Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Why Variance Metrics Cannot Validate ComBat Harmonization: Evidence from Three-Vendor Mammography Radiomics

XU ZHANG
Hangzhou Xixi Hospital
YUNJIANG LI
Hangzhou Xixi Hospital
RUI FENG
Zhejiang Provincial People's Hospital
QINQIN AI
aiqinqin.zju@foxmail.com

Hangzhou Xixi Hospital
YUHAI XIE
Taihe County People's Hospital
WENBIN NI
Hangzhou Red Cross Hospital

Article

Keywords: radiomics, full-field digital mammography, breast cancer, multi-center, ComBat, harmonization, vendor effects, IBSI, circularity

Posted Date: September 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11019976/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Why Variance Metrics Cannot Validate ComBat Harmonization: Evidence from Three-Vendor Mammography Radiomics

Authors and Affiliations

Xu Zhang[1], Yunjiang Li[2], Rui Feng[3], Qinqin Ai[4]*, Yuhai Xie[5]*, Wenbin Ni[6]*

[1] Department of Ultrasound, Hangzhou Xixi Hospital, No. 2 Hengbu Street, Xihu District, Hangzhou, Zhejiang 310023, P. R. China; Email: dr.xu_zhang@qq.com

[2] Department of Radiology, Hangzhou Xixi Hospital, No. 2 Hengbu Street, Xihu District, Hangzhou, Zhejiang 310023, P. R. China; Email: wshlyj@163.com

[3] Department of Interventional Medicine, Zhejiang Provincial People's Hospital, Hangzhou, Zhejiang, P. R. China; Email: 766734529@qq.com

[4] Department of Hepatology, Hangzhou Xixi Hospital, Affiliated to Zhejiang Chinese Medical University (Hangzhou Sixth People's Hospital), Hangzhou 310023, Zhejiang, China; Email: aiqinqin.zju@foxmail.com

[5] Department of Radiology, Taihe County People's Hospital Affiliated to Wannan Medical University, Fuyang, Anhui, China; Address: No. 21 Jiankang Road, Taihe County, Fuyang City, Anhui Province, 236600, China; Email: xyhdoctor@163.com

[6] Department of Radiology, Hangzhou Red Cross Hospital, Hangzhou, Zhejiang 310003, China; Address: 208 Huan Cheng Dong Road, Gongshu District, Hangzhou (310003), Zhejiang Province, China; Email: niwenbin519@outlook.com

Corresponding author: Qinqin Ai, Yuhai Xie, Wenbin Ni .

Corresponding Author

Qinqin Ai : Department of Hepatology, Hangzhou Xixi Hospital, Affiliated to Zhejiang Chinese Medical University (Hangzhou Sixth People's Hospital), Hangzhou 310023, Zhejiang, China; Email: aiqinqin.zju@foxmail.com; ORCID: 0009-0009-8538-7462.

Yuhai Xie : Department of Radiology, Taihe County People's Hospital Affiliated to Wannan Medical University, Fuyang, Anhui, China; Address: No. 21 Jiankang Road, Taihe County, Fuyang City, Anhui Province, 236600, China; Email: xyhdoctor@163.com.

Wenbin Ni : Department of Radiology, Hangzhou Red Cross Hospital, 208 Huan Cheng Dong Road, Gongshu District, Hangzhou 310003, Zhejiang Province, China. Email: niwenbin519@outlook.com

Keywords

radiomics; full-field digital mammography; breast cancer; multi-center; ComBat; harmonization; vendor effects; IBSI; circularity

Funding

This study was supported by:

- Public Welfare Scientific Research Guidance Project of Agriculture and Social Development of Hangzhou Science and Technology Bureau (No. 20241029Y045)
- Hangzhou Medical and Health Science and Technology Plan Project (A20240074, A20241870)

The funders had no role in study design, data collection, analysis, interpretation, or manuscript writing. No commercial sponsor was involved.

Conflicts of Interest

The authors declare no competing interests.

Ethical Approval

This retrospective study was approved by the Ethics Committee of the project-hosting institution (No. 2024KY141). All procedures involving human participants were in accordance with the 1964 Helsinki Declaration and its later amendments.

Consent to Participate

The requirement for informed consent was waived by the Ethics Committee because the study used de-identified, retrospectively collected data.

Consent to Publish

Not applicable.

Patient and Public Involvement

Patients and/or the public were not involved in the design, conduct, reporting, or dissemination plans of this research.

Data and Code Availability

- **Raw imaging data:** Not publicly available due to institutional privacy restrictions on retrospective clinical data.
- **De-identified radiomic feature table:** Available from the corresponding author on reasonable academic request, subject to a data use agreement.

- **Harmonization and evaluation pipeline (Python, PyRadiomics, SimpleITK, scikit-learn):** GitHub to publication.

Author Contributions (CRediT)

Author	Conceptualization	Data curation	Methodology	Formal analysis	Investigation	Writing - original draft	Writing - review & editing	Visualization	Supervision	Funding acquisition
Xu Zhang	✓	✓	✓	N/A	✓	✓	✓	N/A	✓	✓
Yunjiang Li	✓	N/A	✓	N/A	✓	N/A	✓	N/A	N/A	N/A
Rui Feng	N/A	✓	N/A	N/A	✓	N/A	✓	N/A	N/A	N/A
Yuhai Xie	N/A	N/A	✓	✓	N/A	N/A	✓	✓	N/A	N/A
Qinqin Ai	N/A	✓	N/A	N/A	✓	N/A	✓	N/A	N/A	N/A
Wenbin Ni	N/A	N/A	✓	✓	N/A	N/A	✓	N/A	✓	✓

All authors read and approved the final manuscript.

Why Variance Metrics Cannot Validate ComBat Harmonization: Evidence from Three-Vendor Mammography Radiomics

Running title: Validating ComBat harmonization in mammography radiomics

Keywords: radiomics; full-field digital mammography; breast cancer; multi-center; ComBat; harmonization; vendor effects; IBSI; circularity

Plain language summary

Radiomics turns medical scans into hundreds of measurable numbers that computers analyze to help diagnose disease. When the same scan is taken on machines from different manufacturers, those numbers can differ for reasons unrelated to the patient, a problem called "vendor effects." Researchers routinely fix this with a popular statistical method called ComBat and check success with simple statistical tests. We show those tests are misleading: they measure exactly what ComBat was built to change, so "success" is almost automatic. Using a more demanding test, asking a computer to guess which machine took each scan, we found the manufacturer remained perfectly identifiable even after ComBat, and that the method could erase a real difference in tumor size. Our study gives researchers a better way to judge whether harmonization truly worked, and a caution that "fixing" scanner differences can accidentally remove meaningful biology.

Abstract

ComBat is the standard method for removing scanner- and vendor-related variation from radiomic features, yet its success is conventionally judged by variance statistics (ICC, η^2, Kruskal-Wallis) that ComBat aligns by construction, making such validation nearly tautological. Using expert tumor segmentations from three mammography vendors (n=120; analysis subset n=91) and 59 standardized features, we tested ComBat with evidence independent of the moments it targets. After harmonization a linear classifier could no longer identify the vendor (AUC 0.56) but a leakage-free random forest identified it perfectly (AUC 1.00), and inter-vendor covariance structure persisted (0.259→0.244). Naive ComBat also erased a batch-correlated biological covariate, tumor size ($p=4.0\times10^{-8}$→0.10), which covariate-protected ComBat preserved. Variance-based reporting may therefore give false assurance of cross-vendor comparability. Harmonization adequacy should therefore be certified by independent non-linear separability rather than by the variance statistics ComBat directly targets.

1. Introduction

ComBat, an empirical-Bayes location-and-scale correction adopted from genomics, has become the default method for removing site- and vendor-related variation from radiomic features [1]. Its success is

conventionally certified by first- and second-order variance statistics (ICC, η^2, and the Kruskal-Wallis test) that quantify exactly the location and scale ComBat was built to align. A method cannot be validated by the quantities it was built to change: reporting that η^2 approaches zero after ComBat largely restates the correction itself rather than independently testing it[2].

This matters because multi-center radiomics depends on features being comparable across scanners, and radiomic models developed in one setting frequently degrade when applied elsewhere, as a cross-national mammography validation recently demonstrated [3]. Full-field digital mammography (FFDM) is the dominant breast-cancer screening modality, radiomics is increasingly applied to it, and cross-vendor feature variability is substantial: texture features differ materially between mammography manufacturers [4-6], and in multivendor FFDM only a subset of features is reproducible across vendors [7]. Cross-vendor harmonization nonetheless remains far less studied in FFDM than in PET/CT or MRI [8].

That ComBat is not universally applicable is already documented: it requires generalized variants when feature distributions are multi-modal or imaging parameters are unknown [9], it should not be applied when its assumptions are unmet [2], and only a subset of features is genuinely harmonizable [10]. What remains untested is whether vendor information survives in the joint, higher-order feature distribution that ComBat does not align; and whether the standard variance metrics can independently certify harmonization success.

Here, using expert tumor segmentations from three FFDM vendors, we formalize this circularity and evaluate harmonization with five lines of evidence that do not share ComBat's target moments: multivariable vendor separability assessed with a leakage-free non-linear classifier, low-dimensional geometry, biological-information preservation, downstream cross-vendor transfer, and higher-order covariance structure. We show that vendor remains perfectly identifiable after harmonization while a batch-correlated biological covariate can be erased, and we argue that harmonization adequacy must be judged by independent non-linear evidence rather than by the variance statistics ComBat directly targets.

2. Materials and Methods

2.1 Data source and cohort

This harmonization study analyzes the same three-center, three-vendor full-field digital mammography (FFDM) collection originally assembled to predict axillary lymph-node metastasis in invasive breast cancer [11]. The parent cohort enrolled 728 pathologically confirmed invasive-breast-cancer patients across three institutions, each with a distinct vendor: Center A (UIH uMammo 590i; Jul 2016-Mar 2022); Center B (HOLOGIC Selenia Dimensions; Dec 2016-Dec 2021); Center C (GE Senographe Essential; Jan 2018-Jun 2021). The study was approved by the ethics committee of the

project-hosting institution (No. 2024KY141), and individual informed consent was waived under the retrospective design.

Inclusion: (i) histopathologically confirmed invasive carcinoma; (ii) no prior biopsy, surgery, or neoadjuvant chemotherapy before FFDM; (iii) pathological confirmation within 2 weeks after FFDM. Exclusion: (i) indistinct/incompletely depicted lesion precluding ROI delineation; (ii) history of another malignancy.

For the present analysis we used n=120 cases (40/center; image dimensions: A 3530×2812 px @0.085 mm/px, B 4096×3328 @0.070, C 3062×2394 @0.100) with valid tumor ROIs and extractable features; after excluding 29 near-whole-breast GE masks, the primary clean subset was n=91 (UIH 40, HOLOGIC 40, GE 11). Three-center patient IDs were non-overlapping, eliminating cross-center leakage.

2.2 Expert ROI delineation and image-ROI alignment

Tumor ROIs were manually delineated by two board-certified radiologists (readers A and B), each with ≥5 years of breast-imaging experience, on the CC or MLO view showing the largest, clearest lesion; reader A repeated delineation after 2 weeks for intra-observer assessment. The parent cohort established inter-/intra-observer reproducibility (ICC > 0.85 for retained features) [11], so the delineation followed a validated protocol.

Expert ROIs were NIfTI masks in `(Cols, Rows)` layout, whereas DICOM pixel arrays are `(Rows, Cols)`, requiring axis transposition. For each center we evaluated eight candidate orientations by maximizing breast-tissue overlap, confirming transposition (T; overlap = 1.0), cached per center.

2.3 Preprocessing

Images were resampled to isotropic 0.20 mm (SimpleITK; linear intensity / nearest-neighbor mask). FFDM lacks Hounsfield units, so intensity was 1-99 percentile clipped. For Center C (four views/case), only the non-empty mask slice was retained to avoid lesion-absent views; cropping was anchored at the ROI bounding-box center (fixed 1024×1024) to preserve peripheral lesions.

2.4 Radiomics feature extraction

Features were extracted with PyRadiomics [12], using Image Biomarker Standardization Initiative (IBSI) feature definitions [13]. FFDM ROIs are single-slice 2D, so the 3D `shape` class is undefined; we used `shape2D` (9) plus first-order (18), GLCM (24), GLRLM (16), GLSZM (16), GLDM (14), NGTDM (5) (102 features/sample) computed on a tight crop (ROI bounding box + 32 px) per the IBSI volume-of-interest definition. Standard near-zero-variance filtering removed 43 (42.2%) features (e.g., GLCM Correlation ≡ 1.0), leaving **59 analysis-ready features**.

2.5 Quality control and conventional variance-based harmonization assessment

Because each center has a distinct vendor (UIH/HOLOGIC/GE), center and vendor are perfectly collinear, so vendor was the batch variable. Cross-vendor batch effects were corrected with empirical-Bayes ComBat [1], the primary harmonization versus per-vendor Z-score.

Conventional harmonization adequacy was quantified by (i) the one-way random intraclass correlation coefficient ICC(1,1) [14]; (ii) η^2, the proportion of total feature variance attributable to vendor; and (iii) the Kruskal-Wallis test for inter-vendor difference per feature. These are the metrics used throughout the radiomics harmonization literature [2]. **Methodological caveat (see §3.3 and §4): because ICC, η^2, and Kruskal-Wallis are themselves first/second-order variance statistics that ComBat aligns, they cannot independently test whether vendor information was removed.** They are reported for literature continuity and as the motivation for the independent evaluation in §2.6.

2.6 Independent, non-circular evaluation of cross-vendor harmonization (E1-E5)

To overcome the circularity above, we evaluated harmonization with five evidence lines that do **not** share ComBat's target moments.

- **E1: Multivariable vendor separability.** We trained logistic regression (linear) and random forest (non-linear) classifiers to predict vendor from the 59 features, under 5-fold stratified cross-validation. We report accuracy, balanced accuracy, and macro one-vs-rest AUC (random baseline: accuracy 1/3, AUC 0.5). The random forest is decisive: ComBat leaves covariance/higher-order structure intact, exploitable by a non-linear but not a linear model.
- **E1′: Leakage-free separability.** To exclude transductive leakage, ComBat was fitted on the training folds only and then applied to the held-out fold before classifier training/testing (fit/apply separation).
- **E2: PCA vendor silhouette.** Features were reduced by PCA; the silhouette coefficient of vendor membership was computed in the PC1-PC2 plane and in the full feature space.
- **E3: Biological-information preservation.** We measured within-vendor rank preservation, pooled rank preservation, feature-correlation-structure preservation (Spearman correlation between feature-feature correlation matrices), and sample-distance-structure preservation. As a biological anchor we used `original_shape2D_PixelSurface` (tumor pixel surface, a vendor-scale-independent geometric truth) and tested inter-vendor difference by Kruskal-Wallis before and after harmonization.
- **E4: Downstream cross-vendor transfer.** For each held-out vendor, texture features were regressed onto the geometric truth log(tumor pixel surface area at 0.20 mm physical spacing) using Ridge regression (leave-one-vendor-out). Because Ridge diverges under cross-vendor extrapolation ($R^2 < -1$, flagged "exploded"), **primary metrics were Spearman ρ and MAE**, not R^2.
- **E5: Higher-order (covariance) structure.** For each vendor pair we computed the feature-feature correlation matrix; inter-vendor structure was summarized by Spearman correlation between flattened matrices and mean absolute difference, before and after harmonization. Persistence of a vendor-specific difference indicates retained higher-order structure.

Covariate-protected ComBat. Following Orlhac et al. [2], tumor-size tertiles were biological covariates *preserved* in the ComBat design matrix. Feasibility needs vendor-covariate overlap; we ran a covariate-overlap diagnostic (vendor × size-tertile counts) and flagged structural zeros.

2.7 Tumor segmentation sub-study (direction B; details in Supplementary Methods S1)

As a complementary demonstration that tumor boundaries are themselves vendor-sensitive, we trained a 2D U-Net for mammographic mass segmentation on expert ROIs from Centers A (UIH) and B (HOLOGIC). Center C (GE) was excluded from training because its masks were not mass-equivalent (29/40 near-whole-breast annotations, different intensity convention). Full data partitioning, preprocessing, architecture, and training are reported in Supplementary Methods S1.

3. Results

3.1 Pipeline and data characterization (real)

- Source cohort: the three-center FFDM archive underlying [11] yielded 1252 pathologically confirmed malignancies with available imaging across 3 centers / 3 vendors (the parent ALNM-prediction study [11] reported 728 enrolled patients under its own eligibility criteria); from this archive, 120 cases with valid tumor ROIs were used for the present analysis.
- Validation subset with complete imaging + ROI: **120 cases (40/center)**, saved as npz (image+mask). After excluding 29 near-whole-breast GE masks, the **clean subset comprised 91 cases** (UIH 40, HOLOGIC 40, GE 11).
- Alignment: transposition (T) confirmed per center (breast-tissue overlap = 1.0).
- Feature table: **120 samples × 102 IBSI-compliant numeric features**, zero NaN/Inf, zero duplicate IDs, zero all-missing columns; **59 analysis-ready features** after near-zero-variance filtering.
- Segmentation set QC: 120 npz, 0 empty masks, 0 read failures.

3.2 Conventional variance-based harmonization assessment

Before harmonization, vendor was the dominant source of feature variance. The median ICC(1,1) was **0.964**, the median η^2 was **0.583**, and **53/59 (89.8%)** features showed significant inter-vendor differences by Kruskal-Wallis ($p<0.05$).

After ComBat, the median η^2 fell to **0.000** and the median ICC(1,1) to **0.000**, while the proportion of vendor-significant features dropped to **21/59 (35.6%)**. Per-vendor Z-score also reduced median η^2 to ≈0 but drove ICC(1,1) to **−0.500**, indicating over-standardization that destroyed cross-vendor consistency (negative reliability).

A caveat on interpretation. These metrics quantify only first/second-order variance (precisely the moments ComBat aligns), so their reduction is expected by construction (see §4). They are reported for literature continuity; the independent evaluation in §3.3 is the basis for our conclusions.

Sensitivity analysis. Re-running ComBat on the clean subset (n=91) reproduced post-harmonization results exactly (η^2, ICC→0.000; KW-significant 35.6%), while pre-harmonization η^2 fell 0.583→0.341; the headline pre-harmonization vendor effect was partly amplified by GE annotation inconsistency, but the conclusion that ComBat removes the linear batch effect is preserved.

3.3 Independent evaluation of cross-vendor harmonization (E1-E5)

The clean subset (n=91) is the primary analysis; the full set (n=120) is reported for confirmation.

E1: Vendor remains perfectly separable by a non-linear model after ComBat. On raw features, vendor was identifiable by random forest at AUC 0.94 (accuracy 0.79) and by logistic regression at AUC 0.92. After naive ComBat, the *linear* logistic AUC dropped to **0.56** (near chance, suggesting success by conventional reasoning) while the *non-linear* random forest AUC rose to **1.00** (accuracy 0.99) vendor became **perfectly** separable (Table 1). Z-score gave the same pattern (RF AUC 0.999). This demonstrates that ComBat removes linear vendor signal while leaving non-linear/higher-order vendor structure fully intact. The full set reproduced this exactly (raw RF AUC 0.96; ComBat-naive RF AUC 1.00, accuracy 0.99).

Table 1. E1 vendor separability (5-fold CV; clean n=91; full n=120). Chance: accuracy 0.33, AUC 0.50.

Variant	logreg AUC	RF AUC	RF accuracy
****Clean (n=91)****			
Raw	0.922	0.940	0.791
ComBat (naive)	0.556	**1.000**	0.989
ComBat (size-protected)	0.562	0.999	0.978
Z-score	0.463	0.999	0.912
****Full (n=120)****			
Raw	0.946	0.959	0.833
ComBat (naive)	0.559	**1.000**	0.992
ComBat (size-protected)	0.829	1.000	1.000
Z-score	0.494	1.000	0.992

Fig. 1 visualizes the E1 separability pattern across variants and shows that the linear-classifier drop paired with the non-linear-classifier rise is reproducible across both the clean and full sets (`mammo_pipeline/radiomics_features/harmonized/figures_eval/separability_{clean,full}.png`).

E1′: Leakage-free confirmation. Fitting ComBat on training folds only and applying to the held-out fold yielded the same conclusion: clean set ComBat-naive logreg AUC 0.78 / RF AUC 1.00; size-protected logreg 0.87 / RF 1.00. Full set: size-protected logreg 0.93 / RF 1.00. The result is therefore not an artifact of transductive fitting.

E2: Vendor clustering dissolves only in the top principal components. Raw features showed a positive vendor silhouette in PC1-PC2 (+0.13) and full space (+0.12), indicating vendor-driven clustering. After ComBat the silhouette became slightly negative (PC1-PC2 −0.08; full −0.01), i.e., vendor clusters dissolved in the dominant low-dimensional projection; consistent with removal of the largest (linear) variance component, but compatible with retained structure in higher dimensions (as E1 shows). The PCA projection of the clean set is shown in Fig. 2 (`Fig2_publication.png`); the GE cluster (green) is the most clearly separable in the raw panel and is largely absorbed into a common scatter after any of the three harmonization variants.

E3: Naive ComBat erases a batch-correlated biological covariate. Using tumor pixel surface as an anchor, the inter-vendor Kruskal-Wallis p fell from **4.0×10^{-8} (raw, significant)** to **0.10 (ComBat-naive, non-significant)**; the cross-vendor tumor-size difference was erased and became non-significant. Z-score similarly erased it (p=0.72). **Covariate-protected ComBat preserved it ($p=5.8\times10^{-3}$, retained=True)**, consistent with removal of a batch-correlated biological covariate (tumor size) rather than residual batch. We note the cross-vendor tumor-size difference is partly shaped by the cohort's vendor-size structural zeros (GE contributes only large tumors; §2.6), so this anchor demonstrates that naive ComBat can remove a batch-correlated biological covariate, the same risk that motivates covariate-protected ComBat [2]. The full set reproduced this (raw $p=1.25\times10^{-18}$ → naive 0.20 erased; size-protected 9.5×10^{-17} retained).

Covariate-overlap diagnostic (why size-protected ComBat is infeasible here). GE occupied only the largest tumor-size tertile (clean: 0/0/11; full: 0/0/40), while UIH/HOLOGIC lacked the largest tertile (full: 15/25/0 and 25/15/0). Structural zero cells make batch and biological covariate near-collinear, so covariate-protected ComBat is statistically unstable in this cohort despite preserving biology in principle.

E4: Downstream transfer improves, but partly by removing biology. Mean Spearman ρ for texture→log(area) transfer rose from 0.32 (raw) to 0.75 (ComBat-naive) and 0.89 (Z-score); MAE fell 220→17 (ComBat) and 4.0 (Z-score). R^2 diverged for 2/3 vendors under cross-vendor extrapolation (estimator artifact, not a harmonization metric). Crucially, Z-score (which gave the best transfer) is the same method that erased the size anchor (E3) and over-standardized (ICC→−0.5, §3.2); its gain partly reflects removal of the biological variation we wish to retain. Naive ComBat sits between.

E5: Higher-order covariance structure persists after ComBat. Mean inter-vendor feature-correlation-matrix difference was 0.259 raw vs **0.244 after ComBat (clean)** and 0.267 vs 0.251 (full); vendor-specific structure remained (Table 2), mechanistically explaining the retained random-forest separability in E1: ComBat does not align the joint covariance structure.

Table 2. E5 inter-vendor covariance-structure difference (mean |Δ| of pairwise feature-correlation matrices). "Vendor-specific retained = True" means ComBat did not remove higher-order vendor structure.

Variant	Clean (raw→post)	Full (raw→post)	Vendor-specific retained
ComBat (naive)	0.259 → 0.244	0.267 → 0.251	True
ComBat (size-protected)	0.259 → 0.280	0.267 → 0.322	True
Z-score	0.259 → 0.265	0.267 → 0.273	True

Synthesis. Conventional metrics (η^2/ICC/KW) report vendor-effect elimination, but they are the same first/second-order moments ComBat aligns; evaluating ComBat with them is near-tautological. The independent, non-linear separability (E1/E1′) and covariance-structure diagnostics (E5) show vendor information is in fact retained, and E3 shows naive ComBat can erase a batch-correlated biological covariate. Harmonization adequacy must therefore be judged by independent evidence, not by the variance statistics ComBat directly targets.

3.4 Tumor segmentation performance (direction B)

The A+B-trained U-Net achieved mean Dice 0.678 (5-fold CV, n=78), 0.704 on center-stratified hold-out (n=16), and 0.649 on the single clean-mass GE probe (n=1), confirming partial cross-vendor transfer despite the GE intensity shift; full metrics are in Supplementary Table S1. With direction D, this shows radiomics features and pixel-level segmentation are both vendor-sensitive yet partially transferable once vendor-inconsistent annotations are removed.

4. Discussion

Our central finding is methodological and builds directly on the recognized limitations of ComBat rather than contradicting them. ComBat removes the linear (location/scale) vendor component, as the field's standard variance metrics confirm; but those metrics are themselves first/second-order statistics: the exact moments ComBat aligns. Judging ComBat by them is circular: a near-zero η^2 after ComBat is expected by construction, not evidence that vendor information is gone. That ComBat should not be applied when its assumptions are unmet has been documented, but the circularity arising from the choice of evaluation metric has not, to our knowledge, been formalized as a property of the evaluation design itself. This is consequential because the field's own harmonization guide certifies success with ICC and η^2 [2], i.e., with the very moments ComBat aligns, so the circularity is embedded in routine practice rather than a theoretical aside.

Testing vendor identifiability with a non-linear classifier that exploits the joint feature distribution ComBat does not align, vendor became perfectly separable (AUC 1.00, leakage-free confirmed), mechanistically explained by E5: the inter-vendor feature-correlation difference persisted (0.259→0.244), i.e., ComBat does not align higher-order structure. Prior empirical evaluations indicate that ComBat's

benefit is task- and implementation-dependent rather than automatic: in multicenter PET/CT, the choice of ComBat strategy materially changed prognostic performance [8]; in MRI radiomics, its effect on machine-learning tissue classification differed between ComBat variants [15]; and in cardiac MRI, ComBat improved feature robustness to acquisition parameters [16]. We extend this work by quantifying how much vendor information survives in the higher-order distribution and by providing an operational test.

Three practice implications follow. First, harmonization adequacy should be reported with an independent non-linear vendor-separability gate (for example, a leakage-free random-forest vendor classifier), not only ICC/η^2; persistent non-linear separability means features are not yet vendor-comparable; consistent with critical appraisals that identify reproducibility as the central obstacle to oncologic translation [17]. Second, naive ComBat can erase biological signal: the inter-vendor tumor-pixel-surface difference, a batch-correlated biological covariate, became non-significant after naive ComBat, whereas covariate-protected ComBat preserved it. This is a material risk for size-prognostic models, because mammography-based radiomics is increasingly used to predict axillary nodal status [18]. Third, covariate-protected ComBat is correct in principle [2] but data-design-dependent: our structural zeros (GE only large tumors) made protection unstable, so multi-vendor studies should engineer vendor-covariate overlap a design lesson not previously illustrated with a failure case.

Why does ComBat nonetheless appear to help downstream? In E4, texture→area transfer improved after ComBat and most after Z-score; yet Z-score is the method that erased the size anchor and over-standardized (ICC→−0.5), so its gain partly reflects removal of the size variation we aimed to retain. Marginal-moment harmonization thus trades biological fidelity for apparent transportability; a trade that should be explicit. Comparable tensions are reported elsewhere: institution-related variability in MRI radiomics persisted despite harmonization [19], and harmonization improved prediction for some endpoints while leaving residual variability for others [20].

Our results sit within, and do not challenge, the harmonization literature; they add a testable adequacy criterion for the understudied FFDM setting. Several methodological developments bear directly on the residual we describe: generalized ComBat variants extend the approach to multi-modal distributions [9], and nested ComBat addresses multiple batch effects [21]; how batches are defined at all materially affects harmonization and downstream prognostication [22]; and deep-learning-based image harmonization provides a distribution-matching alternative that acts before feature extraction [23]. Whether any of these removes the higher-order vendor structure we identify remains to be tested. The Graphical Abstract (Fig. 3) contrasts conventional variance-metric with independent non-linear assessment on the same data.

More broadly, whether radiomic models generalize at all remains an open question [24], and determining which features are reproducible across variable acquisition settings is an active area of

investigation [25], including phantom studies of scanner-to-scanner robustness [26]. Mammography is no exception: deep-learning and handcrafted radiomics both improve lesion classification on mammographic images [27], yet cross-vendor comparability of the underlying features is rarely assessed. Standardized workflow and reporting guidance has converged on similar recommendations [28].

Limitations. (i) Retrospective, single-lesion ROIs; (ii) FFDM is 2D, so 3D shape features are unavailable and ComBat operates on 59 2D-derived features; (iii) the clean GE subset is small (n=11), mitigated by full-set confirmation (n=120); (iv) absence of clinical endpoints means we demonstrate harmonization *behavior*, not predictive gain; direction A (prediction modeling) awaits endpoint labels; (v) structural zeros limited covariate-protected ComBat; (vi) we evaluated ComBat and Z-score, not newer generalized or distribution-matching harmonization, which may address the higher-order residual and warrant separate study. (vii) Vendor and center are perfectly collinear in this cohort (one vendor per center), so "vendor-identifiable" denotes the combined center/vendor imaging signature; isolating a pure vendor effect would require multiple vendors within at least one center. (viii) We report point estimates without confidence intervals and did not apply multiple-comparison correction to the per-feature Kruskal-Wallis counts, which we present as descriptive rather than confirmatory evidence.

5. Conclusion

ComBat removes linear cross-vendor effects in FFDM radiomics but leaves center/vendor-identifiable higher-order structure intact (vendor and center are perfectly collinear in this cohort) and can erase batch-correlated biological variation when applied naively. Harmonization adequacy must be judged by independent non-linear separability and covariance-structure diagnostics, not by the variance statistics ComBat directly targets. Covariate-protected ComBat is preferred in principle but requires overlap between batch and biological covariates that must be engineered at study design.

Data Availability

The radiomics feature matrices and ROI masks underlying this study are available from the corresponding author upon reasonable request. Source FFDM images cannot be shared publicly owing to the retrospective ethical approval (ethics committee of the project-hosting institution, No. 2024KY141; individual consent waived) and institutional data-sharing restrictions; access may be granted for academic collaboration under a data-use agreement. Inter-reader reproducibility statistics derive from the parent cohort.

Code Availability

The harmonization and evaluation pipeline (Python 3, PyRadiomics, SimpleITK, scikit-learn), including the empirical-Bayes ComBat implementation, the leakage-free train/apply separability test, the covariance-structure diagnostics, and the figure-generation scripts, is available at GitHub publication.

Reporting Compliance

This observational, multi-center methodological study followed the STROBE guideline for observational research. Because the study trains classifier models (logistic regression, random forest) to assess vendor separability, we additionally considered TRIPOD principles for model evaluation, including leakage-free cross-validation and transparent specification of training/test partitioning. The completed Nature Portfolio Reporting Summary accompanies this submission.

References

1. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037. PMID 16632515.

2. Orlhac F, Eertink JJ, Cottereau AS, Zijlstra JM, et al. A guide to ComBat harmonization of imaging biomarkers in multicenter studies. *J Nucl Med*. 2022;63(2):172-179. doi:10.2967/jnumed.121.262464. PMID 34531263.

3. Hou R, Grimm LJ, Marks JR, King LM, et al. Cross-national radiomics validation using mammography to predict occult invasion in ductal carcinoma in situ. *Radiology*. 2025;317(2):e243739. doi:10.1148/radiol.243739. PMID 41251557.

4. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229-263. doi:10.3322/caac.21834. PMID 38572751.

5. Winham SJ, McCarthy AM, Scott CG, Gastounioti A, et al. Radiomic parenchymal phenotypes of breast texture from mammography and association with risk of breast cancer. *Radiology*. 2025;315(2):e240281. doi:10.1148/radiol.240281. PMID 40358450.

6. Mendel KR, Li H, Lan L, Cahill CM, et al. Quantitative texture analysis: robustness of radiomics across two digital mammography manufacturers' systems. *J Med Imaging (Bellingham)*. 2018;5(1):011002. doi:10.1117/1.JMI.5.1.011002. PMID 28948196.

7. Robinson K, Li H, Lan L, Schacht D, et al. Radiomics robustness assessment and classification evaluation: a two-stage method demonstrated on multivendor FFDM. *Med Phys*. 2019;46(5):2145-2156. doi:10.1002/mp.13455. PMID 30802972.

8. Xu H, Abdallah N, Marion JM, Chauvet P, et al. Radiomics prognostic analysis of PET/CT images in a multicenter head and neck cancer cohort: investigating ComBat strategies, sub-volume characterization, and automatic segmentation. *Eur J Nucl Med Mol Imaging*. 2023;50(6):1720-1734. doi:10.1007/s00259-023-06118-2. PMID 36690882.

9. Horng H, Singh A, Yousefi B, Cohen EA, et al. Generalized ComBat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. *Sci Rep*. 2022;12(1):4493. doi:10.1038/s41598-022-08412-9. PMID 35296726.

10. Ibrahim A, Refaee T, Leijenaar RTH, Primakov S, et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS One*. 2021;16(5):e0251147. doi:10.1371/journal.pone.0251147. PMID 33961646.

11. Xie Y, Ma P, Wang X, et al. Radiomics nomogram based on digital mammography for predicting axillary lymph node metastasis of invasive breast cancer: a multicenter study. *Radiol Pract*. 2024;39(1):31-36. doi:10.13609/j.cnki.1000-0313.2024.01.006. [Chinese; verifiable via CNKI/Wanfang]

12. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339. PMID 29092951.

13. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338. doi:10.1148/radiol.2020191145. PMID 32154773.

14. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012. PMID 27330520.

15. Leithner D, Nevin RB, Gibbs P, Weber M, et al. ComBat harmonization for MRI radiomics: impact on nonbinary tissue classification by machine learning. *Invest Radiol*. 2023;58(9):697-701. doi:10.1097/RLI.0000000000000970. PMID 36897814.

16. Priya S, Dhruba DD, Sorensen E, Aher PY, et al. ComBat harmonization of myocardial radiomic features sensitive to cardiac MRI acquisition parameters. *Radiol Cardiothorac Imaging*. 2023;5(4):e220312. doi:10.1148/ryct.220312. PMID 37693205.

17. Horvat N, Papanikolaou N, Koh DM. Radiomics beyond the hype: a critical evaluation toward oncologic clinical use. *Radiol Artif Intell*. 2024;6(4):e230437. doi:10.1148/ryai.230437. PMID 38717290.

18. Wang Q, Lin Y, Ding C, Guan W, et al. Multi-modality radiomics model predicts axillary lymph node metastasis of breast cancer using MRI and mammography. *Eur Radiol*. 2024;34(9):6121-6131. doi:10.1007/s00330-024-10638-2. PMID 38337068.

19. Morelli L, Paganelli C, Marvaso G, Parrella G, et al. Addressing intra- and inter-institution variability of a radiomic framework based on apparent diffusion coefficient in prostate cancer. *Med Phys*. 2024;51(11):8096-8107. doi:10.1002/mp.17355. PMID 39172115.

20. Khajetash B, Hajianfar G, Talebi A, Mahdavi SR, et al. Impact of harmonization on predicting complications in head and neck cancer after radiotherapy using MRI radiomics and machine learning techniques. *Med Phys*. 2025;52(6):5091-5103. doi:10.1002/mp.17793. PMID 40162683.

21. Ziasaeedi N, Lemaréchal Y, Agharazii M, Manem VSK, et al. Radiomics-based kidney lesion classification: mitigating batch effect with nested ComBat harmonization. *Med Phys*. 2025;52(9):e18070. doi:10.1002/mp.18070. PMID 40891042.

22. Wu H, Liu X, Peng L, Yang Y, et al. Optimal batch determination for improved harmonization and prognostication of multi-center PET/CT radiomics feature in head and neck cancer. *Phys Med Biol*. 2023;68(22):225014. doi:10.1088/1361-6560/ad03d1. PMID 37844604.

23. Du D, Lv W, Lv J, Chen X, et al. Deep learning-based harmonization of CT reconstruction kernels towards improved clinical task performance. *Eur Radiol*. 2023;33(4):2426-2438. doi:10.1007/s00330-022-09229-w. PMID 36355196.

24. Benyoucef R, Goubet M, Barrat A, Chauveau B, et al. Do MRI radiomic models truly generalize? External validation of three studies in parotid lesion characterization. *Eur Radiol*. 2026;36(8):6376-6388. doi:10.1007/s00330-026-12512-9. PMID 41944835.

25. Barabino E, Tosques M, Battaglia AP, Ficarra G, et al. Hunting for signal in the noise: identifying robust CT radiomic features across variable scanning settings. *Eur Radiol*. 2026;36(5):3733-3742. doi:10.1007/s00330-025-12215-7. PMID 41417126.

26. Zhu L, Dong H, Sun J, Wang L, et al. Robustness of radiomics among photon-counting detector CT and dual-energy CT systems: a texture phantom study. *Eur Radiol*. 2025;35(2):871-884. doi:10.1007/s00330-024-10976-1. PMID 39048741.

27. Beuque MPL, Lobbes MBI, van Wijk Y, Widaatalla Y, et al. Combining deep learning and handcrafted radiomics for classification of suspicious lesions on contrast-enhanced mammograms. *Radiology*. 2023;307(5):e221843. doi:10.1148/radiol.221843. PMID 37338353.

28. Avanzo M, Soda P, Bertolini M, Bettinelli A, et al. Robust radiomics: a review of guidelines for radiomics in medical imaging. *Front Radiol*. 2025;5:1701110. doi:10.3389/fradi.2025.1701110. PMID 41602202.

Figure Legends

Fig. 1. Vendor-identification macro-OVR AUC across harmonization variants (E1). Logistic regression (blue) and random forest (red) 5-fold stratified cross-validation on the 59 analysis-ready features; chance AUC = 0.5. After ComBat (naive or size-protected) and Z-score, linear separability drops toward chance while non-linear separability rises to or remains at 1.0, indicating that vendor information survives in the joint feature distribution ComBat does not align.

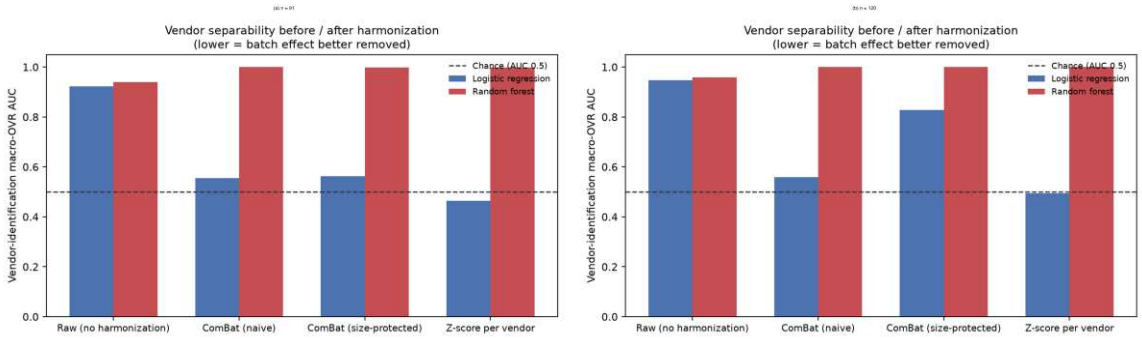


Fig. 2. PCA projection (PC1 vs PC2) of the 59 features colored by vendor across harmonization variants (E2). Clean set, n=91. The GE cluster (green) is the most clearly separable in the raw panel and is largely absorbed into a common scatter after any of the three harmonization variants, consistent with removal of the dominant (linear) variance component.

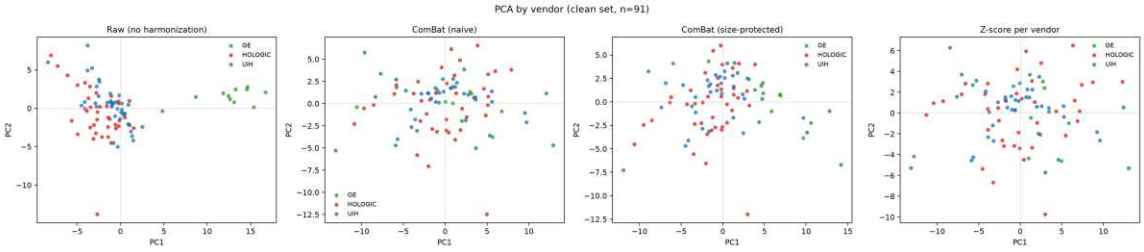
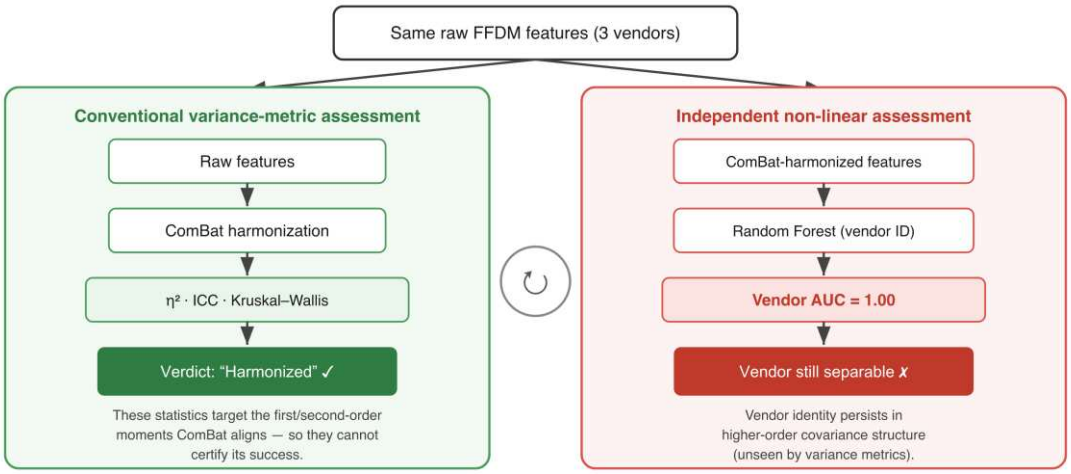


Fig. 3. Graphical abstract: the circularity of ComBat evaluation. Two-panel schematic contrasting the conventional variance-metric assessment (left; η^2, ICC, Kruskal-Wallis report success because they target the moments ComBat aligns) with the independent non-linear separability assessment (right; random-forest vendor AUC = 1.00 reveals that vendor information persists in the higher-order structure).

The Circularity of ComBat Evaluation

Same data · Different tests · Opposite conclusions



Implication: ComBat is judged "successful" by the very variance statistics (η^2, ICC, KW) it was built to align. Harmonization quality should be gated by an independent, non-linear vendor-separability test (here RF AUC = 1.00) — not by the metrics ComBat itself targets.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- npjComBatFFDMSupplementary.docx