PEER-REVIEWED · OPEN ACCESS
Sriwijaya Journal of Forensic and Medicolegal
Vol. 4, No. 1 (2026), pp. 44–52 · e-ISSN 2987-1530 · Article ID: 388
journal homepage: phlox.or.id/index.php/sjfm

Original Research Article

Internal Validation of an NMR Aqueous-Humor Multimetabolite Model for Post-Mortem Interval Estimation

Bambang Sutrisno[1], **Sri Mulyati**[1], **Karina Chandra**[2*], **Jason Willmare**[3]

[1]*Department of Health Sciences, Emerald Medical Center, Jakarta, Indonesia*
[2]*Department of Anesthesiology and Intensive Care, CMHC Research Center, Palembang, Indonesia*
[3]*Department of Neurology, Diffa State Hospital, Diffa, Niger*

ARTICLE INFO

Article history
Received: February 10, 2026
Revised: March 20, 2026
Accepted: April 19, 2026
Available online: May 15, 2026
Published: May 30, 2026

Keywords
Aqueous humor
Metabolomics
Nuclear magnetic resonance
Post-mortem interval
Ridge regression

Corresponding author
Karina Chandra
E-mail:
karina.chandra@cattleyacenter.id
ORCID: 0009-0004-0194-6110

DOI
10.59345/sjfm.v4i1.388

ABSTRACT

Background. Molecular changes in protected ocular compartments may complement conventional post-mortem interval (PMI) estimation, but association does not guarantee accurate individual prediction.

Objective. To internally validate a ridge-regression model that estimates PMI from nuclear magnetic resonance (NMR)–quantified metabolites in human aqueous humor.

Methods. This secondary analysis used 33 human samples and 41 metabolites from Animal Metabolite Database (AMDB) Experiment 242 (human samples). Age, sex, cause of death, temperature, and sampling details were unavailable. Concentrations were log(1 + x)-transformed. Ridge regression was tuned by four-fold inner cross-validation and evaluated by five-fold outer cross-validation repeated 20 times; metrics used the mean of 20 out-of-fold predictions per sample. Spearman associations used Benjamini–Hochberg false discovery rate (FDR) control.

Results. PMI ranged from 5.0 to 58.6 hours. Mean absolute error was 7.04 hours (95% CI 5.06–9.33), root mean squared error (RMSE) 9.48 hours (6.63–12.45), R^2 0.233 (interval extending below zero), and Pearson r 0.485. The mean-only benchmark RMSE was 10.94 hours, so the improvement of 1.46 hours (13.4%) has no interval. Six metabolites passed FDR control (hypoxanthine, $\rho = 0.707$). Predictions were compressed toward the center: samples under 12 hours were overestimated by 8.2 hours on average and the 58.6-hour sample underestimated by 30.6 hours.

Conclusion. These data suggest a detectable PMI-related signal in human aqueous humor, but individual precision is inadequate and the model remains a proof of concept. Locked external validation with standardized sampling and clinical and environmental covariates is required before medicolegal use.

Copyright. © 2026 The Author(s). Published by Phlox Institute: Indonesian Medical Research Organization, on behalf of the Sriwijaya Journal of Forensic and Medicolegal. The authors retain copyright of their work.

Access licence. This is an open access article distributed under the terms of the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 International Licence (CC BY-NC-SA 4.0), which permits non-commercial use, sharing, adaptation and reproduction in any medium, provided the original author and source are credited and any derivative work is distributed under the same licence.

1. Introduction

The post-mortem interval (PMI) is the time between death and examination. It anchors the reconstruction of events, the evaluation of witness accounts, and the interpretation of autopsy and toxicology findings. A single measurement rarely yields it. Body cooling, livor and rigor mortis, ocular changes, and decomposition are combined with environmental information, and each is affected by temperature, clothing, body habitus, cause of death, and the interval before refrigeration.[1,2] A PMI estimate is therefore best understood as an uncertainty-bounded synthesis, not an exact timestamp. That limitation has encouraged the search for objective measurements, including three-dimensional imaging, corneal optical coherence tomography, computed tomography radiomics, and microbiome profiling.[3–7]

Thanatochemistry measures post-mortem biochemical change in body fluids. Vitreous potassium is the best-studied analyte. It rises as membrane gradients collapse,[8,9] and recent human studies continue to evaluate it, often together with other analytes.[10–13] Hypoxanthine, a product of adenine nucleotide degradation, has been measured in human vitreous humor with high-performance liquid chromatography (HPLC).[14,15] A strong correlation between an analyte and PMI does not guarantee a narrow prediction interval for an individual decedent, however. Regression equations depend on analytical procedure, population, and temperature.[9,15]

Vitreous humor fills the vitreous chamber behind the lens. Aqueous humor, by contrast, is a smaller, continuously renewed fluid in the anterior and posterior chambers in front of the lens. The two matrices differ in volume, protein content, turnover, and tissue contact, so a time-dependent signal in vitreous humor cannot be assumed to transfer to

aqueous humor. Aqueous humor deserves matrix-specific study. Its low protein content and protected location suit spectroscopic profiling of small metabolites,[16] and it has been studied far less than vitreous humor;[17] spectroscopic work on human vitreous humor illustrates the approach.[18,19] Concentrations at sampling still reflect ante-mortem physiology, agonal events, post-mortem diffusion, temperature, and handling. A multimetabolite profile is better viewed as a time-sensitive biochemical state than as a clock.

Metabolomics measures many low-molecular-weight compounds at once. It can capture interconnected pathways that change after circulation stops, such as purine degradation, anaerobic metabolism, tricarboxylic acid cycle intermediates, amino acid release, and phospholipid breakdown. Nuclear magnetic resonance (NMR) spectroscopy quantifies such compounds reproducibly without compound-specific ionization. Human work links the metabolome to PMI and shows that sampling time must be built into study design,[20,21] and multi-organ metabolomic models have been developed under controlled conditions.[22,23] NMR metabolomics of ovine vitreous humor combined with potassium has reported interval-specific prediction errors.[24] Ovine aqueous humor has also been profiled by NMR. In that model, the metabolomic profile predicted PMI better than potassium measured in the same experiment during the first 24 hours.[17,25] The closest human precedent is a study that translated an animal NMR model to human autopsy material.[26] The deposit analyzed here derives from a human ^{1}H NMR study of serum, aqueous humor, and vitreous humor,[27] whose sample-level aqueous-humor data are public in AMDB Experiment 242.[28]

Association is not prediction. Association asks whether concentrations rise or fall with PMI. Prediction asks how close an estimate is for a sample not used to build the model. With 33 samples and 41 correlated candidate predictors, unpenalized regression is unstable, and feature selection on the full dataset would inflate performance. Ridge regression with tuning inside the training data and evaluation on held-out samples addresses both problems.[29] Reporting follows TRIPOD+AI. PROBAST+AI guided the risk-of-bias considerations.[30,31]

The data are not new, because this is a re-analysis of a public deposit. The unresolved gap is that time-related metabolic change in human aqueous humor has not been translated into a transparent estimate of out-of-sample error and calibration. This study adds a nested-validation evaluation of the human aqueous-humor profile. It covers sample-level error, calibration, a mean-only benchmark, and sensitivity analyses, and it keeps association analysis separate from prediction. The primary aim was to internally validate a ridge-regression model for continuous PMI estimation from 41 NMR-quantified metabolites. Secondary aims were to compare it with the benchmark, identify FDR-controlled metabolite–PMI associations, and characterize calibration and the error distribution. A final aim was to check how the results changed when the log transformation was omitted and when the maximum-PMI sample was excluded.

2. Methods

2.1 Study design and data source

This secondary analysis developed and internally validated a continuous-outcome prediction model. Its structure followed TRIPOD+AI,[30] and PROBAST+AI domains were considered for the data source, predictors, outcome, and analysis.[31] No new specimens were collected. No participants were contacted. Data came from AMDB Experiment 242, "Human aqueous humor post-mortem," linked to the study by Zelentsova et al.[27] (database record[28]). The samples are human. In the source study, samples were collected by a regional forensic medicine bureau in Novosibirsk.[27] The repository provides a concentration table (metabolites in rows, samples in columns; nmol/g wet tissue) and a property table with time post-mortem in minutes. The concentration matrix was transposed and joined to PMI by repository sample identifier. All 33 samples had PMI and all 41 metabolite concentrations. None was excluded, and PMI was converted to hours only for presentation.

The deposit contains no age, sex, cause or manner of death, setting, body or ambient temperature, storage interval, time from death to sampling, freeze–thaw history, or technical replicates. None of these could be used as a covariate. Identifiers cannot be linked to decedents, so each was treated as an independent sample. If a decedent contributed more than one sample, outer folds could share information, and record-wise validation of this kind can overestimate accuracy.[32] All performance estimates would then be optimistic. PMI was analyzed as a continuous variable and was not dichotomized.

2.2 Preprocessing

Concentrations were transformed as $\log(1 + x)$. This form is defined for zero and positive values, learns no parameter from the data, and reduces right skew. The transformation was chosen after the distributions of all 33 samples had been inspected, without reference to PMI, and a sensitivity analysis omitted it. Standardization of the transformed predictors took place within the ridge fitting procedure. Means and standard deviations were estimated from the training data only. No metabolite filtering, variance screening, or imputation was applied.

2.3 Ridge regression and repeated nested cross-validation

Ridge regression was chosen because predictors outnumbered samples and were expected to be correlated. Models were fitted with the glmnet algorithm[33] using the mixing parameter $\alpha = 0$, which gives a pure ridge (L2) penalty. PMI was modeled in minutes, and the penalty λ depends on that scale. A logarithmically spaced grid from 10^{4} to 10^{-3} supplied the candidate values from which λ was selected.

Validation used five-fold outer cross-validation repeated 20 times. Nested designs avoid the strong optimistic bias that simple K-fold validation shows in small samples.[34] Each repetition followed six steps. (1) Samples were assigned to five outer folds so that the PMI range was distributed across folds. (2) Within each outer training set, four-fold inner cross-validation selected λ. (3) The model was refitted on the whole outer training set. (4) It then predicted the held-out outer fold. (5) After all folds and repetitions, the 20 out-of-fold predictions for each sample were averaged. (6) Performance was calculated from the 33 observed values and 33 averaged predictions. observations played no part in choosing standardization parameters, λ, or coefficients.

The 660 prediction instances (33 samples × 20 repetitions) are repeated predictions of the same 33 samples under different training compositions. They were used to stabilize the sample-level estimate and to describe partition sensitivity, not to increase the effective sample size. The sample is the unit of analysis. Averaging smooths

predictions and can yield lower error than a single model trained once. The reported metrics therefore describe the averaged out-of-fold predictions, and they are not the expected error of one deployed model. The standard deviation of each sample's 20 predictions was retained as a measure of partition sensitivity. It is not an interval for the true PMI.

2.4 Mean-only benchmark

In each outer split, the benchmark predicted every test observation by the mean PMI of the outer training set. It used the same partitions as the ridge model. Improvement was expressed as the RMSE difference in hours and as a percentage of the benchmark RMSE. Mean absolute error was also compared.

2.5 Performance measures and calibration

Mean absolute error (MAE), root mean squared error (RMSE), and predictive R^2 were calculated as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i is the observed PMI, $\hat{y}_i$ the averaged out-of-fold prediction, $\bar{y}$ the mean observed PMI, and n = 33. Predictive R^2 can be negative, and it differs from the squared correlation. Pearson r described linear and Spearman ρ rank association; neither was treated as agreement. Signed error was predicted minus observed PMI, so positive values denote overestimation.

Calibration was assessed by regressing observed on predicted PMI.[35] The slope of this regression equals $r \times SD_{obs}/SD_{pred}$, so a slope near 1 can coexist with compressed predictions when r is low. For this reason, compression was also assessed from the spread of predictions and from the slope of predicted on observed PMI. Residuals were plotted against both observed and predicted PMI. Any imperfect predictor whose predictions vary less than the outcome shows a negative association between error and observed PMI (regression to the mean), which is not by itself evidence of a model-specific defect. Absolute errors were summarized by median, 80th and 95th percentiles, and counts at or below 5, 10, 12, 15, and 20 hours. These are descriptive values, not acceptability thresholds. Signed errors were also summarized for PMI below 12 hours, 12 to 24 hours inclusive, and above 24 hours. These strata are exploratory and defined on the outcome.

2.6 Uncertainty

Percentile bootstrap 95% confidence intervals were calculated by resampling the 33 observed–averaged-prediction pairs for MAE, RMSE, R^2, correlations, and calibration measures. The intervals describe uncertainty in these pairs only. They do not include model-development uncertainty or between-laboratory heterogeneity. They are not prediction intervals for an individual decedent, and they would be too narrow if samples were clustered within decedents.

2.7 Exploratory metabolite–PMI associations

Each metabolite was related to PMI by Spearman rank correlation on the original concentration scale, with a two-sided p value. The Benjamini–Hochberg procedure controlled the false discovery rate across all 41 tests,[36] and adjusted values below 0.05 were called FDR-significant. Bootstrap intervals were calculated for each coefficient. This analysis was separate from model development. All 41 metabolites entered the ridge model regardless of their association, and the ranking was not used to choose features, transformation, folds, or λ.

2.8 Sensitivity and reproducibility analyses

Two sensitivity analyses repeated the nested validation. One used untransformed concentrations, and the other excluded the sample with the maximum PMI (58.6 h; n = 32), which changes both sample size and outcome range. For reproducibility, the primary values were those of the original glmnet run. As a check, metrics were recalculated from the out-of-fold prediction table rounded to one decimal hour, which can differ slightly from unrounded values. An independent Python repeated nested-validation ridge implementation was also run on the public data as reconstructed from the repository tables. The Python run followed the same principles but not the same fold-generation sequence, scaling, or λ parameterization.

2.9 Ethics

Ethical approval was granted by the CMHC Research Center Ethics Committee Indonesia (No. 2025/013). The study used a de-identified public dataset. Consent and collection procedures were those of the source study, which reported conduct in accordance with the Declaration of Helsinki and approval by the ethics committee of the International Tomography Center SB RAS (ECITC-2017-03).[27] The analysis generated no decisions for individual medicolegal cases.

3. Results

3.1 Cohort and outcome distribution

The analytic cohort comprised all 33 samples, each with PMI and 41 metabolite concentrations (1,353 values, none missing). PMI ranged from 300 to 3,516 minutes (5.0–58.6 h). The mean was 20.57 hours (SD 10.99), the median 18.2 hours (first quartile 13.0, third quartile 26.4; interquartile range 13.4), and the distribution was concentrated within the first 30 hours with a maximum at 58.6 hours. Concentration ranges differed markedly across metabolites, and several were right-skewed. Table 1 gives the dataset characteristics and the FDR-significant associations.

Table 1. Dataset characteristics and FDR-significant metabolite–PMI associations.

Panel A. Dataset characteristics

Characteristic	Value	Comment
Samples	33	Human aqueous humor, AMDB Experiment 242
Metabolites	41	NMR-quantified, nmol/g wet tissue
Missing concentration cells	0	1,353 values
PMI, mean (SD), h	20.57 (10.99)	—
PMI, median (IQR), h	18.2 (13.0–26.4)	IQR width 13.4 h
PMI range, h	5.0–58.6	300–3,516 min

Panel B. FDR-significant Spearman associations with PMI

Metabolite	Spearman ρ	Bootstrap 95% CI	FDR-BH
Hypoxanthine	0.707	0.437–0.866	0.00017
Choline	0.644	0.346–0.828	0.00106
Succinic acid	0.545	0.234–0.762	0.01411
Betaine	0.518	0.196–0.764	0.02077
Glutamic acid	0.474	0.089–0.762	0.04206
Fumaric acid	0.467	0.109–0.732	0.04206

Notes: CI, confidence interval; FDR-BH, Benjamini–Hochberg false discovery rate across 41 tests; IQR, interquartile range; PMI, post-mortem interval; SD, standard deviation.

3.2 Primary out-of-fold performance

The log(1 + x) ridge model had an MAE of 7.04 hours (95% CI 5.06–9.33) and an RMSE of 9.48 hours (95% CI 6.63–12.45) (Table 2). R^2 was 0.233, and its bootstrap interval extended below zero. Pearson r was 0.485 and Spearman ρ 0.580, a moderate tendency for higher predictions to accompany higher observed PMI. Predictions occupied a narrower range (6.0–35.3 h) than observed PMI (5.0–58.6 h). The mean predicted PMI (20.58 h) matched the mean observed PMI (20.57 h), but the SD of predictions was 5.85 hours against 10.99 hours.

3.3 Benchmark comparison

The training-fold mean-only benchmark had an MAE of 8.52 hours and an RMSE of 10.94 hours (R^2 about −0.02, calculated from the benchmark RMSE and the observed SD). Ridge regression reduced MAE by 1.48 hours and RMSE by 1.46 hours (13.4% of benchmark RMSE). The RMSE interval for the ridge model (6.63–12.45 h) contains the benchmark RMSE, and no interval was estimated for the difference, so the improvement cannot yet be distinguished from sampling noise.

Table 2. Out-of-fold prediction performance, sensitivity analyses, and range-specific error.

Panel A. Model performance and sensitivity analyses

Analysis	n	MAE, h	RMSE, h	R^2	Pearson r
Primary log(1 + x) ridge	33	7.04 (5.06–9.33)	9.48 (6.63–12.45)	0.233	0.485
Mean-only benchmark	33	8.52	10.94	−0.02*	—
Raw-concentration ridge	33	7.30	10.01	0.145	0.388
Excluding maximum PMI	32	6.55	8.06	0.124	0.358
Independent Python check	33	6.84	9.34	0.256	—

Panel B. Magnitude and direction of error across the observed PMI range

PMI stratum	n	MAE, h	Mean signed error, h	Directional pattern
Overall	33	7.04	+0.01	Near-zero aggregate bias
<12 h	7	9.36	+8.21	Predominant overestimation
12–24 h	14	5.39	+2.34	Mild overestimation
>24 h	12	7.64	−7.49	Predominant underestimation

Notes: Parentheses show sample-level bootstrap 95% confidence intervals. MAE, mean absolute error; RMSE, root mean squared error. Predictive R^2 = 1 − (out-of-fold residual sum of squares)/(total sum of squares around the observed mean). Signed error = predicted − observed PMI. Strata are exploratory and defined on the outcome (12–24 h includes 24.0 h). *Calculated from the benchmark RMSE (10.94 h) and the observed SD (10.99 h); the per-repetition range was not reported.

3.4 Calibration and residual behavior

The calibration slope of observed on predicted PMI was 0.913, with an intercept of 1.83 hours (from the rounded predictions). The slope of predicted on observed PMI was about 0.26 (calculated as $r \times SD_{pred}/SD_{obs}$), which expresses the compression. Mean signed error was +0.01 hours and the median 0.0 hours, but error direction depended on PMI (Figure 1, Table 2, Panel B): among 7 samples under 12 hours, mean signed error was +8.21 hours (MAE 9.36); among 14 samples at 12 to 24 hours, +2.34 hours (MAE

5.39); and among 12 samples above 24 hours, −7.49 hours (MAE 7.64).

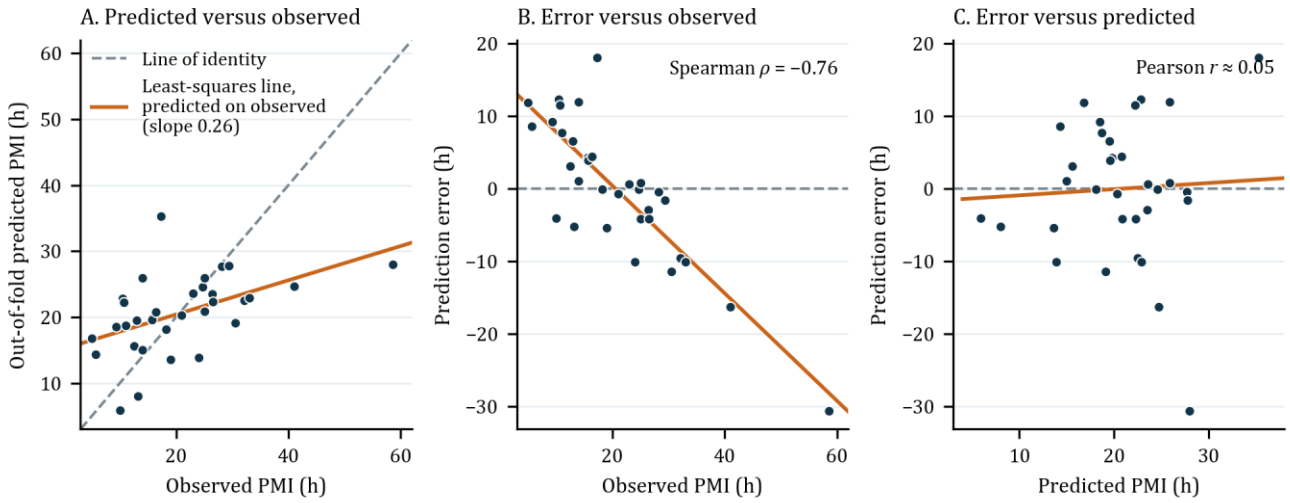


Figure 1. Aggregated out-of-fold predictions and error patterns (n = 33). (A) Predicted versus observed PMI; the dashed line is perfect agreement and the solid orange line is the least-squares fit of predicted on observed PMI (slope 0.26). (B) Prediction error (predicted − observed) versus observed PMI with the least-squares line; the dashed line marks zero error. (C) Prediction error versus predicted PMI with the least-squares line.

The Spearman correlation between observed PMI and signed error was −0.76. A negative correlation of about −0.85 is expected for any predictor with r = 0.485 and these SDs (calculated), so this value largely reflects regression to the mean and does not by itself show systematic miscalibration. Consistent with a slope near 1, error was almost uncorrelated with predicted PMI (r ≈ 0.05, calculated; Figure 1C). The practical consequence remains: early samples were predicted too long and late samples too short.

3.5 Distribution of individual errors

The median absolute error was 5.1 hours. Sixteen of 33 samples had an absolute error of 5 hours or less, 24 of 10 hours or less, 29 of 12 hours or less, 30 of 15 hours or less, and 32 of 20 hours or less; the 80th percentile was 11.52 hours and the 95th 16.98 hours. The largest error was 30.6 hours, for the 58.6-hour sample (predicted 28.0 h). Other large errors were an 18.0-hour overestimate of a 17.3-hour sample and a 16.3-hour underestimate of a 41.0-hour sample. Sixteen predictions were higher than observed, 15 lower, and 2 equal after rounding. The SD of each sample's 20 out-of-fold predictions had a median of 2.3 hours (interquartile range 2.0–3.1; range 1.1–5.9; mean 2.63). Samples with large partition SDs were not necessarily those with large errors (Figure 2).

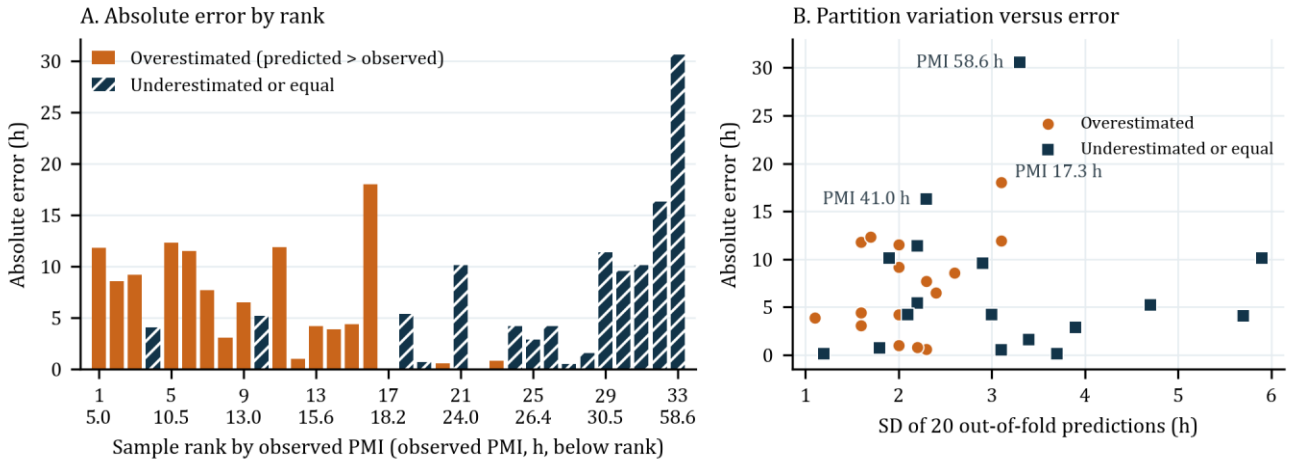


Figure 2. Sample-level error and prediction variation. (A) Absolute error for the 33 samples in rank order of observed PMI (1 = shortest); orange bars denote overestimation and hatched dark-blue bars underestimation or equal prediction. (B) Standard deviation of each sample's 20 out-of-fold predictions versus absolute error; the three largest errors are labeled by observed PMI.

3.6 Exploratory metabolite–PMI associations

Six of 41 metabolites remained associated with PMI after Benjamini–Hochberg correction (Table 1, Panel B; Figure 3): hypoxanthine (ρ 0.707, 95% CI 0.437–0.866; FDR 0.00017), choline (0.644; 0.00106), succinic acid (0.545; 0.01411), betaine (0.518; 0.02077), glutamic acid (0.474; 0.04206), and fumaric acid (0.467; 0.04206). Inosine monophosphate (IMP), creatine, pyroglutamic acid, and lysine were next in rank but did not meet the threshold.

Because metabolites are correlated and the analyses univariable, the ranking does not establish independent effects or a minimal predictive panel.

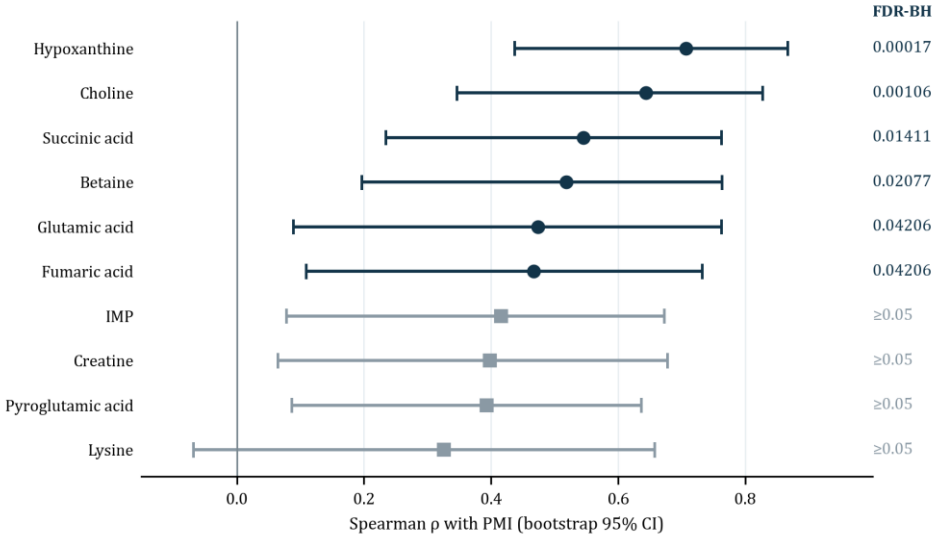


Figure 3. Ten leading exploratory associations between metabolites and PMI. Points are Spearman ρ with bootstrap 95% confidence intervals; dark blue circles denote FDR-BH below 0.05 and grey squares denote FDR-BH of 0.05 or higher.

3.7 Sensitivity and reproducibility

With untransformed concentrations, MAE rose from 7.04 to 7.30 hours and RMSE from 9.48 to 10.01 hours, R^2 fell from 0.233 to 0.145, and r from 0.485 to 0.388. Excluding the maximum-PMI sample gave an MAE of 6.55 hours and an RMSE of 8.06 hours, but R^2 was only 0.124 and r 0.358; the lower RMSE partly reflects the narrower outcome range. Intervals and a recalculated 32-sample benchmark are not available. Recalculation from the rounded out-of-fold table reproduced the primary MAE, RMSE, R^2, correlations, calibration slope, and error quantiles within rounding. The independent Python implementation gave an MAE of 6.84 hours, an RMSE of 9.34 hours, and an R^2 of 0.256 (Table 2), slightly better than the primary run, which indicates that estimates depend on fold assignment and implementation.

4. Discussion

4.1 Principal findings

In this 33-sample deposit, the aqueous-humor profile carries some PMI-related information but is not precise enough for casework. The ridge model improved RMSE by 13.4% over a mean-only benchmark, and six metabolites survived FDR control, which suggests a detectable PMI-related signal. The precision is another matter. R^2 was 0.233 with an interval extending below zero, and the RMSE interval contained the benchmark RMSE. No interval was estimated for the improvement. Out-of-fold RMSE remained 9.48 hours, with an 80th percentile of absolute error above 11 hours. Incremental information and decision-grade precision are separate standards, and this study meets only the first. Because metrics derive from averaged predictions, they are likely more favorable than the performance of one model trained once.

4.2 Calibration, compression, and the extremes

A calibration slope of 0.91 sits beside strongly compressed predictions (SD 5.85 h versus 10.99 h) because r is only 0.485. A regularized model trained on few samples with sparse extremes is expected to compress predictions toward the mean. The strong negative association between error and observed PMI is largely a mathematical consequence of imperfect prediction. The forensically relevant point is the pattern itself. Samples under 12 hours were overestimated by about 8 hours on average, an error comparable to the interval being estimated, while the 58.6-hour sample was underestimated by 30.6 hours. A central estimate near 20 hours can therefore mislead when the true PMI lies in a sparsely represented region. The 58.6-hour sample also carries leverage. Excluding it lowered R^2 to 0.124 and r to 0.358, so the modest performance is partly carried by one observation. The slightly better Python result (RMSE 9.34 h) shows that differences in implementation, fold generation, scaling, and λ parameterization together shift estimates by a few tenths of an hour. Residuals against predicted PMI, calibration curves in larger data, and individual prediction intervals are needed. Intervals for individual estimates will be wider than the bootstrap intervals for average performance reported here. Calibration deserves as much attention as discrimination, because poorly calibrated predictions can mislead decisions.[35] Small samples also give large error bars on cross-validated estimates. In neuroimaging examples they reached about ±10% with 100 samples,[37] so wide intervals with 33 samples are expected.

4.3 Biological interpretation and clinical confounders

Hypoxanthine, the strongest correlate, fits the established role of purine degradation in ocular-fluid thanatochemistry. Adenosine triphosphate depletion generates adenosine, inosine, and hypoxanthine, which can accumulate as clearance declines.[14,15,27] Choline, succinate, fumarate, betaine, and glutamate fit three overlapping domains: loss of energetic homeostasis, membrane disruption, and altered osmotic and amino acid balance. These processes are compatible with the pattern, which generates hypotheses for controlled time-course studies. It cannot establish a biochemical sequence. Because the association analysis was separate from the model, it also cannot show what drives the predictions.

The confounders that matter in practice were not recorded. Ante-mortem hypoxia and agonal duration, terminal illness and organ failure, drugs such as allopurinol[38] or metformin, resuscitation, hospital versus community death, refrigeration before sampling, and body temperature could shift hypoxanthine or succinate independently of PMI. They are the main reason a laboratory model may not transport. Hypoxanthine is also a well-known marker of tissue hypoxia in life. In hypoxemic pigs, hypoxanthine rose in plasma and cerebrospinal fluid before death.[39] Vitreous hypoxanthine has been proposed as an indicator of prolonged hypoxia as well as of PMI, although one series of 92 bodies found a higher correlation of vitreous potassium than of hypoxanthine with time since death.[40]

4.4 Relation to previous work

The results build on the human observations of Zelentsova et al., who showed gradual post-mortem change in ocular fluids.[27] They ask a prediction question within the aqueous-humor subset. Gradual group-level change translated into moderate out-of-fold prediction but not narrow case-level error. The closest precedent is the human aqueous-humor NMR work of Chighine et al.,[26] which should be the direct comparator in future validation. Interval-specific errors of 6.9 h (PMI below 24 h), 7.4 h (24 to 48 h), and 10.3 h (above 48 h) were reported for ovine vitreous humor.[24] The human RMSE of 9.48 hours is numerically similar. Species, matrix, temperature control, and validation design differ, though, so the similarity supports plausibility but not convergence. Established methods supply the context needed to judge whether an error of about 9 hours is competitive.[1,9] For vitreous potassium, 95% limits of about ±34 h up to 120 h after death were reported in one series,[9] and a later re-evaluation found accuracy of about ±23 to ±26 h, while a narrower recommended accuracy was not met in a validation sample of 492 cases.[41] For body cooling, the actual PMI fell within the 95% interval of the rectal nomogram in 72.8% of cases in a series of 100 bodies with a mean PMI of 4.5 h.[2] Body cooling remains the reference approach in practice, and chemical methods play a marginal role in routine work.[42] These figures are not directly comparable with an RMSE. If errors were approximately normal, a 95% limit corresponding to the RMSE of 9.48 hours would be about ±18.6 hours. A hypoxanthine-only or small-panel model within the same nested loop would show whether 41 predictors add anything to the strongest single marker. This was not done. The single-marker correlation ($\rho^2 \approx 0.50$) is not comparable with out-of-fold R^2 because the marker was chosen from 41 using all samples.

Other recent human studies show the range of reported errors. In 38 autopsy cases with PMI of 39.5 to 285 h, a multivariate vitreous model of potassium, ammonium, and albumin gave an RMSE of ±15.5 h, against ±22.6 h for potassium alone.[12] ATR-FTIR spectroscopy of vitreous humor gave a cross-validated RMSE of 15.8 h (R^2 0.53) in 20 cases with PMI of 24.8 to 97.6 h.[19] Blood metabolomics from routine toxicology screening (n = 4,876; PMI 1 to 67 days) gave a mean absolute error of 1.45 days in unseen test cases.[20] The matrices, PMI ranges, and sample sizes differ, so these errors cannot be set directly against the 9.48-h RMSE reported here. They do show that small human studies report errors of this order, and that the large blood cohort needed to narrow them. In the closest human aqueous-humor study, models were less accurate than the ovine model, and taurine and hypoxanthine were PMI-specific metabolites in both species.[26] This fits hypoxanthine leading the associations in the present data. Zelentsova et al. also found that metabolomic change in anatomically isolated ocular fluids is slower and smoother than in serum,[27] which supports the choice of ocular matrices for a time-sensitive profile.

4.5 Implications for model development and forensic use

With 33 samples and 41 predictors, ridge regression stabilizes estimation but does not replace sample size. A continuous-outcome sample-size calculation would require far more observations.[43] A future study should prespecify the target use and acceptable error and plan sample size for the precision of calibration and residual variance. Any predictor reduction should be decided on biological and analytical grounds in a separate development phase, and the pipeline should be locked before external outcomes are examined. A molecular PMI estimate should support, not replace, expert synthesis. A single point prediction without an individual uncertainty interval is insufficient for high-stakes interpretation.

4.6 Strengths and limitations

The strengths are public sample-level data that allowed reconstruction and checking, complete data without imputation, and regularization suited to correlated predictors. Nested tuning, repetition, a fold-specific benchmark, FDR control, sensitivity analyses, and sample-level reporting add further support.

The limitations are considerable. The sample is small and comes from a single source with an unknown cohort. The deposit lacks covariates and details of how PMI was established. Samples may not be independent. Validation is internal only, metrics derive from averaged predictions, and the bootstrap intervals omit model-development uncertainty. The benchmark difference has no interval, per-repetition performance is not reported, and no single-marker comparator was evaluated. The associations are cross-sectional. Several further analyses (grouped folds, per-repetition distributions, hypoxanthine-only and small-panel models, a permutation test, benchmark recalculation for n = 32) require the original code and are not reported here.

4.7 Future research

Progress requires a staged program. It begins with analytical validation of the NMR measurements and moves to biological validation across temperature, cause of death, and sampling conditions. Model development with prespecified predictors follows. The last stage is locked multicenter external validation against established ocular chemistry and contextual variables, reporting individual prediction intervals across the intended PMI range. Paired aqueous and vitreous sampling from the same decedents would show whether each matrix adds information to the other or whether they are redundant. Data, code, preprocessing specifications, and coefficients should be deposited so that independent laboratories can reproduce the results.

5. Conclusion

In 33 human aqueous-humor samples, a ridge model reduced RMSE by 13.4% relative to a mean-only benchmark (MAE 7.04 h, RMSE 9.48 h under repeated nested internal validation). Six metabolites showed FDR-significant associations with PMI. Explained variance was modest, with R^2 0.233 and an interval extending below zero. The improvement has no interval. Predictions were

compressed toward the center, with early samples overestimated and the longest PMI underestimated. These findings suggest a detectable PMI-related signal and support the feasibility of aqueous-humor metabolomic PMI estimation. They do not yet give stable individual estimates for medicolegal use. A necessary next step is a locked, multicenter human validation that compares the metabolomic model with established ocular chemistry and contextual variables and reports individual prediction intervals.

Declarations

Use of generative artificial intelligence. None. The authors declare that no generative artificial intelligence (AI) tools were used in the preparation, drafting, analysis, or editing of this manuscript.

Author contributions. BS: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft. SM: Investigation, Validation, Visualization, Writing – review & editing. KC: Conceptualization, Supervision, Project administration, Writing – review & editing. JW: Resources, Validation, Writing – review & editing. All authors approved final version of manuscript.

Funding. None. The authors received no external funding for this study.

Conflict of interest. The authors declare no conflict of interest.

Ethical approval. Ethical approval was granted by the CMHC Research Center Ethics Committee Indonesia (No. 2025/013) for the secondary analysis of a de-identified public dataset.

Consent to participate. Not applicable. No new participant contact or specimen collection occurred; consent and collection procedures for the original specimens were those of the source study.

Data availability. The source data are publicly available in Animal Metabolite Database Experiment 242 (https://amdb.online/amdb/experiments/242/).[28] Derived sample-level predictions, association results, and the reproducibility specification are available from corresponding author on reasonable request.

Reporting guideline. TRIPOD+AI, used as a transparency framework.

Acknowledgments. None.

References

1. Henssge C, Madea B. Estimation of the time since death in the early post-mortem period. Forensic Sci Int. 2004;144(2-3):167-175. doi:10.1016/j.forsciint.2004.04.051
2. Laplace K, Baccino E, Peyron PA. Estimation of the time since death based on body cooling: a comparative study of four temperature-based methods. Int J Legal Med. 2021;135(6):2479-2487. doi:10.1007/s00414-021-02635-7
3. Wilk LS, Edelman GJ, Roos M, et al. Individualised and non-contact post-mortem interval determination of human bodies using visible and thermal 3D imaging. Nat Commun. 2021;12(1):5997. doi:10.1038/s41467-021-26318-4
4. Nioi M, Napoli PE, Demontis R, et al. The influence of eyelid position and environmental conditions on corneal changes in the early postmortem interval: a prospective, multicentric OCT study. Diagnostics (Basel). 2022;12(9):2169. doi:10.3390/diagnostics12092169
5. Klontzas ME, Leventis D, Spanakis K, et al. Post-mortem CT radiomics for the prediction of time since death. Eur Radiol. 2023;33(11):8387-8395. doi:10.1007/s00330-023-09746-2
6. De-Giorgio F, Cusumano D, Vellini L, et al. Estimation liver radiomics from postmortem CT: development of interpretable models for postmortem interval estimation. Phys Med. 2025;138:105186. doi:10.1016/j.ejmp.2025.105186
7. Hu L, Xing Y, Jiang P, et al. Predicting the postmortem interval using human intestinal microbiome data and random forest algorithm. Sci Justice. 2021;61(5):516-527. doi:10.1016/j.scijus.2021.06.006
8. Sturner WQ. The vitreous humour: postmortem potassium changes. Lancet. 1963;281(7285):807-808. doi:10.1016/S0140-6736(63)91509-5
9. Madea B, Henssge C, Hönig W, Gerbracht A. References for determining the time of death by potassium in vitreous humor. Forensic Sci Int. 1989;40(3):231-243. doi:10.1016/0379-0738(89)90181-3
10. Kurup SS, Bharathi M, Narayan G, et al. Estimation of time since death from potassium levels in vitreous humor in cases of unnatural death: a facility-based cross-sectional study. Cureus. 2023;15(5):e39572. doi:10.7759/cureus.39572
11. Kavin G, Chaudhari VA, Patra AP, et al. Correlation of vitreous and serum biochemical markers with postmortem interval and cold storage duration: a prospective observational study. J Forensic Leg Med. 2025;114:102932. doi:10.1016/j.jflm.2025.102932
12. Focardi M, Defraia B, Bianchi I, et al. A step forward in post-mortem interval estimation: multivariate analysis of ammonium, albumin, and potassium levels in vitreous humor. Diagnostics (Basel). 2026;16(13):1970. doi:10.3390/diagnostics16131970
13. Cecchi R, Camatti J, Schirripa ML, et al. Postmortem biochemistry of GFAP, NSE and S100B in cerebrospinal fluid and in vitreous humor for estimation of postmortem interval: a pilot study. Forensic Sci Med Pathol. 2025;21(2):589-598. doi:10.1007/s12024-024-00874-9
14. Rognum TO, Hauge S, Øyasæter S, Saugstad OD. A new biochemical method for estimation of postmortem time. Forensic Sci Int. 1991;51(1):139-146. doi:10.1016/0379-0738(91)90214-4
15. Rognum TO, Holmen S, Musse MA, et al. Estimation of time since death by vitreous humor hypoxanthine, potassium, and ambient temperature. Forensic Sci Int. 2016;262:160-165. doi:10.1016/j.forsciint.2016.03.001
16. Snytnikova OA, Khlichkina AA, Yanshole LV, et al. Metabolomics of the human aqueous humor. Metabolomics. 2017;13(1):5. doi:10.1007/s11306-016-1144-0
17. Locci E, Stocchero M, Gottardo R, et al. Comparative use of aqueous humour 1H NMR metabolomics and potassium concentration for PMI estimation in an animal model. Int J Legal Med. 2021;135(3):845-852. doi:10.1007/s00414-020-02468-w
18. Notarstefano V, Santoni C, Montanari E, et al. A new approach to assess post-mortem interval: a machine learning-assisted label-free ATR-FTIR analysis of human vitreous humor. Spectrochim Acta A Mol Biomol Spectrosc. 2025;327:125326. doi:10.1016/j.saa.2024.125326
19. Țurlea IR, Curca GC, Mernea M, et al. FTIR spectroscopy of vitreous humor for postmortem interval estimation: a multivariate regression approach. Int J Mol Sci. 2026;27(8):3468. doi:10.3390/ijms27083468
20. Magnusson R, Söderberg C, Ward LJ, et al. The human metabolome and machine learning improves predictions of the post-mortem interval. Nat Commun. 2026;17(1):1504. doi:10.1038/s41467-026-69158-w
21. Steuer AE, Wartmann Y, Schellenberg R, et al. Postmortem metabolomics: influence of time since death on the level of endogenous compounds in human femoral blood. Necessary to be considered in metabolome study planning? Metabolomics. 2024;20(3):51. doi:10.1007/s11306-024-02117-y
22. Lu XJ, Li J, Wei X, et al. A novel method for determining postmortem interval based on the metabolomics of multiple organs combined with ensemble learning techniques. Int J Legal Med. 2023;137(1):237-249. doi:10.1007/s00414-022-02844-8
23. Fan W, Dai X, Ye Y, et al. Estimation of postmortem interval under different ambient temperatures based on multi-organ metabolomics and machine learning algorithm. Int J Legal Med. 2025;139(5):2561-2575. doi:10.1007/s00414-025-03523-0
24. Locci E, Stocchero M, Gottardo R, et al. PMI estimation through metabolomics and potassium analysis on animal vitreous humour. Int J Legal Med. 2023;137(3):887-895. doi:10.1007/s00414-023-02975-6
25. Locci E, Stocchero M, Noto A, et al. A 1H NMR metabolomic approach for the estimation of the time since death using aqueous humour: an animal model. Metabolomics. 2019;15(5):76. doi:10.1007/s11306-019-1533-2

26. Chighine A, Stocchero M, De-Giorgio F, et al. Translating metabolomic evidence gathered from an animal model to a real human scenario: the post-mortem interval issue. Metabolomics. 2025;21(5):125. doi:10.1007/s11306-025-02321-4

27. Zelentsova EA, Yanshole LV, Melnikov AD, et al. Post-mortem changes in metabolomic profiles of human serum, aqueous humor and vitreous humor. Metabolomics. 2020;16(7):80. doi:10.1007/s11306-020-01700-3

28. Animal Metabolite Database. Experiment 242: Human aqueous humor post-mortem [Internet]. Created 2022 Dec 16; last modified 2023 Sep 7]. Available from: https://amdb.online/amdb/experiments/242/

29. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55-67. doi:10.1080/00401706.1970.10488634

30. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine-learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378

31. Moons KGM, Damen JAA, Kaul T, et al. PROBAST+AI: an updated quality, risk-of-bias, and applicability assessment tool for prediction models using regression or artificial-intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505

32. Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. GigaScience. 2017;6(5):gix019. doi:10.1093/gigascience/gix019

33. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1-22. doi:10.18637/jss.v033.i01

34. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS One. 2019;14(11):e0224365. doi:10.1371/journal.pone.0224365

35. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. doi:10.1186/s12916-019-1466-7

36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

37. Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage. 2018;180:68-77. doi:10.1016/j.neuroimage.2017.06.061

38. Pacher P, Nivorozhkin A, Szabó C. Therapeutic effects of xanthine oxidase inhibitors: renaissance half a century after the discovery of allopurinol. Pharmacol Rev. 2006;58(1):87-114. doi:10.1124/pr.58.1.6

39. Poulsen JP, Øyasæter S, Sanderud J, et al. Hypoxanthine, xanthine, and uric acid concentrations in the cerebrospinal fluid, plasma, and urine of hypoxemic pigs. Pediatr Res. 1990;28(5):477-481. doi:10.1203/00006450-199011000-00012

40. Madea B, Käferstein H, Hermann N, Sticht G. Hypoxanthine in vitreous humor and cerebrospinal fluid—a marker of postmortem interval and prolonged (vital) hypoxia? Remarks also on hypoxanthine in SIDS. Forensic Sci Int. 1994;65(1):19-31. doi:10.1016/0379-0738(94)90296-8

41. Madea B, Rödig A. Time of death dependent criteria in vitreous humor: accuracy of estimating the time since death. Forensic Sci Int. 2006;164(2-3):87-92. doi:10.1016/j.forsciint.2005.12.002

42. Madea B. Methods for determining time of death. Forensic Sci Med Pathol. 2016;12(4):451-485. doi:10.1007/s12024-016-9776-y

43. Riley RD, Snell KIE, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: part I - continuous outcomes. Stat Med. 2019;38(7):1262-1275. doi:10.1002/sim.7993