A DECODABILITY-GROUNDED AUDIT OF CLINICAL PREDICTION CLAIMS IN PPG FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

What does held-out clinical prediction from photoplethysmography (PPG) reveal about physiology? We audit frozen representations through concept decodability, covariate-adjusted and incremental prediction, and post-erasure predictability after probe refitting, with encounter-matched ECG comparisons. Across four encoders and $37{,}784$ MC-MED encounters, AnyPPG decodes pulse-derived concepts well. For potassium on $6{,}348$ matched encounters, its correlation falls from 0.21 to 0.05 after age and renal adjustment, whereas ECGFounder retains $0.32 \rightarrow 0.27$ (paired adjusted difference $+0.22$; 95% CI $[0.19, 0.26]$). Creatinine likewise attenuates under target-derivative-free adjustment. On a separate $5{,}160$-encounter complete-case cohort, adding ECGFounder to the same age, estimated glomerular filtration rate (eGFR), and blood urea nitrogen (BUN) baseline raises potassium correlation by about 0.087 under ridge and nonlinear predictors; direct AnyPPG increments have confidence intervals including zero. Erasure loss gives the opposite modality ordering from interval decodability: after rank-five subspace removal and refitting, AnyPPG loses 0.0156 in adjusted correlation versus ECGFounder's 0.0081. AnyPPG starts from a much lower adjusted correlation; its loss is near the heart-rate control. Together, these results identify adjustment-sensitive PPG potassium prediction and stronger retained ECG association, while decodability and erasure capture complementary aspects of the representations.

1 INTRODUCTION

Photoplethysmography (PPG), the optical pulse signal recorded by smartwatches and pulse oximeters, is being studied with foundation encoders (Pillai et al., 2025; Saha et al., 2025; Nie et al., 2026). Other PPG work examines hemoglobin measurement and blood-glucose estimation (Abuzairi et al., 2024; Lombardi et al., 2024). Related ECG studies report potassium and multi-laboratory prediction (Galloway et al., 2019; Chiu et al., 2024; Lopez Alcaraz & Strodthoff, 2025). What evidence supports a physiological interpretation of a PPG prediction beyond held-out performance?

Potassium illustrates the distinction. A PPG score may reflect potassium-related physiology, patient characteristics such as age or renal status, or both. Cohort-level risk prediction, agreement with a potassium reference, and tracking within-person changes call for different evidence. Shortcut learning can affect medical predictions: hospital-source information and pneumonia prevalence influenced chest-radiograph generalization (Geirhos et al., 2020; Zech et al., 2018). In PPG, patient characteristics and target-related physiology share the waveform; an audit must examine both concept accessibility and predictive comparisons.

Figure 1 shows the shared setup and three audit measurements. **Concept decodability:** how well can selected physiological reference labels be decoded from frozen representations? PPG-derived pulse features provide within-waveform controls; intervals measured on the paired ECG test access to electrophysiological timing. **Adjusted and incremental prediction:** how much association remains after specified covariate adjustment, and what predictive gain does the representation add over measured covariates? We exclude target-derived adjustment variables. **Post-erasure predictability:** how much potassium prediction remains after removing selected concept-correlated directions and refitting, compared with random-direction, shuffled-concept, and heart-rate controls? Paired ECG comparisons hold the evaluated encounters fixed. For each claim, we report the reference labels, comparisons, estimates, and uncertainty.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

On 6,348 matched potassium encounters, adjustment for age and renal markers reduces AnyPPG correlation from 0.21 to 0.05, whereas ECGFounder retains 0.32 → 0.27 (paired adjusted difference +0.22; 95% CI [0.19, 0.26]). In a separate 5,160-encounter complete-case comparison, ECGFounder adds about 0.087 over the same demographic and renal baseline under ridge and nonlinear predictors; direct AnyPPG increments have confidence intervals including zero. AnyPPG also decodes pulse-related concepts well; erasure loss ranks the modalities differently from interval-concept decodability.

Contributions. (1) A three-axis audit of decodability, adjusted and incremental prediction, and post-erasure predictability, with a target-derivative rule and paired patient-cluster contrasts. (2) An evaluation of four frozen encoders on 37,784 MC-MED encounters and a 30-candidate inventory, with targeted controls and sensitivity checks. (3) A worked potassium audit comparing reference and decoded intervals with full representations over the same demographic and renal baseline, revealing different modality orderings for interval decodability and erasure loss.

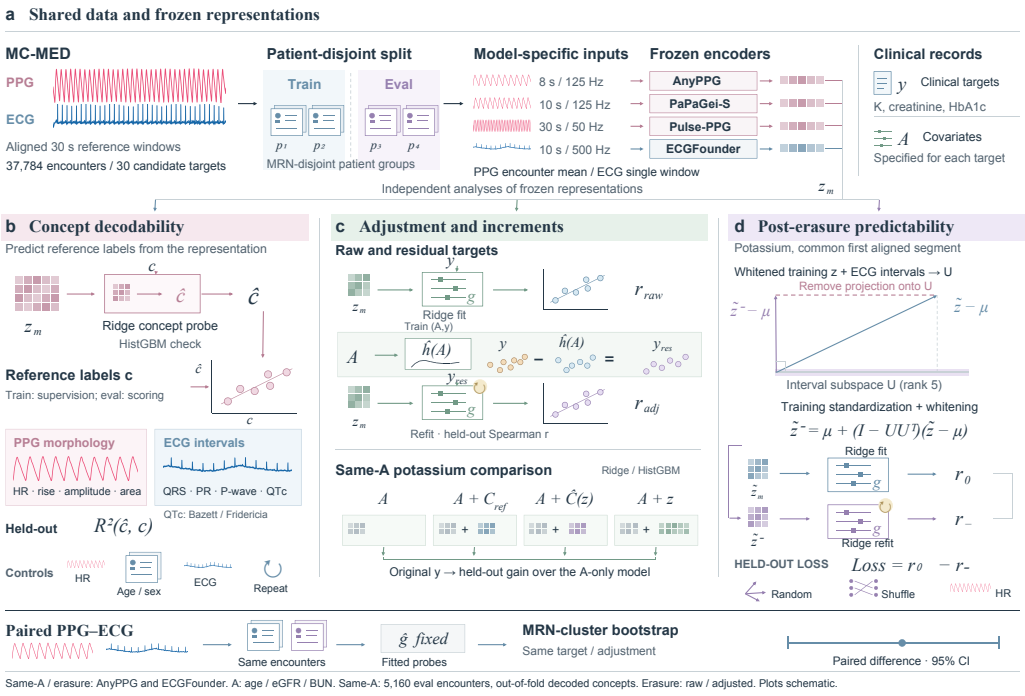


Figure 1: **Overview of the clinical prediction audit.** (a) Patient-disjoint MC-MED cohort and frozen PPG/ECG encoders. (b) Held-out R^2 for waveform-derived concept labels. (c) Residual-target association and a separate potassium comparison of reference intervals, decoded intervals, and full representations over the same baseline $A = \{\text{age, eGFR, BUN}\}$. (d) Potassium predictability after removing a rank-five linear ECG-interval subspace and refitting the target probe. The same-A and erasure comparisons use AnyPPG and ECGFounder; paired differences use matched encounters and patient-cluster bootstrap intervals with fitted probes fixed. Waveform traces are real; scatterplots, projection geometry, and confidence intervals are schematic.

2 Related Work

Biosignal foundation models. PPG encoders (Pillai et al., 2025; Saha et al., 2025; Ding et al., 2024; Nie et al., 2026) and ECG encoders (Li et al., 2025) are commonly benchmarked on downstream tasks; newer PPG and cross-modal models broaden this landscape (Chen et al., 2025; Guo et al., 2026; Zhou et al., 2026; Gu et al., 2026). We also examine what physiological labels their frozen representations make accessible.

Biosignal prediction audits. ECG potassium work adjusts for demographics and renal markers (Galloway et al., 2019); EEG/ECG audits test spectral nuisances (Bindra & Panwar, 2026). Our

PPG–ECG comparisons bring concept decoding, target-specific adjustment, incremental prediction, and potassium erasure-and-refitting together.

Concepts, shortcuts, and interventions. Concept methods quantify directions and explanatory completeness (Kim et al., 2018; Yeh et al., 2020); medical-image audits examine spurious cues (Winkler et al., 2019; Hill et al., 2024). Vision tests vary texture, background, or viewpoint (Geirhos et al., 2019; Xiao et al., 2020; Barbu et al., 2019). Our representation-level check instead removes a selected linear concept subspace and refits the target probe. For potassium, this joins concept accessibility, adjusted and incremental prediction, and post-erasure predictability in one evidence profile, with stated labels, fits, and evaluation populations.

3 METHOD

Audit object and units. For each encoder–target claim, we measure concept *decodability*, *adjusted and incremental prediction*, and *post-erasure predictability*. A frozen encoder f_m maps its model-specific input to $z_m = f_m(x) \in \mathbb{R}^{d_m}$; concept and target probes are trained on training patients. Results specify the target and label, encoder and input window, probe, and evaluation cohort. Candidate concepts define the physiological labels tested; adjustment set A and baseline set C define the prediction comparisons.

AnyPPG uses the central 8 s at 125 Hz, PaPaGei-S the central 10 s at 125 Hz, Pulse-PPG a length-agnostic 50 Hz version of the archived 30 s segment, and ECGFounder one central 10 s window at native 500 Hz, with a 125 Hz sensitivity. Concept labels use the full 30 s segment. Splits are patient-disjoint by medical record number (MRN). Target analyses include encounters with a measured target; n denotes encounters, with distinct-patient counts where relevant.

Reference labels. Primary laboratory tasks match MC-MED component names exactly and retain encounters with exactly one row for that component. For nine extension tasks, numeric repeats use the encounter median; ABO/Rh requires concordant records. Missing labels are excluded target-wise. Disease-state labels use past-medical-history ICD prefixes at the patient level and apply to that patient's encounters; the target is documented historical diagnosis. Concept preprocessing includes quality checks, band-pass filtering, resampling, and within-window standardization. Encoder inputs follow each released model's native adaptation.

3.1 CONCEPT DECODABILITY

For concept c, a probe of class p predicts its reference label from the frozen representation. PPG concept labels are derived from the same waveform used by PPG encoders; ECG intervals are delineated algorithmically with NeuroKit on the aligned ECG. Decodability is the held-out coefficient of determination,

$$D_{m,p}(c) = 1 - \frac{\sum_{i \in \mathcal{I}_c} (c_i - \hat{c}_{m,p,i})^2}{\sum_{i \in \mathcal{I}_c} (c_i - \bar{c}_{\text{eval}})^2}, \tag{1}$$

where $\mathcal{I}_c$ contains concept-valid evaluation rows, $\hat{c}_{m,p,i}$ is the trained probe's prediction, and $\bar{c}_{\text{eval}}$ is their mean reference label. Larger R^2 means less squared error; zero equals the evaluation-mean reference error, and negative values exceed it. We call the primary ridge score *linear concept decodability*; HistGBM tests nonlinear decoding.

A matched-segment control fits the same ridge specification to AnyPPG and ECGFounder on the first aligned 30 s segment: central 8 s for AnyPPG and central 10 s of archived 125 Hz ECG resampled to 500 Hz for ECGFounder. Heart-rate-only and age-plus-sex controls use the same concept-valid rows. Spearman correlation and absolute-agreement ICC(2,1) between the first two nonoverlapping valid 30 s windows measure internal label repeatability. Appendix N lists candidate target–concept links and measured coverage.

3.2 ADJUSTED AND INCREMENTAL PREDICTION

Target-specific adjustment. Adjustment set A excludes measured covariates deterministically derived from the target on the same specimen. Lab-reported eGFR uses that specimen's creatinine (CKD-EPI 2021); log-creatinine and eGFR have Pearson $r = -0.85$ ($n = 31{,}510$). Thus creatinine

uses $A_{\mathrm{Cr}} = \{\text{age, sex, BUN}\}$ (independently measured BUN), whereas potassium uses $A_{\mathrm{K}} = \{\text{age, eGFR, BUN}\}$, with an additional eGFR-removal sensitivity.

Residual predictability and attenuation. For target y, the raw probe gives $\hat{y} = g_\theta(z_m)$ and $r_{\mathrm{raw}} = \rho_S(\hat{y}, y)$. A training-fitted nuisance model $\hat{h}_A$ produces residuals $y^{\mathrm{res}} = y - \hat{h}_A(A)$. A separate target probe predicts these residuals, with held-out score $r_{\mathrm{adj}} = \rho_S(\hat{y}^{\mathrm{adj}}, y^{\mathrm{res}})$. Ridge is the primary nuisance model; double residualization also removes linear A-predictable variation from z_m. A sensitivity uses five-fold MRN-cross-fitted ridge and HistGradientBoosting nuisance models with the same fixed ridge representation probe: potassium, creatinine, and CRP on PPG, plus potassium on ECGFounder on identical encounters, with an MRN-cluster paired interval for ECG−PPG.

The Excess Correlation Fraction summarizes relative attenuation between separately fitted original- and residual-target probes:

$$\mathrm{ECF} = \frac{r_{\mathrm{raw}} - r_{\mathrm{adj}}}{|r_{\mathrm{raw}}|}. \tag{2}$$

It can be negative; we report it for prespecified $|r_{\mathrm{raw}}| \geq 0.15$.

Increment over measured baseline variables. Residual-target association removes the fitted contribution of A from the target. Incremental prediction instead retains the original target and tests whether embeddings improve prediction alongside baseline C, which comprises demographics and triage vitals. Let $M_y(\hat{y}, y)$ be Spearman correlation for laboratories and AUROC for binary targets. On held-out encounters,

$$\Delta_m(y; C) = M_y\big(g(C, z_m), y\big) - M_y\big(g(C), y\big). \tag{3}$$

Linear features use training-set standardization. Primary regression uses ridge ($\alpha = 1$); classification uses logistic regression ($C_{\mathrm{LR}} = 1$). Because feature concatenation and regularization affect fitting, Δ_m compares these fitted procedures. For potassium, creatinine, and CRP, a sensitivity reserves 20% of training MRNs for validation, tunes ridge $\alpha \in \{.01, .1, 1, 10, 100, 1000\}$ separately for $g(C)$ and $g(C, Z)$, and selects the baseline or combined model by validation performance before evaluation.

Direct prediction over the same adjustment set. Unlike the broad increment over C, this potassium comparison uses the same adjustment set A as the baseline for interval features and full representations. We compare A, $A + C_{\mathrm{ref}}$, $A + \widehat{C}_m$, and $A + Z_m$ on one complete-case cohort. Here $A = \{\mathrm{age}, \mathrm{eGFR}, \mathrm{BUN}\}$, and C_{ref} contains QRS duration, PR interval, P-wave duration, and Bazett- and Fridericia-corrected QT. Ridge concept probes ($\alpha = 1$) predict $\widehat{C}_m$ from Z_m, using five patient-level folds for out-of-fold training predictions and all training patients for evaluation predictions. Ridge and HistGradientBoosting target predictors are tuned separately by configuration within training patients, then refitted on all training patients. The cohort has 20,734 training and 5,160 evaluation encounters (3,996 evaluation patients), each with potassium, A, and all five interval labels. We report held-out Spearman correlation and its difference from the A-only model; paired 95% intervals use 1,000 shared MRN-cluster bootstrap resamples with fitted models fixed (Appendix Q).

3.3 Post-erasure predictability

For potassium, we remove selected concept-correlated directions and refit the target probe. Both modalities use the first aligned segment: AnyPPG takes its central 8 s and ECGFounder a native-500-Hz central 10 s. Training-only standardization and PCA whitening yield $\widetilde{Z}_m$. Let C_{phys} contain jointly valid QRS duration, PR interval, P-wave duration, and Bazett- and Fridericia-corrected QT. The columns of U_m form the rank-five left singular basis of $\mathrm{Cov}(\widetilde{Z}_m, C_{\mathrm{phys}})$. For training mean μ_m in whitened coordinates, we project

$$\widetilde{z}_m^{-} = \mu_m + (I - U_m U_m^\top)(\widetilde{z}_m - \mu_m). \tag{4}$$

We refit a potassium probe with the same fixed ridge specification. The baseline held-out correlation minus the post-erasure correlation, on the same target and evaluation encounters, is *erasure sensitivity*; post-erasure correlation measures retained predictability after this linear projection and refit.

We also residualize each concept linearly on age, eGFR, and BUN using training data, and analyze raw potassium and five-fold MRN-cross-fitted ridge- and HistGBM-adjusted residuals. Rank-matched random directions, row-shuffled concepts, and a heart-rate subspace are controls. Before/after concept-probe scores track decodability; MRN-cluster bootstrap intervals hold fitted probes fixed (Appendix M).

Paired comparisons and inference. On identical encounters, we compare PPG and ECG raw and adjusted correlations, increments, and potassium erasure losses. Paired 95% CIs resample MRN clusters with fitted probes fixed, quantifying evaluation-sampling uncertainty conditional on the fitted pipeline. A paired CI excluding zero supports a directional comparison under that procedure. The broad incremental table uses encounter-bootstrap intervals and BH correction across target–modality rows; conditional-association q-values are corrected within each reported target family. Archived conditional-table p-values use a two-sided Fisher-z approximation for Spearman correlation, with $\mathrm{SE} = 1.03/\sqrt{n-k-3}$ and $k = 3$ adjusted variables. That calculation treats encounters as independent and is retained for implementation provenance; labeled paired results use MRN clusters.

Per-claim interpretation. Each claim's profile reports available measurements alongside its reference labels, candidate concepts, covariates, cohorts, probes, controls, and uncertainty. $D_{m,p}(c)$ measures selected-concept accessibility; r_{adj} measures residual-target association and ECF summarizes relative attenuation; $\Delta_m(y; C)$ measures fitted gain over C; and post-erasure correlation and loss measure retained predictability and erasure sensitivity. Paired contrasts compare matched evaluation encounters. We report disagreements among these measurements and target–concept coverage (Appendix N).

Secondary expert-prior analysis. We compare increments with archived clinician-consensus mechanism tiers for 26 analyzable targets. Consensus was frozen before tier computation, but individual ratings are unavailable and authors had seen earlier analyses. This comparison is exploratory (Appendices A and C).

4 Experiments

4.1 Experimental setup

Data. MC-MED (Kansal et al., 2025) provides bedside PPG (Pleth, 125 Hz) and ECG (lead II, 500 Hz) with laboratories and demographics for emergency-department patients. Our build has 37,784 encounters (5,521 evaluation patients); the potassium evaluation has 6,348 encounters from 4,783 patients. An archived VitalDB (Lee et al., 2022) build has 6,119 of 6,388 public cases passing availability and quality filters, but contributes no evidentiary results (§4.7). WildPPG (Meier et al., 2024) and PPG-DaLiA (Reiss et al., 2019) supply wearable PPG for concept probes only, without laboratory labels. Dataset versions and licenses appear in the Ethics Statement and Appendix P.

Signal pairing and timestamp audit. Within each encounter identifier (CSN), PPG and ECG are aligned by WFDB `base_date+base_time`; we trim the earlier channel before extracting common relative windows. Of 53,843 successfully read, expected-rate PPG–ECG pairs, header-only reconstruction shows that all 16,059 without an analyzed segment lack a complete aligned 30 s window. The remaining 37,784 CSNs yield 106,983 analyzed segments, each with a parsed aligned start. PPG- and ECG-derived heart rates differ by a median 0.66 bpm (IQR 0–1.75; 83.5% within 3 bpm), a functional check of record-relative pairing, not specimen timing (Appendix G). Target probes average PPG segment embeddings within CSN; ECGFounder uses one native central 10 s window per CSN. Paired target contrasts match encounters, retaining these modality-specific extraction protocols.

Encoders and exposure. AnyPPG (Nie et al., 2026), PaPaGei-S (Pillai et al., 2025), Pulse-PPG (Saha et al., 2025) (PPG, 512-d), and ECGFounder (Li et al., 2025) (ECG, 1024-d) are frozen. AnyPPG has MC-MED pretraining exposure and PaPaGei reports VitalDB exposure; patient-level overlap is unresolved in released identifiers (Appendix O). Probes use training-standardized ridge (α=1) or logistic regression (C=1) with fixed hyperparameters. HistGBM tests nonlinear fitting; shuffled-label ($\rho = -0.01$) and random-initialization controls check the pipeline. Two identically architected supervised CNNs provide a separate matched-architecture control (§4.7).

Reading the potassium comparisons. The 6,348 matched potassium encounters support raw and covariate-adjusted correlations. A separate 5,160-encounter complete-case comparison, requiring all five ECG intervals, tests reference intervals, decoded intervals, and full embeddings against the same age/eGFR/BUN baseline for prediction of the original potassium label. Erasure returns to the matched potassium encounters, using a common first segment to measure what remains after linear subspace removal and target-probe refitting. The reported changes therefore have different baselines and fitting procedures.

4.2 Conditional association: potassium and creatinine

Figure 2(a) summarizes the primary laboratory results with patient-cluster CIs. Here adjusted correlation measures prediction of potassium after removing the component predicted by age, eGFR, and BUN on training patients. AnyPPG correlation falls from r_{raw}=0.21 to r_{adj}=0.05 after adjustment for {age, eGFR, BUN} (ECF 0.77; 0.75 under double residualization); ECGFounder retains 0.32 → 0.27 (ECF 0.16). On the same 6,348 encounters, the ECG−PPG adjusted difference is +0.22 (cluster 95% CI $[0.19, 0.26]$); the raw difference is +0.11 ($[0.08, 0.14]$). The broad baseline-increment comparison also favors ECG by +0.10 ($[0.08, 0.12]$); the same-A comparison below is a distinct complete-case analysis. The adjusted contrast persists with ECG band-limited to 125 Hz (r_{adj}=0.25). Across three PPG backbones, potassium increments over demographics are at most 0.01 in magnitude or have CIs overlapping zero.

Creatinine illustrates the target-derivative rule. Including lab-reported eGFR gives an adjusted correlation of 0.014, but eGFR is computed from the same specimen's creatinine. With the derivative-free {age, sex, BUN} set, adjusted correlation is 0.063 (cluster 95% CI $[0.039, 0.087]$; ECF 0.75): substantial attenuation with a small retained association. The difference between these analyses shows the effect of adjusting for a function of the target.

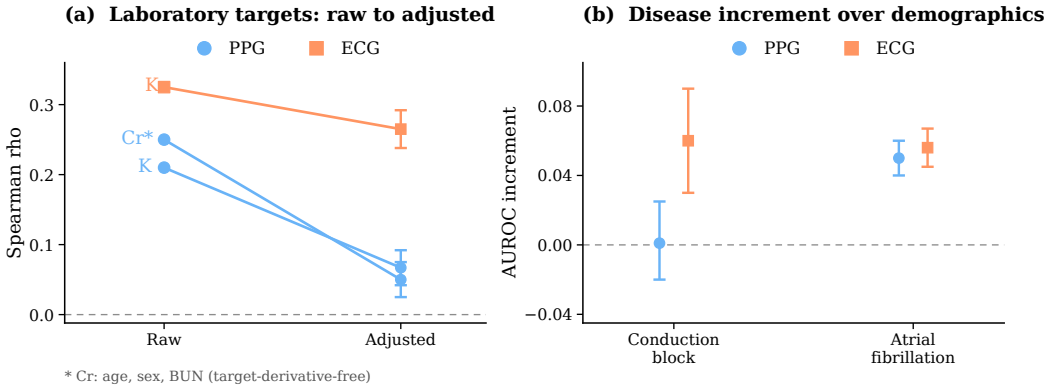


Figure 2: **Conditional association and paired modality contrasts.** (a) Laboratory targets: raw and covariate-adjusted Spearman ρ with patient-cluster 95% CIs on the adjusted values. Potassium uses {age, eGFR, BUN}; creatinine (†) uses the target-derivative-free set {age, sex, BUN}. Potassium's paired adjusted difference is ECG−PPG = 0.221 (cluster 95% CI $[0.186, 0.257]$). (b) Historical disease-state targets: incremental skill over demographics with bootstrap CIs. The panels use laboratory correlation and disease-classification increments, respectively.

4.3 Incremental potassium prediction

Unlike the residual-target analysis above, this comparison predicts the original potassium label from age, eGFR, and BUN plus, separately, reference intervals, decoded intervals, or full representations. All four feature sets use the same A-only baseline. Requiring all five interval labels yields a separate, interval-complete evaluation cohort of 5,160 encounters from 3,996 patients (Table 1).

Table 1: **Potassium prediction over the same adjustment set.** Held-out Spearman correlations on the common complete-case cohort. A contains age, eGFR, and BUN; C_{ref} contains five ECG interval labels; $\widehat{C}_m$ contains their predictions from each representation. Ridge and HistGBM denote potassium predictors; both use ridge concept decoders. The first two rows are shared baselines.

	AnyPPG		ECGFounder	
Inputs	Ridge	HistGBM	Ridge	HistGBM
A	.3521	.3570	.3521	.3570
$A + C_{\mathrm{ref}}$	.3568	.3628	.3568	.3628
$A + \widehat{C}_m$	.3598	.3643	.3571	.3604
$A + Z_m$	.3437	.3647	.4391	.4442

The A-only correlations are .3521 with ridge and .3570 with HistGBM; adding ECGFounder raises them to .4391 and .4442. Its paired increments over A are .0870 with ridge (95% CI [.0647, .1103]) and .0872 with HistGBM ([.0684, .1054]). Adding AnyPPG directly changed correlation by $-.0084$ ([$-.0193$, .0029]) and .0077 ([$-.0003$, .0153]), respectively. Reference intervals added .0047 ([.0005, .0085]) with ridge and .0058 ([$-.0010$, .0120]) with HistGBM. Ridge gains from decoded intervals were .0077 ([.0019, .0135]) for AnyPPG and .0050 ([.0005, .0089]) for ECGFounder. The full ECG representation adds more than the tested interval sets; for AnyPPG, the decoded-interval gain shows that the estimated increment depends on how its representation is used.

4.4 Linear physiological concept decodability

The interval comparisons above measure gains in potassium prediction; decodability instead tests how accurately the representations encode the reference labels themselves. On MC-MED, ridge probes decode PPG-derived pulse and morphology labels from AnyPPG (rise time R^2=0.90, heart rate 0.83, systolic amplitude and pulse area 0.5–0.8). They decode ECG-derived intervals less accurately: QRS duration 0.04, Bazett QTc 0.07, P-wave duration 0.11, PR interval 0.14, and Fridericia QTc 0.22 (which includes heart-rate correction). Group means are 0.74 for pulse features, 0.55 for PPG morphology, and 0.12 for ECG intervals. A HistGBM probe barely changes the ECG-interval group mean ($0.115 \rightarrow 0.116$); PaPaGei-S also has a low interval mean (0.06).

On the same first segment, the five interval labels have mean R^2 0.0951 from AnyPPG versus 0.4883 from ECGFounder; heart-rate-only and age/sex baselines reach 0.0529 and 0.0286. First-to-second-window label repeatability is 0.742–0.873 by Spearman correlation (ICC 0.739–0.875). These comparisons place PPG decodability alongside ECG readouts, simple baselines, and internal label consistency (Appendix L).

4.5 Decodability and erasure loss give different modality orderings

Returning to the 6,348 potassium encounters used for conditional association, the first-segment erasure design starts from raw correlations of 0.1909 for AnyPPG and 0.3203 for ECGFounder. After ridge adjustment, they are 0.0334 and 0.2696 (HistGBM: 0.0339 and 0.2670). On the jointly valid erasure cohort, the five intervals have mean linear-decoding $R^2 = 0.0950$ from AnyPPG versus 0.3490 from native-500-Hz ECGFounder.

Erasure gives the opposite ordering in absolute loss. After residualizing the concepts on age, eGFR, and BUN, rank-five erasure and refitting reduce ridge-adjusted potassium correlation from 0.0334 to 0.0178 for AnyPPG (loss 0.0156, cluster 95% CI $[0.0078, 0.0232]$) and from 0.2696 to 0.2615 for ECGFounder (loss 0.0081, $[0.0043, 0.0123]$). The paired ECG$-$PPG loss difference is -0.0076 (cluster 95% CI $[-0.0139, -0.0016]$); HistGBM-adjusted losses are 0.0186 and 0.0084. Mean residualized-concept R^2 falls from 0.0612 to -0.0003 for AnyPPG and from 0.3142 to -0.0003 for ECGFounder. Both adjusted losses exceed 100 rank-matched random-direction losses (empirical one-sided $p = .0099$) and row-shuffled-concept losses. AnyPPG's loss is near its heart-rate-subspace control (0.0128); ECGFounder's exceeds its heart-rate control (0.0021). The starting scores and controls place the larger absolute PPG loss alongside its much weaker retained association and comparable heart-rate sensitivity (Appendix M).

4.6 HISTORICAL DISEASE-STATE TARGETS AND CONTROLS

Conduction block (I44/I45), a historical diagnosis associated with QRS/PR morphology, shows a small PPG increment over demographics ($+0.003$, $[-0.023, 0.026]$) versus a larger ECG increment ($+0.06$, $[0.03, 0.10]$). These encounter-bootstrap intervals support descriptive comparisons. Historical atrial fibrillation has positive increments on both modalities (PPG $+0.05$ $[0.04, 0.06]$). ABO and Rh shams have embedding-only AUROCs of 0.513 ($[0.480, 0.546]$) and 0.546 ($[0.483, 0.603]$). The blood-type results calibrate the pipeline; the AF result indicates additional association with documented history.

4.7 ROBUSTNESS OF THE CONDITIONAL-ASSOCIATION PATTERN

Matched-architecture control. We train one 1D-CNN per modality, with identical depth, width, 512-d embedding, seed, budget, 125 Hz input, and *potassium supervision*, then freeze both. The PPG encoder changes $0.17 \rightarrow 0.039$ (ECF 0.77) and ECG $0.31 \rightarrow 0.26$ (ECF 0.14), retaining the cross-modal contrast under a shared architecture and training regime.

Finite nuisance-form sensitivity. With five-fold MRN-cross-fitted nuisance predictions, PPG potassium adjusted correlation is 0.0488 under ridge versus 0.0494 under HistGBM (attenuation 0.768 vs. 0.765). On identical encounters ECGFounder retains 0.2696 and 0.2670: paired ECG$-$PPG differences are $+0.2208$ (cluster 95% CI $[0.1872, 0.2558]$) and $+0.2176$ ($[0.1843, 0.2543]$). Creatinine PPG also attenuates (0.0631 vs. 0.0497; attenuation 0.748 vs. 0.802). CRP is more nuisance-form sensitive: adjusted correlation falls from 0.1191 to 0.0478 and attenuation rises from 0.232 to 0.691. The potassium contrast persists under both nuisance families, whereas the CRP estimate depends more on that choice (Appendix E).

When both modalities instead use their central crop from the same first aligned 30 s segment without within-CSN averaging, the HistGBM-adjusted ECG$-$PPG difference remains $+0.2105$ (cluster 95% CI $[0.1761, 0.2462]$). This matches segment selection but retains native input durations (PPG 8 s; ECG 10 s after 125-to-500 Hz resampling).

Timing and external scope. Evidentiary laboratory analyses join by encounter identifier and align waveforms relative to their recorded starts. The earlier timing analysis is withdrawn: it substituted order time for specimen collection time and used an unsuitable date parser; specimen collection times are unavailable. The archived VitalDB analysis is excluded because of flawed renal adjustment, incomplete paired inference, and encoder pretraining exposure. Timing-resolved and external laboratory validation remain open (Appendices G and O).

4.8 INCREMENTAL VALUE OVER TRIVIAL BASELINES

Across the 17 nontrivial targets in the original task set (excluding age and sex, and with two all-negative disease labels not estimable), the AnyPPG embedding adds a median $+0.033$ over demographics-plus-triage-vitals, and covariates alone outperform the embedding alone on $12/17$ targets. The appendix ladder covers six prespecified exemplars. Under fixed probes, combined models score below the covariate baseline for creatinine (0.431 covariates, 0.250 embedding, 0.369 combined) and CRP (0.437, 0.155, 0.272). In the validation sensitivity, tuned combined scores are 0.384 versus a 0.432 baseline for creatinine and 0.420 versus 0.438 for CRP; validation selects the baseline for both, yielding guarded increments of zero. For potassium, tuning changes the fixed -0.0116 increment to $+0.0003$. These checks show how tuning and baseline fallback affect the three tested targets, so we report both fixed-probe and validation-selected comparisons.

5 DISCUSSION AND LIMITATIONS

What the potassium comparisons show. The five ECG intervals are more accurately decoded from ECGFounder, yet adding their reference or decoded values to age, eGFR, and BUN yields smaller potassium gains than adding the full ECG representation. The 0.087 full-representation increment is measured on the $5{,}160$-encounter interval-complete cohort, distinct from the $6{,}348$-encounter adjusted-correlation comparison; both comparisons favor ECG under their respective metrics. For AnyPPG, decoded intervals yield a small ridge gain while direct feature concatenation has an incre-

ment whose CI includes zero. Thus the tested intervals account for only a small part of the measured ECG increment, while feature use matters for AnyPPG. Erasure asks a different question: after the selected linear projection and probe refit, AnyPPG has the larger absolute loss, but starts from far lower adjusted correlation and has a loss near its heart-rate control. Decodability, incremental gain, and erasure sensitivity provide complementary comparisons of the same potassium claim.

Connecting evidence to intended use. These profiles measure held-out, encounter-level prediction and selected physiological readouts. The renal-adjusted potassium comparison uses eGFR and BUN from the same encounter to characterize residual-target association; a prediction-time application without those laboratories requires a baseline of variables available at that time. Agreement with a contemporaneous potassium measurement calls for specimen-timed references, while tracking within-person change calls for repeated longitudinal references. External cohorts can test transportability.

Secondary expert priors. The pooled tier–increment association ($\rho = 0.454$, $p = .021$) weakens under metric-stratified normalization ($p = .440$), removal of blood-type controls from binary targets ($p = .330$), and restriction to historical disease states ($p = .502$). The archived clinician tiers thus guide hypothesis generation (Appendix C).

Diagnostic coverage. Concept scores depend on the selected labels and probe families: PPG-derived pulse labels share the waveform with their encoder, while ECG intervals are algorithmically delineated. Window-to-window repeatability measures internal label consistency. The matched-segment ECG concept control resamples archived 125 Hz ECG to 500 Hz; erasure instead uses native 500 Hz ECG and a rank-five subspace estimated from jointly valid intervals. Random, shuffled, and heart-rate controls contextualize the resulting losses. Patient-cluster intervals resample evaluation patients with fitted probes fixed; the broad incremental table instead uses encounter resampling. Adjustment checks cover ridge and HistGBM for PPG potassium, creatinine, and CRP and for ECG potassium; CRP is more sensitive to nuisance-model choice. The same-A comparison requires all five interval references and therefore uses a smaller complete-case cohort. Paired comparisons match evaluation encounters; input windows and pretraining exposure remain model-specific.

Clinical and audit-validation coverage. The evidence comes from an emergency-department cohort with encounter-linked laboratories and historical diagnosis labels. AnyPPG has MC-MED pretraining exposure; ECGFounder's exposure is unresolved. Waveform alignment is record-relative, and specimen collection times are unavailable. The excluded timing and VitalDB analyses are documented in §4.7. Further validation can use overlap-safe external cohorts, specimen-timed and longitudinal references, and adjudicated labels; an audit-decision benchmark would additionally need known claim validity, a comparator, and a prespecified decision metric.

6 Conclusion

Clinical prediction claims become clearer when concept accessibility, conditional and incremental prediction, and post-erasure predictability are evaluated together. AnyPPG encodes PPG-derived pulse features, while target-specific adjustment substantially attenuates PPG potassium and creatinine associations. On matched potassium encounters ECGFounder retains a stronger adjusted association; on a separate interval-complete cohort its full representation adds more over the same demographic and renal baseline than the tested ECG intervals. Linear interval decodability and absolute erasure loss rank the modalities differently, with AnyPPG's larger loss starting from much lower adjusted correlation and near its heart-rate control. Together, the available measurements form evidence profiles for evaluated claims, with their labels, cohorts, and controls stated explicitly.

AI Use Statement

AI assistants provided feedback on experimental design and result interpretation, assisted with drafting and language editing based on the authors' ideas and content, and supported code implementation and debugging. The authors independently reviewed and verified the incorporated work and take responsibility for the final paper and code.

Reproducibility Statement

Splits are patient-disjoint. Primary probes use fixed hyperparameters; specified sensitivity analyses tune on training data. The adjustment sets, target-derivative rule, ECF reporting threshold ($|r_{\text{raw}}| \geq 0.15$), and frozen consensus tiers are specified in §3 and the appendices. We distinguish encounter bootstrap, archived unclustered Fisher-z approximations, and MRN-cluster bootstrap where reported.

The target inventory, consensus-tier summary, per-target results, and scripts accompany the anonymous artifact. Raw MC-MED and VitalDB waveform access follows the source licenses; derived tables can be provided. `run_revision_audits.sh` reruns the revision-added analyses from frozen structured MC-MED artifacts when the dataset is mounted. Raw ingestion, encoder inference, the original target inventory, and figure generation are separate pipeline stages. The entry point runs the revision-2/3/4/5/6 analyses and `revision4_artifact_audit.py`, which checks archived JSON-to-manuscript values and records file hashes in `revision4_artifact_audit.json`. These hashes identify the archived artifact snapshot; the withdrawn timing analysis remains a provenance record. The same-adjustment-set comparison is implemented in `revision6_concept_target_ladder.py`, with results in `revision6_concept_target_ladder_results.json`; its reported environment is Python 3.12, scikit-learn 1.8.0, and NumPy 2.x.

Ethics Statement

The existing patient datasets are de-identified and used under their sources' terms: MC-MED v1.0.1 (PhysioNet credentialed access), VitalDB (open research license), WildPPG (CC BY-NC-SA 4.0), and PPG-DaLiA (UCI Machine Learning Repository, CC BY 4.0); per-dataset citations and versions are in Appendix P. The expert grading task presented only target names and definitions, without patient data, to consenting clinician collaborators. The retained expert artifact is the signed consensus summary; individual ratings are unavailable. Subgroup analyses report target-specific encounter counts and disease-positive counts where available (Appendix K); these are descriptive within-stratum performance and calibration summaries.

References

Tomy Abuzairi, Ester Vinia, Muhammad Arkana Yudhistira, Mia Rizkinia, and Winda Eriska. A dataset of hemoglobin blood value and photoplethysmography signal for machine learning-based non-invasive hemoglobin measurement. *Data in Brief*, 52:109823, 2024. doi: 10.1016/j.dib.2023.109823.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Joshua Tenenbaum, and Boris Katz. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jasmeet Singh Bindra and Siddharth Panwar. A spectral audit framework reveals task-dependent aperiodic reliance across EEG and ECG deep learning. *arXiv preprint arXiv:2606.08583*, 2026. doi: 10.48550/arXiv.2606.08583.

Zhaoliang Chen, Cheng Ding, Saurabh Kataria, Runze Yan, Minxiao Wang, Randall J. Lee, and Xiao Hu. GPT-PPG: A GPT-based foundation model for photoplethysmography signals, 2025.

I-Min Chiu, Po-Jung Wu, Huan Zhang, J. Weston Hughes, Albert J. Rogers, Laleh Jalilian, Marco Perez, Chun-Hung Richard Lin, Lee Chien-Te, James Zou, and David Ouyang. Serum potassium monitoring using AI-enabled smartwatch electrocardiograms. *JACC: Clinical Electrophysiology*, 10(12):2644–2654, 2024. doi: 10.1016/j.jacep.2024.07.023.

Cheng Ding, Zhicheng Guo, Zhaoliang Chen, Randall J. Lee, Cynthia Rudin, and Xiao Hu. SiamQuality: A ConvNet-based foundation model for photoplethysmography signals. *Physiological Measurement*, 45(8):085004, 2024. doi: 10.1088/1361-6579/ad6747.

Conner D. Galloway, Alexander V. Valys, Jacqueline B. Shreibati, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiology*, 4(5):428–436, 2019. doi: 10.1001/jamacardio.2019.0640.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Robert Geirhos, Joern-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Xiao Gu, Wei Tang, Jinpei Han, Veer Sangha, Fenglin Liu, Shreyank N. Gowda, Antonio H. Ribeiro, Patrick Schwab, Kim Branson, Lei Clifton, Antonio Luiz P. Ribeiro, Zhangdaihong Liu, and David A. Clifton. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, 8:220–233, 2026. doi: 10.1038/s42256-026-01180-5.

Zongheng Guo, Tao Chen, Yang Jiao, Yi Pan, Xiao Hu, and Manuela Ferrario. SIGMA-PPG: Statistical-prior informed generative masking architecture for PPG foundation model, 2026.

B. G. Hill, F. L. Koback, and P. L. Schilling. The risk of shortcutting in deep learning algorithms for medical imaging research. *Scientific reports*, 2024. doi: 10.1038/s41598-024-79838-6.

Aman Kansal, Emma Chen, Tom Jin, Pranav Rajpurkar, and David Kim. MC-MED: Multimodal clinical monitoring in the emergency department. *PhysioNet*, 2025. doi: 10.13026/wvyw-g663. Version 1.0.1.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677, 2018.

Hyung-Chul Lee et al. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022. doi: 10.1038/s41597-022-01411-5.

Jun Li, Aaron D. Aguirre, Valdery Moura, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7), 2025. doi: 10.1056/AIoa2401033.

Sara Lombardi, Leonardo Bocchi, and Piergiorgio Francia. Photoplethysmography and artificial intelligence for blood glucose level estimation in diabetic patients: A scoping review. *IEEE Access*, 12:178982–178996, 2024. doi: 10.1109/ACCESS.2024.3508467.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Abnormality prediction and forecasting of laboratory values from electrocardiogram signals using multimodal deep learning. *Scientific Reports*, 15:39362, 2025. doi: 10.1038/s41598-025-26715-5.

Manuel Meier, Berken Utku Demirel, and Christian Holz. WildPPG: A real-world PPG dataset of long continuous recordings. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Guangkun Nie, Xiaocheng Fang, Gongzheng Tang, Yujie Xiao, Jun Li, Bo Liu, Hongyan Li, and Shenda Hong. AnyPPG: An ECG-guided PPG foundation model trained on over 100,000 hours of recordings for holistic health profiling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026. doi: 10.1145/3770855.3818911.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2410.20542.

Attila Reiss, Ina Indlekofer, and Philip Schmidt. PPG-DaLiA. UCI Machine Learning Repository, 2019.

Mithun Saha, Maxwell A. Xu, Wanting Mao, Sameer Neupane, James M. Rehg, and Santosh Kumar. Pulse-PPG: An open-source field-trained PPG foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):126:1–126:35, 2025. doi: 10.1145/3749494.

Julia K. Winkler et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10):1135–1141, 2019. doi: 10.1001/jamadermatol.2019.1735.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. doi: 10.48550/arXiv.2006.09994.

Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20554–20565, 2020.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018. doi: 10.1371/journal.pmed.1002683.

Hao Zhou, Simon A. Lee, Cyrus Tanade, Keum San Chun, Juhyeon Lee, Migyeong Gwak, Megha Thukral, Justin Sung, Eugene Hwang, Mehrab Bin Morshed, Li Zhu, Viswam Nathan, Md Mahbubur Rahman, Subramaniam Venkatraman, and Sharanya Arcot Desai. Physiology-aware masked cross-modal reconstruction for biosignal representation learning, 2026.

A Expert grading rubric and consensus

The archived rubric summary (v1.1, dated 2026-09-02, before the panel saw reported result tables) defines five tiers, a half-step rule for co-dominant pathways, and a claim-type classification (measurement / state / baseline-absorbed). It records ten clinicians independently grading 30 targets on both modalities and a consensus frozen as `tier_signed_v1` on 2026-09-16 (mode plus named adjudications, including conduction block PPG L_4 under the rubric's low-pass ruling). The retained expert artifact is the signed consensus summary, not the individual ratings, complete immutable rubric, or adjudication log needed to reconstruct rating distributions, agreement coefficients, and adjudications. Authors had inspected earlier model results before this exercise; the consensus tiers are therefore secondary, hypothesis-generating annotations.

B Target inventory and tier-analysis scope

The 30-candidate inventory covers four analyses with different inclusion rules (Table 2): the 26-target tier test (T), the 17-target nontrivial baseline summary (B), threshold-reportable laboratory conditional-association results (C), and the negative-control analysis (E). Age and sex serve as baseline variables; cardiac-device and sleep-breathing labels have zero positives in the extracted task table. Excluding these four from the tier analysis leaves 26 targets; applying the same exclusions within the original task family leaves 17 baseline-summary targets.

Table 2: **Complete candidate inventory and analysis inclusion.** T/B/C/E denote inclusion in the tier, nontrivial baseline, threshold-reportable conditional, and negative-control analyses. "Threshold" denotes a result below the prespecified $|r_{\mathrm{raw}}| \geq 0.15$ reporting rule.

Candidate	Label source / exclusion	T	B	C	E
Age	Visit field; baseline variable	–	–	–	–
Sex	Visit field; baseline variable	–	–	–	–
ED disposition	Visit disposition; non-disease binary outcome	yes	yes	–	–
Potassium	Single exact-match laboratory row	yes	yes	yes	–
Calcium	Single exact-match laboratory row	yes	yes	yes	–
eGFR	Single exact-match laboratory row; target-specific C not run	yes	yes	–	–
Glucose	Single exact-match laboratory row	yes	yes	yes	–

Continued on next page

Table 2: **Complete candidate inventory and analysis inclusion (continued).**

Candidate	Label source / exclusion	T	B	C	E
Hemoglobin	Single exact-match laboratory row	yes	yes	yes	–
Albumin	Single exact-match laboratory row; not in pre-specified C family	yes	yes	–	–
BUN	Single exact-match laboratory row; target-specific C not run	yes	yes	–	–
Sodium	Single exact-match laboratory row	yes	yes	yes	–
Atrial fibrillation	Patient-level PMH ICD prefix	yes	yes	–	–
Heart failure	Patient-level PMH ICD prefix	yes	yes	–	–
Cardiac device	Patient-level PMH ICD prefix; zero positives	–	–	–	–
CKD	Patient-level PMH ICD prefix	yes	yes	–	–
Diabetes	Patient-level PMH ICD prefix	yes	yes	–	–
Sleep breathing disorder	Patient-level PMH ICD prefix; zero positives	–	–	–	–
Anemia	Patient-level PMH ICD prefix	yes	yes	–	–
Respiratory compromise	Patient-level PMH ICD prefix	yes	yes	–	–
VTE	Patient-level PMH ICD prefix	yes	yes	–	–
Conduction block	Patient-level PMH ICD prefix	yes	yes	–	yes
Creatinine	Encounter median of numeric repeats	yes	–	yes	–
Magnesium	Encounter median; below C threshold	yes	–	–	–
NT-proBNP	Encounter median of numeric repeats	yes	–	yes	–
HbA1c	Encounter median; below C threshold	yes	–	–	–
CRP	Encounter median of numeric repeats	yes	–	yes	–
TSH	Encounter median; below C threshold	yes	–	–	–
Lactate	Encounter median of numeric repeats	yes	–	yes	–
ABO type	Concordant categorical records; sham control	yes	–	–	yes
Rh factor	Concordant categorical records; sham control	yes	–	–	yes

Table 3 lists matched increment outputs for the 26 tier-test targets after these exclusions.

C SECONDARY EXPERT-PRIOR ANALYSIS

Expert mechanism tiers. The clinicians represented emergency medicine, cardiology, nephrology, endocrinology, pulmonology, hematology, critical care, anesthesiology, clinical physiology, and geriatrics. Their rubric assigned L_0 to direct substrate, L_1 one-step mechanical/event coupling, L_2 fast shared regulation, L_3 chronic covariation only, L_4 no physical pathway; half-steps allowed for co-dominant pathways. Tiers are *priors*, coded as $\text{closeness} = 4 - L$ (higher = mechanistically closer). The 26-target tier–increment test follows the exclusions in Appendix B; the consensus record and its exploratory status are described in Appendix A.

Figure 3(a–b) relates the consensus tiers to measured increments for the 26 analyzable targets. The pooled Spearman association is ρ=0.454 (permutation p=0.021). All 11 AUROC targets—the subset previously labeled "disease-only"—include ED disposition and the ABO and Rh controls (ρ=0.628, p=0.042). Excluding the blood-type controls leaves nine binary targets (ρ=0.368, exact permutation p=0.330); restricting to the eight historical disease-state targets gives ρ=0.279 (p=0.502). The 15 continuous laboratory targets yield ρ=0.290 (p=0.327), and rank-normalizing within metric type with claim-stratified permutation yields ρ=0.178 (p=0.440). These sensitivities limit the pooled result to exploratory association with the recorded priors.

The tier–attenuation disagreements also merit attention. Calcium, sodium, and glucose were graded L_3 ("chronic covariation only") yet retain adjusted associations with low ECF (Figure 3(c)); NT-proBNP has high ECF and a small residual under the archived unclustered Fisher-z/BH approximation. Vascular or systemic relationships compressed by the L_3 category are possible explanations to test for these MC-MED associations. Appendix B supplies target codings, metrics, and inclusion flags.

D CONDITIONAL-ASSOCIATION ANALYSES

Table 4 reports the prespecified five-target family meeting the $|r_{\text{raw}}| \geq 0.15$ embedding-correlation reporting threshold; albumin is outside this family. BUN and eGFR need separate adjustment sets that exclude the target itself and, for eGFR, target-derived renal quantities. Those reruns are unavailable, so their ECF values are not reported. The frozen scripts compute two-sided Spearman p-values using a Fisher-z normal approximation with $\text{SE} = 1.03/\sqrt{n-k-3}$ and $k = 3$. BH correction is within the stated family. The primary script uses in-memory p-values; the extension script rounds

Table 3: **Rated targets with matched increment outputs.** Closeness $= 4 - L$; $\Delta = \text{skill}(C, Z) - \text{skill}(C)$. The released columns Perf_C and Perf_Z are the covariate-only and embedding-only scores; the combined score is $\text{Perf}_C + \Delta$. Intervals in this broad table describe encounter-bootstrap uncertainty; the resampling unit is the encounter.

Target	Claim	Metric	L	Close	Delta [95% CI]	Demo	Embedding
ED disposition	D	AUROC	3	1	+0.020 [0.014, 0.027]	0.722	0.709
Potassium	M	rho	3	1	-0.012 [-0.027, 0.003]	0.287	0.210
Calcium	M	rho	3	1	+0.055 [0.032, 0.077]	0.190	0.184
eGFR	M	rho	3	1	+0.013 [0.007, 0.019]	0.677	0.542
Glucose	M	rho	3	1	+0.044 [0.028, 0.062]	0.320	0.339
Hemoglobin	M	rho	0	4	+0.058 [0.044, 0.071]	0.451	0.355
Albumin	M	rho	3	1	+0.046 [0.034, 0.059]	0.438	0.393
BUN	M	rho	3	1	-0.008 [-0.019, 0.004]	0.511	0.388
Sodium	M	rho	3	1	+0.033 [0.016, 0.048]	0.242	0.230
Atrial fibrillation	S	AUROC	1	3	+0.050 [0.039, 0.060]	0.790	0.806
Heart failure	S	AUROC	2.5	1.5	+0.074 [0.057, 0.092]	0.732	0.783
CKD	S	AUROC	3	1	+0.033 [0.016, 0.048]	0.720	0.725
Diabetes	S	AUROC	2	2	+0.047 [0.035, 0.061]	0.652	0.681
Anemia	S	AUROC	1	3	+0.032 [0.018, 0.047]	0.658	0.654
Resp compromise	S	AUROC	2	2	-0.045 [-0.070, -0.022]	0.770	0.664
VTE	S	AUROC	3	1	-0.012 [-0.034, 0.007]	0.657	0.614
Conduction block	S	AUROC	4	0	+0.003 [-0.023, 0.026]	0.689	0.667
Creatinine	M	rho	3	1	-0.063 [-0.078, -0.047]	0.431	0.250
Magnesium	M	rho	3	1	-0.072 [-0.141, -0.014]	0.152	0.066
NT-proBNP	M	rho	3	1	+0.071 [0.011, 0.135]	0.249	0.326
HbA1c	M	rho	3	1	-0.074 [-0.172, 0.036]	0.182	0.090
CRP	M	rho	3	1	-0.165 [-0.259, -0.067]	0.437	0.155
TSH	M	rho	3	1	-0.001 [-0.112, 0.098]	0.059	0.060
Lactate	M	rho	2.5	1.5	+0.009 [-0.031, 0.045]	0.214	0.184
ABO type	M	AUROC	4	0	-0.046 [-0.087, -0.009]	0.582	0.513
Rh factor	M	AUROC	4	0	-0.050 [-0.104, 0.007]	0.675	0.546

p to five decimals before BH. These archived, encounter-level approximations are unclustered and serve as descriptive provenance. `revision3_statistics_audit.json` reproduces every displayed q under the frozen implementations.

Table 4: **Prespecified five-target attenuation table.** r are Spearman correlations before/after the indicated fixed linear adjustment; ECF is descriptive relative attenuation. BH q-values use the archived Fisher-z approximation described above within exactly these five rows. Potassium's paired ECG values are in Table 11.

Target	Adjustment set	r_{raw}	r_{adj}	ECF	BH q	n
Potassium	age, eGFR, BUN	.2099	.0487	.7681	.0002	6,348
Glucose	age, eGFR, BUN	.3387	.2590	.2354	$<$.0001	6,118
Hemoglobin	age, eGFR, BUN	.3548	.2845	.1983	$<$.0001	6,566
Sodium	age, eGFR, BUN	.2302	.2106	.0849	$<$.0001	6,114
Calcium	age, eGFR, BUN	.1838	.1789	.0269	$<$.0001	6,339

Table 5 reports the separate extension analysis. Its BH family contains seven laboratory targets (creatinine, magnesium, NT-proBNP, HbA1c, CRP, TSH, lactate); only the four meeting the prespecified $|r_{\text{raw}}| \geq .15$ threshold are shown. The target-derivative-free creatinine rerun retained correlation estimates but not p/q values; its entry is therefore unavailable. The invalid eGFR-adjusted run is not used for that entry.

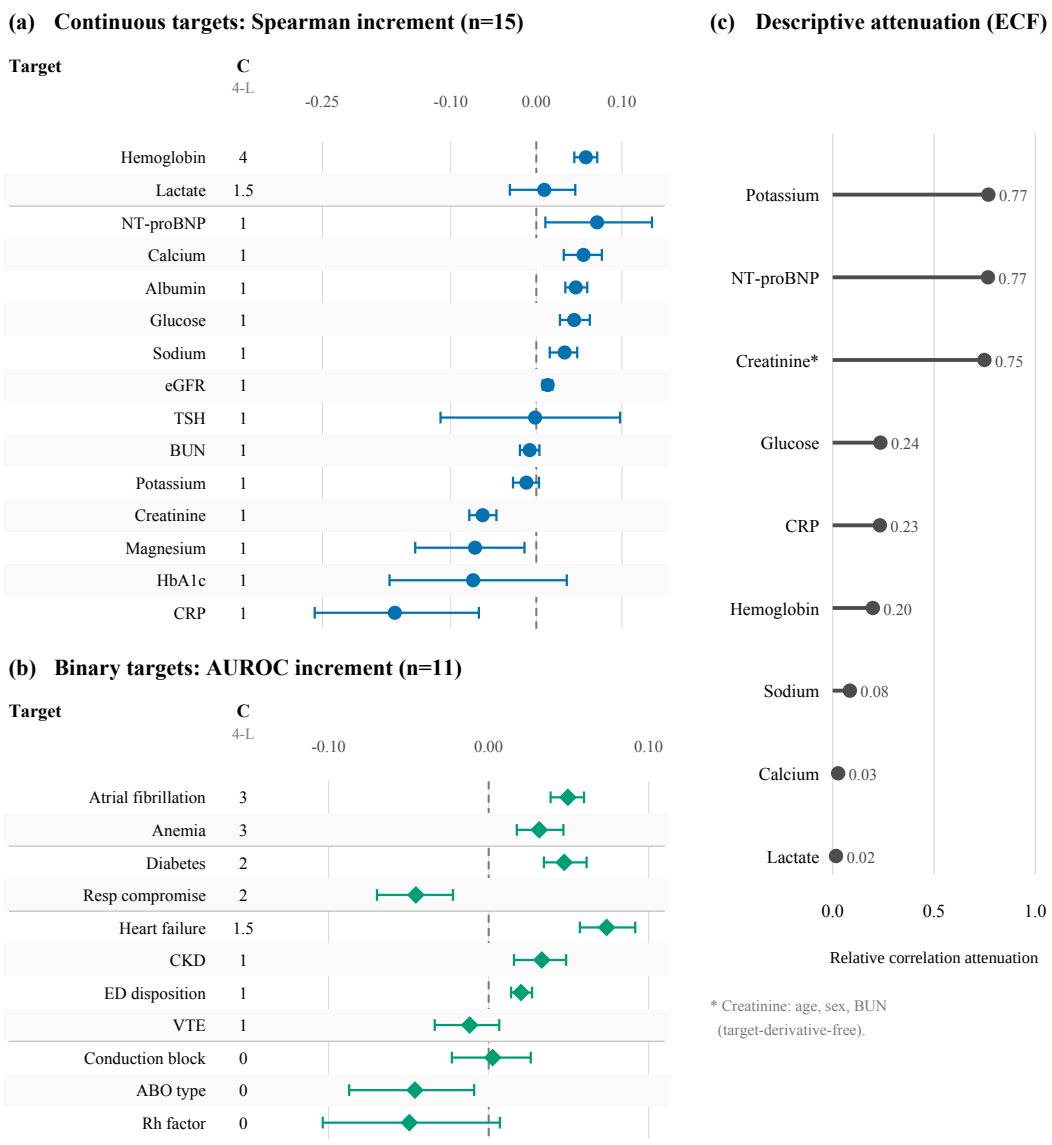


Figure 3: **Frozen mechanism priors and measured increments, stratified by outcome type.** Closeness $= 4 - L$ (higher $=$ mechanistically closer) against incremental skill for (a) continuous targets evaluated by Spearman correlation and (b) binary targets evaluated by AUROC. Each target occupies one forest-plot row with its 95% CI; the adjacent $C = 4 - L$ column displays the frozen tier information alongside each target. The pooled association is sensitive to metric and claim-type handling (statistics in this appendix). (c) Descriptive ECF per target (only targets with $|r_{\text{raw}}| \geq 0.15$; creatinine under the derivative-free set). Disagreements identify cases for subsequent physiological validation.

Table 5: **Threshold-reportable extension targets.** Except for the clean creatinine rerun, BH q-values come from the separate seven-target extension family.

Target	Adjustment set	r_{raw}	r_{adj}	ECF	BH q	n
Creatinine	age, sex, BUN	.2504	.0630	.7484	n.a.	6,960
NT-proBNP	age, eGFR, BUN	.3262	.0763	.7661	.0284	1,278
CRP	age, eGFR, BUN	.1550	.1189	.2329	.0422	384
Lactate	age, eGFR, BUN	.1842	.1811	.0164	$<$.0001	2,617

E PROBE AND NUISANCE-MODEL SENSITIVITIES

These analyses use frozen encounter representations and the unchanged evaluation split. The validation-tuned probe covers three PPG regression targets, with hyperparameters selected on a deterministic 20% subset of training MRNs. The nuisance-model check uses five MRN folds for out-of-fold training residuals and adds paired ECGFounder results for potassium. The three-target PPG results are point estimates; paired potassium ECG−PPG differences include MRN-cluster bootstrap intervals.

Table 6: **Validation-tuned ridge and baseline fallback.** "Guarded" is the model (baseline or combined) selected by validation performance before evaluation. "Fixed" is the post-review re-fit under the recorded scikit-learn 1.7.2 environment. The archived broad table retains its earlier frozen-run CRP value $-.1653$; the re-fit here yields $-.1663$. Each table retains its own run provenance.

Target	Fixed Δ	Tuned base	Tuned combined	Tuned Δ	Selected	Guarded Δ
Potassium	$-.0116$	.2880	.2882	$+.0003$	combined	$+.0003$
Creatinine	$-.0627$	.4322	.3844	$-.0478$	baseline	.0000
CRP	$-.1663$	.4378	.4204	$-.0174$	baseline	.0000

Table 7: **Five-fold MRN-cross-fitted nuisance sensitivity.** The same fixed $\alpha = 1$ representation probe is fitted to cross-fitted residuals.

Target	Adjustment	r_{raw}	Ridge r_{adj}	Ridge attn.	HistGBM r_{adj}	HistGBM attn.
Potassium	age, eGFR, BUN	.2099	.0488	.7675	.0494	.7646
Creatinine	age, sex, BUN	.2504	.0631	.7481	.0497	.8016
CRP	age, eGFR, BUN	.1550	.1191	.2318	.0478	.6913

Table 8: **Cross-fitted paired potassium modality sensitivity.** All rows use the same 6,348 encounters and 4,783 patients. Differences are ECG−PPG; 95% CIs resample MRN clusters 1,000 times with fitted probes held fixed.

Nuisance procedure	PPG r	ECG r	Difference	Cluster 95% CI
None (raw target)	.2099	.3203	+.1104	[.0806,.1403]
Five-fold ridge	.0488	.2696	+.2208	[.1872,.2558]
Five-fold HistGBM	.0494	.2670	+.2176	[.1843,.2543]

Table 9: **Same-first-segment potassium sensitivity.** Both modalities use one central crop from aligned `seg_idx=0`, evaluated individually within each CSN; PPG uses 8 s and ECGFounder uses 10 s after 125-to-500 Hz resampling. Differences are ECG−PPG with 1,000-resample MRN-cluster CIs.

Nuisance procedure	PPG r	ECG r	Difference	Cluster 95% CI
None (raw target)	.1910	.3061	+.1151	[.0847,.1468]
Five-fold ridge	.0334	.2509	+.2174	[.1840,.2548]
Five-fold HistGBM	.0361	.2465	+.2105	[.1761,.2462]

The scripts and machine-readable outputs are:

Revision-2 sensitivity
Script: `revision2_sensitivity.py`
Output: `revision2_sensitivity_results.json`

Matched-window extension
Script: `revision4_matched_window_sensitivity.py`
Output: `revision4_matched_window_results.json`

F INFERENCE SCOPES AND WITHDRAWN EQUIVALENCE ANALYSIS

The broad target table uses encounter bootstrap. Conditional-table q-values retain the archived unclustered Fisher-z approximation with family-specific BH correction. Estimates labeled patient-cluster or MRN-cluster resample patients with fitted probes held fixed. Potassium modality con-

trasts, including Table 8, use the same cluster resamples for both modalities. These intervals quantify held-out evaluation uncertainty conditional on the fitted probes.

Repeated encounters leave the broad-table intervals unsuitable for chance-equivalence inference. The previous CI-inversion equivalence analysis for conduction block and blood-type controls is therefore withdrawn. Reported p-values retain their specified testing procedures; bootstrap-tail fractions are not p-values from those tests.

Table 10: **Expert-tier sensitivity.** The 11-target row covers all binary outcomes, including ED disposition and blood-type controls. Exact two-sided permutation tests are used for the controls-excluded and state-only rows.

Subset	n	Spearman ρ	p
All analyzable targets	26	0.454	0.021
All binary targets (includes ED disposition, ABO, Rh)	11	0.628	0.042
Binary targets excluding ABO/Rh	9	0.368	0.330
Historical disease-state targets only	8	0.279	0.502
Continuous laboratory targets	15	0.290	0.327
Rank-normalized, claim-stratified	26	0.178	0.440

Table 11: **Paired modality differences for potassium** (same 6,348 encounters and 4,783 patients; MRN-cluster bootstrap, 1,000 resamples; probes re-fit once and held fixed).

Quantity	PPG [95% CI]	ECG [95% CI]	Paired ECG - PPG [95% CI]
Raw rho	0.210 [0.183, 0.233]	0.320 [0.296, 0.342]	+0.110 [0.081, 0.142]
Adjusted rho	0.049 [0.023, 0.075]	0.270 [0.243, 0.294]	+0.221 [0.186, 0.257]
Increment Delta	--	--	+0.100 [0.075, 0.124]

G TIMESTAMP AND SYNCHRONIZATION AUDIT

MC-MED documents randomly shifted dates with preserved temporal order; future years reflect this de-identification procedure. In the frozen analysis table, all 106,983 segments from 37,784 CSNs have a parsed aligned start; years range from 2200 to 2302. A representative CSN has visit arrival `2294-09-18 06:46:47`, PPG start `06:47:47.015`, ECG start `06:48:20.721`, and potassium `Order_time 06:49:47`. The preprocessor combines each channel's WFDB `base_date` and `base_time`, trims the earlier channel to the later start, and then extracts common relative windows. PPG–ECG heart-rate disagreement has median 0.66 bpm (IQR 0–1.75) across 106,587 segments with both estimates; 83.5% are within 3 bpm and 88.7% within 5 bpm. This comparison checks functional agreement after record-relative alignment, not independent hardware-clock synchronization.

The processing ledger marks 53,843 CSNs complete, whereas 37,784 contribute at least one analyzed segment. A header-only reconstruction of the frozen build accounts for every difference: all 16,059 completed CSNs without a row have less than 30 s of common signal after start-time alignment (median 0.0 s, 95th percentile 22.7872 s, maximum 29.9840 s). All have readable headers and expected sampling rates, and every analyzed CSN appears in the completion ledger. Thus $53{,}843 - 16{,}059 = 37{,}784$; the duration/alignment rule accounts for the difference. SQI is recorded as an audit field. The reconstruction script and counts are `revision3_selection_audit.py` and `revision3_selection_audit.json`. Segment-level concept extraction succeeds for 99.63% of analyzed windows.

The former timing table is withdrawn. Its code called `pandas.to_datetime(..., errors=coerce)` with nanosecond timestamps; the exact supported upper bound is `2262-04-11 23:47:16.854775807`. The archived run lacks a pandas-version record. Dates beyond that bound became missing, leaving 3,833/6,348 potassium encounters (60.4%) in that timing table. MC-MED defines `Order_time` as order placement, so the table measured order-to-arrival intervals in a sample selected by the parser boundary. The percentage describes retained coverage; record-specific exclusion causes are unavailable in the archived log. This parser was confined to the withdrawn timing script: the main analyses join target labels by CSN and pair wave-

forms using WFDB headers. A wide-range parser permits waveform-to-order gap reconstruction, but specimen collection times are unavailable, so waveform-to-specimen timing cannot be assessed.

H IMPLEMENTATION SPECIFICATION

Table 12 specifies the frozen implementation. Quality fractions are audit fields; target encoders are evaluated without SQI thresholding. Concept decodability uses the listed label-validity masks, whereas target prediction uses the available labels under each task's construction rule.

Table 12: **Label and waveform implementation details.**

Component	Frozen implementation
Primary laboratory labels	Exact `Component_name`: POTASSIUM, CALCIUM, EGFR REFIT WITHOUT RACE (2021), GLUCOSE, HEMOGLOBIN (HGB), ALBUMIN, BLOOD UREA NITROGEN (BUN), SODIUM. Retain a CSN only when exactly one matching row exists; potassium string `>10.0` is set missing.
Extension labels	Numeric repeats are aggregated by CSN median for CREATININE, MAGNESIUM, BNP NT-PRO, HEMOGLOBIN A1C, C-REACTIVE PROTEIN (CRP), TSH, and lactate components `POC:LACTATE, ISTAT` and `LACTATE.WHOLE BLD`. Strings with leading $<$ or $>$ are parsed at the recorded boundary. ABO/Rh requires concordant categorical records.
Historical labels	Patient-level PMH ICD-prefix match: I48, I50, Z95.0/Z95.2/Z95.3, N18, E10/E11, G47.3/E66.2, D50–D53/D64, J96/R09.02/R06.0, I26/I80/I82, and I44/I45; a positive MRN is propagated to all of that patient's CSNs.
Split and aggregation	Deterministic MD5 hash of MRN gives an 80/20 train/eval split. Segment embeddings are averaged within CSN; the first MRN and split fields are retained. Missing labels are removed target-wise.
Pairing and segmentation	WFDB `base_date+base_time`; trim the earlier channel to the later start; retain common duration. Expected rates are ECG 500 Hz and PPG 125 Hz. The archived build retained up to three nonoverlapping complete 30 s windows per CSN; pairs with less than 30 s common duration yield no row.
Concept preprocessing	Fourier resample both channels to 125 Hz; third-order zero-phase Butterworth filters (PPG 0.5–8 Hz, ECG 0.5–40 Hz); within-window z-score. Missing, flatline, and 1st/99th-percentile clipping fractions are recorded. Validity masks are HR 30–220 bpm; IPA/SVRI 0–5; AI -1–2; RI 0–1.5; SQI 0–1; QRS 50–200 ms; QT 250–600 ms; QTc 280–550 ms; systolic amplitude 10^{-6}–100; rise time 0.05–0.60 s; VPG slope ratio 0.05–20; notch position 0.10–0.90; SDNN 5–500 ms; PR 80–320 ms; P-wave duration 40–160 ms.
AnyPPG	Center 8 s at 125 Hz (1,000 samples), per-segment z-score; 512-d segment embeddings, then CSN mean.
PaPaGei-S	For each segment, attempt NeuroKit `ppg_clean`; if NeuroKit is unavailable or raises an exception, silently retain the uncleaned signal, then z-score and take the center 10 s at 125 Hz (1,250 samples). Fallback counts are unavailable in the archived extraction record; segment embeddings are averaged by CSN.
Pulse-PPG	Z-score the 125 Hz PPG and polyphase-resample to 50 Hz; length-agnostic global-max-pool embedding, then CSN mean.
ECGFounder	Native aligned lead-II ECG; center 10 s at 500 Hz (5,000 samples), z-score. One central window is extracted per CSN.

I MATCHED-ARCHITECTURE CONTROL DETAILS

Design. Two identically-architected 1D-CNNs (four conv blocks, $64\rightarrow512$ channels, global average pooling, 512-d embedding; identical seeds, AdamW 3×10^{-4}, cosine schedule, 60 epochs, batch 256) trained *supervised on serum potassium* on the same $85{,}680$ training segments at 125 Hz; only the input modality differs. After training, encoders were frozen and passed through the standard protocol.

Table 13: **Matched CNN results** (n=6,348 eval encounters).

Encoder	Raw rho	Adjusted rho	ECF	n eval
Matched PPG	0.1675	0.0390	0.7672	6,348
Matched ECG	0.3072	0.2644	0.1391	6,348

J SELECTED BASELINE LADDER

Table 14: **Six prespecified baseline exemplars** (held-out performance; Spearman ρ for laboratories, AUROC for binary targets). Columns are a random-label probe, handcrafted waveform features, demographics plus triage vitals, a randomly initialized encoder, and three pretrained PPG encoders. The separate 17-target summary uses the full structured target outputs; this table supplies the six-exemplar comparison.

Target	Rand. label	Handcraft	Demo	Rand. init	PaPaGei-S	AnyPPG	Pulse-PPG
Potassium	-0.010	0.150	0.290	0.170	0.130	0.210	0.161
Conduction block	--	0.600	0.690	0.630	0.600	0.670	0.594
Atrial fibrillation	--	0.650	0.790	0.720	0.680	0.810	0.720
Hemoglobin	--	0.220	0.450	0.250	0.210	0.360	0.280
Glucose	--	0.260	0.320	0.270	0.220	0.340	0.259
Calcium	--	0.070	0.190	0.080	0.050	0.180	0.095

K DESCRIPTIVE SUBGROUP CHECKS

Table 15: **Subgroup point estimates.** Potassium uses Spearman ρ; conduction block uses AUROC, with disease positives in parentheses. Sex is coded M$=0$/F$=1$ in the task table; the archived subgroup-performance script swapped those two display labels, corrected here. The retained results provide descriptive within-stratum performance; subgroup contrast intervals and inferential fairness assessments remain unavailable.

	Potassium			Conduction block		
Stratum	n	PPG	ECG	n (positive)	PPG	ECG
Male	3,039	.243	.301	3,670 (164)	.697	.795
Female	3,308	.190	.327	3,879 (122)	.627	.678
Age <59	3,038	.126	.270	3,645 (64)	.597	.644
Age ≥ 59	3,310	.136	.281	3,905 (222)	.629	.748

L PHYSIOLOGICAL CONCEPT DECODABILITY AND LABEL REPEATABILITY

Held-out R^2 from fixed linear ridge probes measures concept decodability. Each comparison uses the first aligned 30 s segment and the same concept-valid rows across probes. AnyPPG receives its central 8 s at 125 Hz; for this matched-segment control, ECGFounder receives the central 10 s of archived 125 Hz ECG resampled to 500 Hz. This differs from the native 500 Hz ECG used for target prediction and the erasure analysis below. The HR-only baseline uses PPG-derived HR; the demographic baseline uses age and sex. Repeatability compares `seg_idx=0` and 1 within CSN, measuring internal stability of the extracted labels.

Table 16: **Linear concept decodability and label repeatability.** R^2 values come from fixed $\alpha = 1$ ridge probes on the patient-disjoint split. Electrophysiological eval counts range from $6{,}542$ to $7{,}420$ encounters. ρ_{01} and ICC compare the first two valid 30-s windows. Group means use all concepts in each group; repeatability is displayed for the five ECG interval labels. The pulse-feature group corresponds to the archived recoverable category.

Concept/group	PPG	ECGFounder	HR only	Age/sex	ρ_{01}	ICC
Pulse-feature mean	.7200	.4032	.5144	.0154	–	–
PPG morphology mean	.5342	.1940	.1858	.0619	–	–
Electrophysiology mean	.0951	.4883	.0529	.0286	–	–
QRS duration	.0243	.5217	.0131	.0204	.8731	.8749
QTc Fridericia	.1980	.5509	.1607	.0319	.8262	.8223
QTc Bazett	.0479	.4530	.0219	.0159	.7841	.7726
PR interval	.1243	.4768	.0685	.0581	.7421	.7394
P-wave duration	.0808	.4393	.0001	.0167	.7517	.7561

The HR-only baseline nearly matches AnyPPG's rise-time decodability (.8928 vs. .8876), showing comparable prediction from rate alone in this comparison. Full per-concept values and valid fractions are in `revision4_concept_controls_results.json`.

M RESTRICTED LINEAR CONCEPT-SUBSPACE ERASURE AND REFITTING

This stress test starts from the 37,784 encounters with a common first aligned segment and the frozen patient-disjoint split; potassium results use the target-available encounters. It measures the change in held-out potassium correlation after removing a selected linear subspace and refitting the target probe. AnyPPG receives its central 8 s at 125 Hz; ECGFounder receives the native-500 Hz central 10 s, unlike the resampled ECG in Table 16. The concept basis uses the five jointly valid ECG interval labels. After training-only PCA-whitening, the rank-five left singular subspace of the representation–concept cross-covariance is removed. The conditional basis first residualizes each concept on age, eGFR, and BUN. A linear-decoding manipulation check measures how accessible the selected labels remain after erasure. In the raw-concept analysis, mean held-out concept R^2 changes from .0950 to .0021 for AnyPPG and from .3490 to .0148 for ECGFounder; for covariate-residualized concepts it changes from .0612 to $-.0003$ and from .3142 to $-.0003$, respectively. The remaining training cross-covariance is below 8×10^{-16} of its original Frobenius norm.

Table 17: **Potassium probe after linear concept-subspace erasure.** Entries under the two erasures are post-erasure Spearman r with loss from the baseline in parentheses. C_{phys} is the five-interval concept vector and $C_{\mathrm{phys}} \perp A$ residualizes it on age, eGFR, and BUN. Shuffled and HR columns report loss; random is the 95% range of losses across 100 rank-matched random subspaces. Residual targets use five-fold MRN-cross-fitted nuisance models. Cluster intervals for all observed concept-erasure losses exclude zero; for every displayed target and modality, both concept losses exceed all 100 random losses (empirical one-sided $p = .0099$).

Modality	Target	Baseline	Erase C_{phys}	Erase $C_{\mathrm{phys}} \perp A$	Shuffled	HR	Random 95%
AnyPPG	Raw	.1909	.0966 (.0944)	.1697 (.0212)	.0006	.0183	[−.0015,.0042]
AnyPPG	Ridge residual	.0334	.0169 (.0165)	.0178 (.0156)	.0005	.0128	[−.0022,.0034]
AnyPPG	HistGBM residual	.0339	.0126 (.0212)	.0153 (.0186)	.0006	.0165	[−.0021,.0036]
ECGFounder	Raw	.3203	.2843 (.0359)	.3037 (.0165)	.0019	.0054	[−.0014,.0031]
ECGFounder	Ridge residual	.2696	.2617 (.0079)	.2615 (.0081)	.0016	.0021	[−.0011,.0035]
ECGFounder	HistGBM residual	.2670	.2588 (.0082)	.2587 (.0084)	.0018	.0025	[−.0013,.0034]

For the covariate-residualized basis and ridge-adjusted target, AnyPPG's correlation falls from .0334 to .0178 (loss .0156), whereas ECGFounder's falls from .2696 to .2615 (loss .0081). The paired ECG-minus-PPG loss difference is $-.0076$ (MRN-cluster 95% CI [−.0139,−.0016]); for the HistGBM-adjusted target it is $-.0102$ ([−.0177,−.0033]). ECGFounder has higher linear interval decodability but a smaller absolute loss in this comparison. AnyPPG's .0156 loss is close to its HR-subspace loss of .0128, so this places its erasure sensitivity near that of rate-related directions. The starting correlations, selected concept basis, and probe refit define the comparison of absolute losses. Full precision and bootstrap outputs are in `revision5_concept_subspace_ablation_results.json`.

N TARGET–CONCEPT MAP AND AUDIT COVERAGE

Table 18: **Per-claim concept links and diagnostic coverage.** Candidate concepts specify physiological hypotheses. Completed diagnostics state the measured evidence; follow-up validation identifies the next empirical question. Concept scores use fixed linear ridge probes unless specified otherwise.

Target	Candidate concepts	Completed diagnostics	Follow-up validation
Potassium	QRS, PR, P-wave duration, QTc; broader ECG morphology	PPG/ECG concept scores; adjusted association; same-A concept and representation increments; paired contrasts; linear erasure and refit	Interval specificity beyond rate; broader concept readouts and targeted predictor interventions
Creatinine	Proximal waveform concepts await validation	Target-derivative-free adjustment; ridge/HistGBM nuisance sensitivity	Reference-labeled concept set linked to the clinical target
Conduction block	QRS/PR morphology	Interval concept scores; increments for diagnosis history	Contemporaneous diagnosis and target-linked interval interventions
Atrial fibrillation	Beat timing and pulse irregularity	HR concept scores; increments for diagnosis history	Irregularity labels and concurrent rhythm adjudication
Hemoglobin	Optical amplitude and pulse morphology	Selected PPG concept scores; target association	Target-linked concept interventions and external reference validation

The potassium example covers decodability, incremental prediction, and erasure. For other targets, the table distinguishes candidate concepts from completed diagnostics and the remaining target-linked validation.

O ENCODER–DATASET EXPOSURE AUDIT

Table 19: **Dataset-level pretraining exposure.** "Unresolved" means the reviewed model report does not establish whether the encoder used the dataset. Documented entries establish dataset-level exposure; patient-level overlap remains unknown under the available released identifiers.

Encoder	MC-MED	VitalDB	WildPPG	PPG-DaLiA
AnyPPG	documented exposure	via PulseDB	unresolved	unresolved
PaPaGei-S	unresolved	documented exposure	unresolved	unresolved
Pulse-PPG	unresolved	unresolved	unresolved	unresolved
ECGFounder	unresolved	unresolved	n/a	n/a

AnyPPG's pretraining data include PulseDB, which incorporates VitalDB waveforms (Nie et al., 2026); its VitalDB exposure is therefore through PulseDB. The reviewed records also document AnyPPG pretraining exposure to MC-MED and PaPaGei-S exposure to VitalDB. ECGFounder's MC-MED exposure remains unresolved. Cohort-specific probe refitting evaluates association patterns conditional on these exposure records; it does not establish patient-level nonoverlap. Table 20 identifies the local weights. Upstream release tags and repository commits are absent from the archived extraction record; checkpoint hashes supply the available weight provenance.

Table 20: **Frozen checkpoint identifiers.** Full hashes are in `checkpoint_manifest.json`.

Encoder	Checkpoint file	SHA-256 prefix
AnyPPG	`anyppg_ckpt.pth`	`99b9bb0a3c2b`
PaPaGei-S	`papagei_s.pt`	`79d68671e51b`
Pulse-PPG	`checkpoint_best.pkl[net]`	`485ade5033b3`
ECGFounder	`1_lead_ECGFounder.pth`	`f863a38897fb`

P DATASET SOURCES, VERSIONS, AND LICENSES

MC-MED v1.0.1, PhysioNet credentialed access (`physionet.org/content/mc-med/1.0.1`) (**??**); de-identified emergency-department waveforms and laboratories used under its data use agreement. **VitalDB**, `vitaldb.net/dataset` (open research license); 6,119 of 6,388 public cases pass availability and quality filters. **WildPPG**, Hugging Face `eth-siplab/WildPPG` (CC BY-NC-SA 4.0). **PPG-DaLiA**, UCI Machine Learning Repository (CC BY 4.0). **Encoders**: AnyPPG (KDD 2026), PaPaGei-S (ICLR 2025), Pulse-PPG (IMWUT 2025), ECGFounder (NEJM AI 2025), used as released. The expert rating task involved no patient data and is described in Appendix A.

Q CONCEPT AND REPRESENTATION INCREMENTS OVER THE SAME ADJUSTMENT SET

Cohort and comparisons. The potassium comparison uses the common complete-case cohort with the target, adjustment variables, and all five reference concepts available: 20,734 training encounters and 5,160 evaluation encounters from 3,996 evaluation patients. The adjustment set is $A = \{\text{age}, \text{eGFR}, \text{BUN}\}$. The reference vector C_{ref} contains QRS duration, PR interval, P-wave duration, and Fridericia- and Bazett-corrected QT, measured by NeuroKit from record-aligned ECG. The four main feature sets are A, (A, C_{ref}), $(A, \widehat{C}_m)$, and (A, z_m), where $\widehat{C}_m$ denotes the five reference concepts decoded from modality m and z_m its full representation. Reference-concept-only, representation-only, and $(A, \text{shuffle}(C_{\text{ref}}))$ models were also fitted. The concept-free models are evaluated on the same complete-case rows as the concept models. The 5,160-encounter same-A evaluation cohort is distinct from the 6,348-encounter matched potassium cohort used for the

Table 21: **Potassium increments over the same adjustment set.** Entries are paired differences in held-out Spearman correlation relative to A alone, with MRN-cluster 95% CIs. All rows use the common 5,160-encounter evaluation cohort (3,996 patients). The two panels specify the target predictor; concept decoding uses ridge with $\alpha = 1$ throughout.

Added features	AnyPPG	ECGFounder
Ridge target predictor		
C_{ref}	$+.0047$ $[.0005, .0085]$	$+.0047$ $[.0005, .0085]$
$\widehat{C}_m$	$+.0077$ $[.0019, .0135]$	$+.0050$ $[.0005, .0089]$
z_m	$-.0084$ $[-.0193, .0029]$	$+.0870$ $[.0647, .1103]$
$\text{shuffle}(C_{\text{ref}})$	$+.0000$ $[-.0010, .0011]$	$+.0000$ $[-.0010, .0011]$
HistGBM target predictor		
C_{ref}	$+.0058$ $[-.0010, .0120]$	$+.0058$ $[-.0010, .0120]$
$\widehat{C}_m$	$+.0073$ $[-.0004, .0157]$	$+.0034$ $[-.0025, .0086]$
z_m	$+.0077$ $[-.0003, .0153]$	$+.0872$ $[.0684, .1054]$
$\text{shuffle}(C_{\text{ref}})$	$+.0010$ $[-.0033, .0052]$	$+.0010$ $[-.0033, .0052]$

residual-target and paired-modality analyses. The same-A increment predicts original potassium, whereas the adjusted-correlation analysis predicts potassium residualized on A.

Fitting and model selection. Concept decoding uses ridge regression with $\alpha = 1$ and does not use the potassium labels. Five MRN-hash folds supply out-of-fold concept predictions for the training encounters; a concept probe fitted on all training encounters supplies evaluation predictions. Each second-stage target model selects its hyperparameter on MRN-hash fold 0 of the training split and is then refitted on the full training split. Ridge selects α; HistGBM selects `max_leaf_nodes`. The ridge search was extended to $\alpha = 10^7$ after initial fits selected the earlier upper limit of 1000; the reported selected values were below the extended upper limit. The ridge and HistGBM columns therefore refer to the target predictor, while the concept decoder remains ridge in both cases.

Paired uncertainty. Table 21 reports the increment in held-out Spearman correlation relative to the model fitted using A alone. Each of 1,000 bootstrap replicates resamples evaluation MRN clusters, using the same resample across the compared feature sets. Fitted probes remain fixed, so these intervals describe evaluation-sampling uncertainty conditional on the fitted models. Identical entries for the reference-concept and shuffled-reference-concept additions are shared covariate-based comparisons, not independent modality replications.

Exploratory extension. The extension also considered creatinine with $A = \{\text{age}, \text{sex}, \text{BUN}\}$ and CRP with $A = \{\text{age}, \text{eGFR}, \text{BUN}\}$. The creatinine HistGBM run showed a positive increment for the shuffled reference-concept block, complicating interpretation of small concept increments. For CRP, complete-case evaluation contained only 268 encounters from 257 patients, and the ridge shuffled-concept increment was $+.0636$. These control results remain exploratory diagnostics for the two targets.

Implementation record. The analysis script is `revision6_concept_target_ladder.py`, with outputs in `revision6_concept_target_ladder_results.json`; it is included in `run_revision_audits.sh`. The recorded environment uses Python 3.12, scikit-learn 1.8.0, and NumPy 2.x on CPU.

R Additional results and controls

Baseline summary provenance. The six-exemplar ladder in Table 14 compares Pulse-PPG with random-label, random-initialization, and handcrafted controls. The 17-target summary in §4.8 uses all-target side-by-side covariate, embedding, and combined scores in the structured tier and demo-dominance files.

Confound-set and probe ablations. Figure 4 shows potassium/conduction increments across four adjustment sets and under the HistGBM probe.

Subgroup checks. Within sex and age strata, ECG exceeds PPG for potassium and conduction block in every reported point estimate; historical-label classifiers have ECE < 0.05. Because stratum-

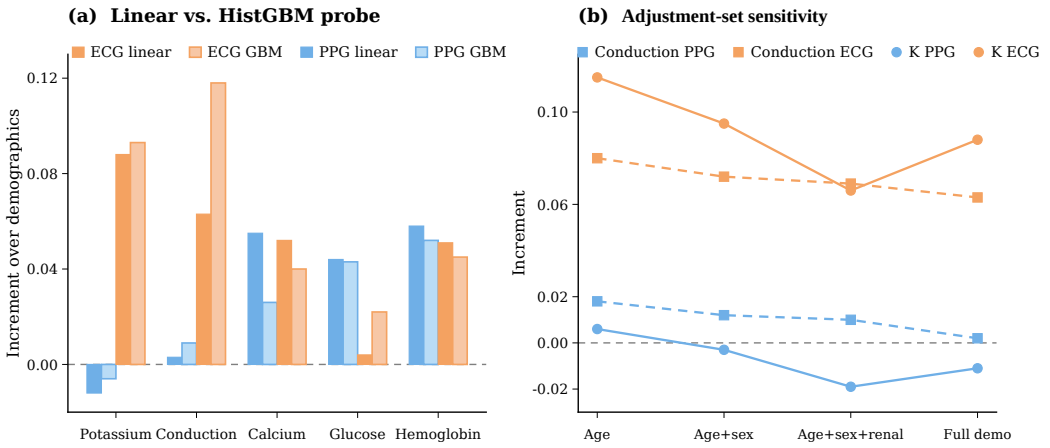


Figure 4: **Ablations.** (a) Probe ablation: potassium and conduction PPG increments remain near zero under the nonlinear HistGBM probe. (b) Confound-set sensitivity: the point estimates preserve the PPG–ECG ordering across four adjustment sets. The released ablation table reports point estimates; the displayed ordering summarizes the tested settings.

specific contrast intervals are unavailable, these are descriptive ordering and calibration checks only (Table 15).

Wearable concept probes. On PPG-DaLiA, pulse-timing labels retain positive decodability (rise time R^2=0.84, HR 0.73). This wearable analysis covers the available timing labels; noisy wearable ECG and unavailable laboratory references constrain its physiological coverage.

SIGMA-PPG inspection. The inspected checkpoint uses an 18-block LaBraM-style ViT with 240 s/50 Hz inputs. Its audit is deferred pending release of the authors' forward code alongside the checkpoint.