lutsey PL, Purcell SM, Redline S, Seshadri S, Stone KL, Yaffe K, Ancoli-Israel S, Xiao Q, Vaou EO, Himali JJ, Pase MP. Sleep architecture and dementia risk in adults: an analysis of 5 cohorts from the sleep and dementia consortium. Sleep 2025;48(9):zsaf129.

[3] George B. Moody PhysioNet Challenge. Predicting cognitive impairment from overnight polysomnograms. https://moody-challenge.physionet.org/2026/, 2026. Official ranking metric is age-conditioned AUROC with a two-year age gap.

[4] Perslev M, Darkner S, Kempfner L, Nikolic M, Jennum PJ, Igel C. U-sleep: resilient high-frequency sleep staging. npj Digital Medicine 2021;4:72.

[5] Takada T, Nijman S, Denaxas S, Snell KIE, Uijl A, Nguyen TL, Soares MO, Payne RA, Buchan I, Damen JAA, Moons KGM, Debray TPA. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. Journal of Clinical Epidemiology 2021;137:83–91.

[6] Thapa R, Kjær MR, He B, Covert I, Moore H, Hanif U, Ganjoo G, Westover MB, Jennum P, Brink-Kjær A, Mignot E, Zou J. A multimodal sleep foundation model for disease prediction. Nature Medicine 2026;32(2):752–762.

[7] Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T. A comparison of AUC estimators in small-sample studies. In Proceedings of the 3rd International Workshop on Machine Learning in Systems Biology, volume 8 of Proceedings of Machine Learning Research. 2010; 3–13.

[8] Blum A, Hardt M. The ladder: a reliable leaderboard for machine learning competitions. In Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research. 2015; 1006–1014.

Address for correspondence:

Weijie Sheng
School of Software Engineering
Yangzhou University, Yangzhou, China
wjsheng@yzu.edu.cn