Research Square

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

Estimating balloon-centre insertion depth for fluoroscopy-free REBOA with minimal body-surface measurements: development and external evaluation of a prediction model using CT angiography

Jianhan Yin
Nankai University
Feiyang Zhong
Nankai University
Yangyang Ge
The First Medical Center of Chinese PLA General Hospital
Long Cao
The 983rd Hospital of Joint Logistics Support Forces of the PLA
Haibo Wang
Hainan Hospital of Chinese PLA General Hospital
Shaofan Wang
Nankai University
Tianfeng Ma
Eastern Theater Command General Hospital of PLA
Dan Rong
The First Medical Center of Chinese PLA General Hospital
Wei Guo
guoweiplagh@sina.com

Nankai University

Research Article

Keywords: resuscitative endovascular balloon occlusion of the aorta, computed tomography, anatomical landmarks, prediction model, intravascular length, external evaluation

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11812826/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

In battlefield and austere emergency settings, resuscitative endovascular balloon occlusion of the aorta (REBOA) can temporarily control non-compressible haemorrhage below the diaphragm. Without fluoroscopy, the depth to the centre of the target aortic zone must be estimated. The gain in accuracy from additional body-surface measurements is uncertain. To evaluate midpoint error and simulated positioning with the smallest of the fixed input sets compared for REBOA.

Methods

In this retrospective study, of 10,453 patients screened across centres after CT angiography (CTA), 547 of 8,722 screened for development and 94 of 1,731 screened for external evaluation were included. Forward selection among 14 measurements and sex identified body thickness, suprasternal notch-to-pubic symphysis distance and body width. Ridge regression estimated four total distances to zone midpoints and six component distances. We compared continuous errors across models and also evaluated calibration and reference-window hits for the four total distances. Comparators included constants, larger input sets and other models. Intervals used patient-level bootstrapping for continuous measures and exact binomial limits for hit rates. Paired t-tests and exact McNemar tests were used in the relevant comparisons, with Holm adjustment.

Results

With three body-surface inputs, mean absolute errors in the external cohort were 15.8 and 16.3 mm on the right and left in zone I and 11.6 and 12.2 mm in zone . The three-input model and constants hit all zone I windows on both sides (94/94); the three-input model hit 88/94 right and 90/94 left zone windows.

Conclusions

Among the fixed input sets compared, the three body-surface measurements balanced measurement burden and midpoint error. They estimated the zone midpoint more accurately than constants even in zone I, where reference-window hit rates were identical. These CT findings may inform depth estimation for fluoroscopy-free REBOA in battlefield and austere emergency settings.

Background

Non-compressible torso haemorrhage (NCTH) is a life-threatening combat injury. In a US review of potentially survivable battlefield deaths before arrival at a treatment facility, the torso was the most

common site of lethal bleeding [1]. In the UK Joint Theatre Trauma Registry, mortality was 85.5% among 296 casualties with NCTH [2]. For bleeding below the diaphragm, resuscitative endovascular balloon occlusion of the aorta (REBOA) can provide temporary haemorrhage control [3]. US special operations medical personnel and a far-forward coalition unit have reported using it [4, 5]. The UK-REBOA randomised trial, however, did not show a survival benefit. The investigators suggested that delayed or unsuccessful definitive haemostasis might explain the early mortality signal [6]. The 2025 Eastern Association for the Surgery of Trauma guideline conditionally recommended against REBOA in haemodynamically unstable trauma patients with suspected subdiaphragmatic bleeding or pelvic fractures [7].

When REBOA is used without fluoroscopy, the operator must estimate the depth needed to reach the intended aortic zone. The usual targets are zone I, between the left subclavian artery and the coeliac trunk, and zone □, between the lowest renal artery and the aortic bifurcation. Zone II, between the coeliac trunk and the lowest renal artery, is contraindicated. Insufficient advancement may leave the bleeding source perfused; excessive advancement may occlude branches that should remain patent. The margin for error is particularly narrow in zone III. In battlefield and emergency care, depth must often be estimated from few measurements, without relying on the casualty to answer questions.

Previous studies have used fixed distances, surface or skeletal landmarks and morphometric estimates to locate aortic zones [8–14]. The location of a zone is not itself an insertion depth: one segmentation study mapped zones onto the body surface without estimating insertion distance [15]. Studies that did report length also used different origins. Chien et al. included an assumed oblique skin-to-artery segment and used the umbilical level and a ratio of surface distances to locate zone III [16]. By contrast, van de Voort et al. generally measured from within the femoral artery and reported lengths by age, sex and access side [17].

Using a common CT reference, we asked how accurately three body-surface measurements, the smallest of three fixed input sets, could estimate total distance to the zone midpoint and its intravascular and extravascular components. In an independent external cohort, we evaluated midpoint error and reference-window hits separately by side and zone. Comparison with development-cohort constants tested what individual measurements added; comparisons with models using the same or more inputs examined the contributions of model complexity and additional measurements.

Methods

Study design and participants

This retrospective study included patients who underwent contrast-enhanced aortic CT angiography (CTA) at our institution and external centres. The CT images and body-surface measurements used in the analysis came from routine clinical care. Reporting was guided by the TRIPOD + AI statement [18], and risk of bias and applicability were considered with reference to the PROBAST + AI framework [19].

The Ethics Review Committee of the Chinese People's Liberation Army General Hospital approved the study (2026 – 784) across all participating centres and the eligible age range and waived informed consent. The study followed the Declaration of Helsinki. Eligible outpatients were aged 16 to 60 years and had CTA coverage from the suprasternal notch to the pubic symphysis. Exclusions included congenital or acquired aortic disease, a CTA Kauppila score of 3 or higher, immune-mediated or connective tissue disease, restricted iliofemoral access, intra-abdominal or retroperitoneal lesions, previous acute aortic syndrome, aortic anatomical variants and pregnancy (Supplementary Methods 1.1 to 1.3).

Data sources and cohort composition

The development centre screened 8,722 consecutive patients who underwent contrast-enhanced aortic CTA from January 2015 to July 2025. Of 558 patients who met the anatomical and disease-related criteria, 11 (2.0%) were excluded because their autoencoder reconstruction error exceeded the 98th percentile, leaving 547 patients for analysis. All fields required for analysis were complete, and no imputation was performed.

The independent external cohort came from three provincial medical centres in northern, eastern and southern China. A total of 1,731 patients who underwent CTA were screened across two periods: January to December 2014 and August 2025 to March 2026. Of these, 94 met the same inclusion and exclusion criteria; all had paired right and left measurements and were included in the primary analysis. The cohort and analysis flow are shown in Supplementary Fig. S1.

Measurements and reference targets

The 15 candidate predictors comprised sex and 14 distances or body dimensions. Starting with an intercept-only model, cumulative forward selection added one variable at a time in the development cohort. At each step, three-fold cross-validation evaluated the largest mean absolute error (MAE) across the ten outputs under ridge penalties of 1, 10 and 100. The candidate addition with the lowest value entered the model; penalties were then selected by output. The selection order and input-count versus error curve informed the three-measurement FAST configuration (Fig. 1; Supplementary Methods 2.1 and Supplementary Table S1). After exploration, the input counts (three, six and fifteen) were fixed for comparison and were not reselected in the inner or outer folds. Age, hypothetical skin puncture-point measurements, vertical skin-to-artery depth and vascular endpoints were not candidate inputs.

The first three predictors to enter were body thickness (BT), suprasternal notch-to-pubic symphysis distance (SNP) and body width (BW). BT and BW were the anteroposterior and transverse diameters of the outer body contour on supine axial CT images at the umbilical level.

PRECISE added three distances to FAST: xiphoid to left anterior superior iliac spine, suprasternal notch to left anterior superior iliac spine, and umbilicus to pubic symphysis. FULL used sex and all 14

distances or body dimensions. Input burden counted distinct raw information items, not procedure time.

For each access side, the intravascular coordinate started at the centre of the femoral artery at the level of the midpoint of the femoral head, above the arterial bifurcation. The perpendicular distance from this origin to its projection on the skin gave the CT-measured depth. The skin entry point need not coincide with that projection: the idealised extravascular length (EVL) from skin entry to the arterial origin was calculated offline as the measured depth multiplied by the square root of two.

Zone I was bounded by the coeliac trunk and left subclavian artery levels on CT, and zone □ by the lowest renal artery and aortic bifurcation. Intravascular length (IVL) was the distance from the femoral origin to the midpoint between the two boundaries on that coordinate. Total reference distance (T) from skin entry to the zone midpoint was the sum of EVL and IVL. The four T, four IVL and two EVL outputs were fitted directly and separately.

A reference-window hit (successful simulated blind-placement positioning) required predicted T to fall within the closed interval formed by adding EVL to each intravascular boundary distance and insetting each end by 20 mm. This corresponds to an assumed axial length of 40 mm and leaves the midpoint unchanged. Empty windows counted as misses but retained their midpoint labels for fitting; singleton windows required exact equality.

Prediction models and comparison methods

FAST, PRECISE and FULL used classical ridge regression [20] with a fixed penalty of 10 and an unpenalised intercept. Predictors were standardised using training-set means and standard deviations to put them on a common scale for penalisation [21]. Each fit minimised the residual sum of squares plus the penalty times the squared norm of the slope vector. Because T is the sum of the component labels, the identical design matrix and penalty make the direct T prediction equal to the sum of the two fitted components (Supplementary Methods 6 and Supplementary Note 2).

The development-cohort mean and median of each target served as constants. Ten model families used FAST inputs: ridge regression, quadratic polynomial ridge regression, elastic net, support vector regression (SVR) with a radial basis function (RBF), random forest, extremely randomised trees, histogram gradient boosting, CatBoost, XGBoost and LightGBM. Nested cross-validation shared five patient-level outer folds and three inner folds across families. We fitted preprocessing within each training fold and selected among three configurations per family by pooled MAE across four T targets, weighting patients and targets equally. Out-of-fold evaluation covered four *T* targets; final development-cohort tuning and refitting generated all ten external outputs. Among ridge penalties of 0.1, 10 and 100, 10 was selected in every outer fold and in final tuning. The final external ridge fit reproduced FAST.

Variable selection was not repeated within the folds. Development-only analyses examined internal learning curves, selection by target-specific MAE or reference-window hits, and a finite squared penalty for predictions outside the window (Supplementary Methods 3 and 7).

The extended comparison evaluated ordinary least squares (OLS), ridge regression, radial-basis-function SVR, CatBoost, spline additive regression, Gaussian process regression, explainable boosting machine and TabPFN v2 with the same three inputs and outer folds. OLS involved no hyperparameter search; TabPFN v2 used fixed pretrained weights. Each other family compared nine configurations in development. All models were then fitted to 547 patients and evaluated on ten outputs in the same 94 external patients. The development correlation matrix, variance inflation factors, condition number of the standardised design matrix and 1000 patient bootstrap samples characterised OLS and fixed-penalty ridge coefficient and prediction stability (Supplementary Methods 3.2 and 3.5).

For ridge regression and SVR from the ten-family comparison and all eight extended models, 26 hypothetical input shifts tested changes in predictions and errors for the four T targets and in zone reference-window hits. Offline single-threaded central processing unit (CPU) inference time was measured on the same machine (Supplementary Methods 3.5 and Supplementary Table S9).

For the ten-family learning curves, development sample sizes were 60, 120, 200, 300, 400, 475 and 547 patients. For the first six sizes, 20 shared nested subsamples were drawn without replacement, with preprocessing refitted and three-fold tuning repeated in each subset. The full cohort contributed a single fit. All sizes were evaluated in the same 94 external patients. A non-negative power-law saturation curve was fitted from 120 to 547 patients; the 60-patient point assessed low-sample-size behaviour. An exponential fit tested sensitivity to curve form. We recorded crossings in either direction and absence of crossing relative to ridge regression over hypothetical training sizes of 547 to 2735 and up to 5470 (Supplementary Methods 3.3).

Outcomes and statistical methods

Signed error was the prediction minus its reference. We calculated MAE, root mean squared error (RMSE), bias and 90th-percentile absolute error separately by output. Reference-window hit proportions were calculated only for the four T targets, separately by side and zone. Absolute error is a scoring loss for point predictions [22]. The equally weighted mean MAE across the four T targets was the tuning criterion and a secondary summary.

Nonparametric bootstrap resampling [23] used complete patient records to preserve pairing between models and sides. Pointwise 95% confidence intervals (CIs) used bootstrap percentiles from 5000 resamples for FAST and paired MAE analyses and 2000 for model-comparison cells. We compared FAST with the mean, median, PRECISE and FULL across six IVL/EVL outputs using two-sided paired t-tests with Holm correction across 24 tests; the sensitivity analysis covered 40 tests across all ten outputs. Record-level sign-flip tests used 100,000 repetitions. Pointwise CIs were reported alongside adjusted tests. Hit proportions used two-sided Clopper-Pearson intervals [24]. The significance level was 0.05 (Supplementary Methods 4).

For each model, absolute errors were averaged first across the four T targets within each patient, then across external patients. Within each comparison, differences were candidate minus ridge regression. We used 10000 shared patient bootstrap resamples for pointwise percentile 95% CIs. Simultaneous intervals based on the maximum standardised absolute deviation were constructed separately for the nine comparators in the ten-family comparison and the seven extended comparators. Learning-curve intervals held training repetitions fixed: each external-patient resample refitted both curve forms and located their crossings. Interquartile ranges across training subsamples were described separately (Supplementary Methods 3.3 and 3.4).

Calibration regressed observed T on predicted T without external recalibration, using 3000 whole-patient bootstrap resamples. Sex comparisons were exploratory. For paired hits, exact two-sided McNemar tests [25] conditioned on the number of discordant pairs. Holm adjustment covered the family of 52 tests across 13 comparators and four targets [26]. We derived pointwise and Bonferroni simultaneous bounds for hit differences from Clopper-Pearson bounds for the two discordant proportions, separately from the McNemar tests.

To examine the contributions of IVL and EVL, we combined the two predicted components, substituted the development mean for either component in turn, or used both means. We propagated hypothetical Gaussian errors, common shifts and bounded-box input errors through the fixed linear model and observed residuals. For geometric sensitivity, predictions were held fixed while each boundary inset was replaced by half an assumed axial length ranging from 20 to 100 mm (Supplementary Methods 6 and 8). Analyses used Python 3.12.14 with NumPy 2.5.4, pandas 3.1.0, SciPy 1.18.1 and scikit-learn 1.9.1.

Offline prediction software implemented the frozen FAST equations and returned all ten distances from BT, SNP and BW. Synthetic inputs checked its numerical agreement with the equations (Supplementary Methods 12 and 13; Supplementary Fig. S7).

Use of generative AI

GPT-6 Astra via DeepSeek Harness assisted with proofreading and grammar checking. The authors are responsible for verifying the text and references and for the final manuscript.

Results

Cohorts and reference targets

The development cohort comprised 401 male and 146 female patients, and the external cohort comprised 72 male and 22 female patients. The three body-surface measurements and ten CT reference distances are summarised in Table 1. In the 94 external patients, the median reference-window half-width was 98.9 mm in zone I and 29.9 mm in zone III when each side was assessed separately.

Table 1
Patient characteristics, the three measurements and reference distances

Characteristic	Summary measure	Development set	External set	SMD (external − development)
Cohort size	n	547	94	Not applicable
Sex, male / female	n	401 / 146	72 / 22	0.076
Body thickness	Mean (SD); median $[Q_1,Q_3]$, mm	218.60 (35.68); 217.12 [195.88, 239.56]	219.49 (29.46); 218.06 [198.79, 242.72]	0.027
Suprasternal notch-to-pubic symphysis distance	Mean (SD); median $[Q_1,Q_3]$, mm	553.58 (27.74); 553.81 [535.17, 569.54]	548.02 (28.13); 547.93 [532.42, 563.14]	−0.199
Body width	Mean (SD); median $[Q_1,Q_3]$, mm	319.20 (35.19); 317.07 [296.61, 340.62]	317.86 (29.65); 318.40 [300.37, 335.62]	−0.041
T, right zone I	Mean (SD); median $[Q_1,Q_3]$, mm	500.44 (32.43); 496.88 [479.18, 518.70]	495.48 (29.93); 496.62 [479.78, 517.91]	−0.159
T, right zone	Mean (SD); median $[Q_1,Q_3]$, mm	294.38 (24.37); 291.34 [279.03, 308.22]	292.92 (21.35); 291.37 [277.45, 306.14]	−0.064
T, left zone I	Mean (SD); median $[Q_1,Q_3]$, mm	495.90 (32.29); 493.35 [473.57, 515.81]	491.28 (29.58); 492.44 [476.84, 509.95]	−0.149

Continuous variables are in mm and are presented as mean (SD) and median [Q1,Q3]; sex is presented as counts. The three measurements are body thickness, suprasternal notch-to-pubic symphysis distance and body width; body thickness and body width are the anterior-posterior and transverse diameters at the umbilical level on supine CT images. Sex is reported here as a cohort characteristic; it was an input to FULL but not to FAST or PRECISE. T denotes total CT reference distance, IVL intravascular reference length and EVL idealised extravascular reference length. Standardised mean differences compare the external and development cohorts using an equally weighted mean of the two group variances and are descriptive measures.

Abbreviations: SD, standard deviation; Q1/Q3, first/third quartile; SMD, standardised mean difference.

Characteristic	Summary measure	Development set	External set	SMD (external − development)
T, left zone □	Mean (SD); median $[Q_1,Q_3]$, mm	289.84 (24.08); 287.96 [273.98, 303.17]	288.73 (21.39); 288.86 [274.77, 299.67]	−0.049
IVL, right zone I	Mean (SD); median $[Q_1,Q_3]$, mm	456.56 (25.71); 455.45 [438.57, 471.25]	454.51 (24.61); 455.24 [439.50, 467.38]	−0.081
IVL, right zone □	Mean (SD); median $[Q_1,Q_3]$, mm	250.50 (17.20); 249.09 [239.14, 259.58]	251.96 (15.93); 250.32 [241.54, 260.34]	0.088
IVL, left zone I	Mean (SD); median $[Q_1,Q_3]$, mm	452.00 (25.75); 450.04 [433.47, 467.90]	450.21 (24.62); 452.34 [434.69, 461.78]	−0.071
IVL, left zone □	Mean (SD); median $[Q_1,Q_3]$, mm	245.93 (17.04); 245.60 [235.00, 254.50]	247.66 (16.20); 246.32 [235.95, 255.09]	0.104
Idealised EVL, right	Mean (SD); median $[Q_1,Q_3]$, mm	43.88 (16.49); 40.96 [32.49, 52.61]	40.97 (12.74); 40.33 [32.50, 47.10]	−0.198
Idealised EVL, left	Mean (SD); median $[Q_1,Q_3]$, mm	43.91 (16.24); 41.86 [33.08, 50.87]	41.07 (13.36); 38.78 [31.87, 48.07]	−0.191

Continuous variables are in mm and are presented as mean (SD) and median [Q1,Q3]; sex is presented as counts. The three measurements are body thickness, suprasternal notch-to-pubic symphysis distance and body width; body thickness and body width are the anterior-posterior and transverse diameters at the umbilical level on supine CT images. Sex is reported here as a cohort characteristic; it was an input to FULL but not to FAST or PRECISE. T denotes total CT reference distance, IVL intravascular reference length and EVL idealised extravascular reference length. Standardised mean differences compare the external and development cohorts using an equally weighted mean of the two group variances and are descriptive measures.

Abbreviations: SD, standard deviation; Q1/Q3, first/third quartile; SMD, standardised mean difference.

Midpoint error and simulated positioning

All FAST estimates fell within the zone I reference window on each side (94/94; 100%, 95% CI 96.2 to 100), although MAE was 15.8 mm (95% CI 13.4 to 18.3) on the right and 16.3 mm (14.1 to 18.6) on the left. In zone III, MAE was lower at 11.6 mm (9.7 to 13.8) and 12.2 mm (10.3 to 14.3), respectively. The 90th-percentile absolute errors in zone III were 24.3 mm on the right and 25.2 mm on the left. Of 94 estimates per side, 88 right-sided estimates (93.6%, 95% CI 86.6 to 97.6) and 90 left-sided estimates (95.7%, 89.5 to 98.8) fell within the narrower zone III windows (Fig. 2, Table 2).

Table 2
Estimation accuracy of the three-measurement model in the external cohort

Direct output	n	MAE [95% CI], mm	RMSE [95% CI], mm	Bias [95% CI], mm	Absolute error P_{90}, mm	Reference-window hits/ n	Hit rate [95% CI], %
T, right zone I	94	15.78 [13.40, 18.29]	19.82 [16.92, 22.63]	2.78 [− 1.31, 6.64]	32.02	94/94	100.00 [96.15, 100.00]
T, right zone	94	11.62 [9.66, 13.77]	15.35 [12.85, 17.83]	0.44 [− 2.74, 3.37]	24.25	88/94	93.62 [86.62, 97.62]
T, left zone I	94	16.33 [14.11, 18.59]	19.57 [17.12, 22.04]	2.37 [− 1.72, 6.21]	29.75	94/94	100.00 [96.15, 100.00]
T, left zone III	94	12.24 [10.29, 14.34]	15.69 [13.34, 18.02]	0.03 [− 3.26, 3.07]	25.20	90/94	95.74 [89.46, 98.83]
IVL, right zone I	94	16.30 [14.02, 18.66]	19.90 [17.29, 22.55]	−0.47 [− 4.64, 3.45]	31.60	Not applicable	Not applicable
IVL, right zone III	94	11.23 [9.37, 13.32]	14.97 [12.32, 17.59]	−2.81 [− 5.80, 0.02]	22.67	Not applicable	Not applicable
IVL, left zone I	94	16.66 [14.41, 19.02]	20.09 [17.60, 22.70]	−0.70 [− 4.88, 3.25]	31.03	Not applicable	Not applicable
IVL, left zone III	94	11.67 [9.67, 13.82]	15.47 [12.81, 18.12]	−3.04 [− 6.14, − 0.15]	23.88	Not applicable	Not applicable
Idealised EVL, right	94	6.88 [5.82, 8.07]	8.80 [7.46, 10.19]	3.25 [1.54, 4.85]	13.47	Not applicable	Not applicable
Idealised EVL, left	94	7.15 [5.90,	9.59 [7.93,	3.07 [1.21,	14.90	Not applicable	Not applicable

All estimates are from the same 94 external patients. Bias is the prediction minus the reference value. Intervals for MAE, RMSE and bias are 95% percentile intervals from 5000 bootstrap resamples of complete records; 90th-percentile absolute errors are point estimates without intervals. Reference-window hit proportions and their Clopper–Pearson intervals are shown for the four total-distance targets, by side and zone. Estimates are displayed to two decimal places, with full precision retained in the source tables.

Abbreviations: MAE, mean absolute error; RMSE, root mean squared error; CI, confidence interval; P90, the 90th percentile of absolute error (90th percentile); T, total reference distance; IVL, intravascular length; EVL, idealised extravascular length.

Direct output	n	MAE [95% CI], mm	RMSE [95% CI], mm	Bias [95% CI], mm	Absolute error P_{90}, mm	Reference-window hits/ n	Hit rate [95% CI], %
		8.47]	11.23]	4.85]			

All estimates are from the same 94 external patients. Bias is the prediction minus the reference value. Intervals for MAE, RMSE and bias are 95% percentile intervals from 5000 bootstrap resamples of complete records; 90th-percentile absolute errors are point estimates without intervals. Reference-window hit proportions and their Clopper–Pearson intervals are shown for the four total-distance targets, by side and zone. Estimates are displayed to two decimal places, with full precision retained in the source tables.

Abbreviations: MAE, mean absolute error; RMSE, root mean squared error; CI, confidence interval; P90, the 90th percentile of absolute error (90th percentile); T, total reference distance; IVL, intravascular length; EVL, idealised extravascular length.

Individualised estimates versus constants

Development-cohort means and medians also fell within all 94 zone I windows on both sides. Midpoint error distinguished individualised estimates from these constants even where hit proportions did not. Compared with the mean constants, FAST reduced right and left zone I MAE by 7.6 mm (95% CI 4.4 to 11.0) and 6.3 mm (3.0 to 9.7), respectively. In zone the corresponding reductions were 4.7 mm (2.1 to 7.2) and 3.7 mm (1.3 to 6.2). Comparisons with the median constants showed the same direction (Supplementary Table S18, panel A, and Supplementary Fig. S8).

In zone , mean constants gave 76/94 right-sided and 73/94 left-sided hits, and median constants gave 79/94 and 75/94. For left zone , paired comparisons with FAST gave Holm-adjusted McNemar *P* values of < 0.001 for the mean constant and 0.003 for the median. The conservative Bonferroni simultaneous bounds for the hit differences nevertheless included zero for all 52 comparisons. These bounds were calculated separately from the Holm-adjusted tests; the hit comparisons did not establish superiority (Supplementary Table S18, panels B and C, and Supplementary Table S19).

Number of measurements and choice of model

Along the forward-selection path with a fixed penalty of α = 10, increasing the number of inputs from one to three reduced the mean MAE across the four T targets from 16.72 to 14.76 mm in the development cohort and from 15.74 to 13.99 mm in the external cohort. Mean MAE reached its minimum with four inputs in development and six externally, whereas the highest mean hit rates occurred with three or four inputs in development and fourteen externally (Fig. 1, Supplementary Table S1). Of the three fixed configurations, FAST (three inputs), PRECISE (six) and FULL (fifteen) had external mean MAEs of 13.99, 13.92 and 14.00 mm. The paired difference for PRECISE minus FAST was − 0.07 mm (95% CI − 0.44 to 0.29) and that for FULL minus FAST was + 0.01 mm (− 0.47 to 0.49; Supplementary Table S2).

In the nested comparison of ten model families using the same three measurements, ridge regression had the lowest development mean MAE across the four T targets (14.65 mm), followed by elastic net and quadratic polynomial ridge regression (both 14.66 mm). The ranking changed externally after the

final models were fitted to all 547 development patients (Supplementary Table S20). RBF-SVR (gamma=scale) had a mean MAE of 13.810 mm versus 13.992 mm for ridge regression, a paired difference of − 0.182 mm (pointwise 95% CI − 0.869 to 0.458; simultaneous 95% CI − 1.080 to 0.716). Simultaneous intervals included zero for all nine comparisons (Fig. 3a,b; Supplementary Table S10A).

The lower mean MAE for SVR did not translate into more zone hits. SVR hit 87/94 right and 88/94 left windows, compared with 88/94 and 90/94 for ridge regression. CatBoost had a higher right zone MAE than ridge regression but one more hit; on the left it had a lower MAE but one fewer hit. Every model hit every external zone I window (Supplementary Tables S9A1, S9A2 and S20).

In the extended comparison, external mean MAE was 13.984 mm for OLS, 13.992 mm for ridge regression, 13.946 mm for TabPFN v2 and 14.002 mm for RBF-SVR (C = 100, gamma = 0.1). Paired differences from ridge regression were − 0.008 mm for OLS (simultaneous 95% CI − 0.045 to 0.029) and − 0.047 mm for TabPFN v2 (− 0.393 to 0.300); the simultaneous intervals contained zero for all seven comparisons (Fig. 3c,d; Supplementary Tables S7A and S7B). Across 1000 development-patient bootstrap samples, the mean standard deviation of predictions at fixed external inputs ranged from 1.260 to 1.659 mm across the four T targets for OLS and from 1.230 to 1.617 mm for ridge regression (Supplementary Table S8B).

Shifting each input by ± 5 mm in eight hypothetical corner-point scenarios changed individual T predictions by up to 4.682 mm for ridge regression and 11.703 mm for SVR from the ten-family comparison (Supplementary Table S9E2). For the same models, median single-patient offline inference times for ten outputs were 0.144 and 1.532 ms, respectively, versus 52794.249 ms for TabPFN v2. For TabPFN v2, single-patient predictions differed from predictions in a 94-patient batch by up to 0.063 mm, exceeding the prespecified numerical consistency tolerance. Its reported external performance and perturbation results used batch calls (Supplementary Tables S9D1 and S9D2).

Increasing the training set from 60 to 547 patients reduced mean MAE across the four T targets for ridge regression from 14.497 to 13.992 mm in the same external patients (Supplementary Table S10B). Extrapolating from 547 to 2735 patients gave different results for CatBoost relative to ridge regression. The power-law curve predicted a crossing in favour of CatBoost at about 2232 patients, but 60.60% of external-patient resamples showed no crossing in either direction. The exponential curve had no point-estimate crossing in this range, and 85.79% of resamples showed no crossing in either direction. Neither curve identified a stable training size at which CatBoost would overtake ridge regression (Supplementary Tables S10C2a and S10C2b; Supplementary Figs. S4 and S5).

Component distances, calibration and sensitivity

FAST estimated idealised EVL with an MAE of 6.9 mm on the right and 7.2 mm on the left. Right and left IVL MAEs were 16.3 and 16.7 mm in zone I and 11.2 and 11.7 mm in zone III (Table 2). Compared with the development-cohort means, EVL MAE decreased by 3.2 mm (95% CI 1.4 to 5.0) on the right and 3.6

mm (1.9 to 5.4) on the left. For zone III IVL, the intervals for the corresponding reductions of 0.9 mm (− 0.4 to 2.3) and 0.8 mm (− 0.5 to 2.2) both contained zero (Supplementary Table S13).

In zone III, overestimation of EVL partly offset underestimation of IVL. Right and left EVL biases of + 3.3 and + 3.1 mm accompanied IVL biases of − 2.8 and − 3.0 mm, leaving *T* biases of + 0.4 and 0.0 mm (Table 2, Supplementary Fig. S10). Replacing predicted IVL with its development mean increased right and left zone III MAE from 11.6 and 12.2 mm to 13.2 and 13.1 mm. Replacing predicted EVL instead increased MAE to 14.7 and 14.6 mm, and using both means gave 16.3 and 15.9 mm (Supplementary Table S14).

PRECISE had lower EVL MAE than FAST on the right (difference − 0.43 mm, pointwise 95% CI − 0.85 to − 0.04) and left (− 0.57 mm, − 1.02 to − 0.14); FULL also had lower left EVL MAE (− 0.92 mm, − 1.77 to − 0.11). Holm-adjusted P values across 24 component comparisons were 0.643, 0.284 and 0.607, respectively, with consistent sign-flip results (Supplementary Table S11). Right and left calibration slopes were 1.12 (95% CI 0.90 to 1.32) and 1.11 (0.89 to 1.32) in zone I, and 1.01 (0.79 to 1.26) and 1.01 (0.77 to 1.26) in zone III. All four intervals allowed departures from unity (Supplementary Table S23).

In right zone □, two estimates fell below the reference window and four above; on the left, all four misses were below the window (Supplementary Table S24). Under hypothetical independent Gaussian errors with an SD of 5 mm in each input, expected right and left zone □ MAEs were 11.7 and 12.3 mm, with expected hit rates of 93.3% and 94.9%. When input errors were bounded by ± 5 mm, 85/94 cases on each side remained hits under every permitted perturbation (Supplementary Table S16). With predictions held fixed, increasing assumed axial length from 40 to 50 to 60 mm reduced zone □ hit rates from 93.6% to 88.3% to 79.8% on the right and from 95.7% to 86.2% to 72.3% on the left (Supplementary Table S17, Supplementary Fig. S6).

In the development cohort, selecting ridge configurations by hit count rather than shared MAE changed zone □ MAE by less than 0.1 mm on each side. Right-sided hits remained at 505/547, while left-sided hits increased from 505 to 506. For quadratic polynomial ridge regression, hit-prioritised selection yielded one fewer hit per side. Adding the strongest boundary penalty tested ($\eta = 1$) produced one additional left zone □ hit at the cost of higher MAE; selection by MAE retained a zero penalty for both zone □ targets (Supplementary Tables S4 to S6; Supplementary Fig. S3).

The offline software returned ten direct distance estimates from three body-surface inputs. Its outputs agreed with the frozen FAST equations on synthetic inputs (Supplementary Methods 12 and Supplementary Fig. S7).

Discussion

Principal findings

Three body-surface measurements improved continuous estimation of CT reference distances compared with development-cohort constants, whereas reference-window hit rates alone did not capture this improvement. In zone I, the wider window accommodated both individualised predictions and constants despite their different midpoint errors. Zone distinguished the methods more clearly, but rankings based on continuous error, tail error and reference-window hits remained inconsistent.

The aortic zone midpoint has equal longitudinal margins within the reference window. Error relative to this midpoint describes accuracy along the CT reference coordinate.

Comparison with previous studies

Fixed distances require no individual measurements, but individual variation matters in zone . In a retrospective trauma CT study, Eliason et al. assumed a 4 cm balloon and fixed distances of 48 cm for zone I and 28 cm for zone . Simulated boundary-crossing rates were 0.4% and 33.3%, respectively. Adjustment for torso length reduced the zone rate to 26.4% [13]. In 57 patients, Weng et al. used the midpoint between the xiphoid process and umbilicus as a surface landmark; predicted zone accuracy on CTA was 84.2% on the left and 86.0% on the right [11].

Access geometry also differs between studies (Supplementary Table S25). Chien et al. incorporated a 45° skin-to-artery assumption into total insertion length [16], whereas van de Voort et al. used an intravascular femoral origin [17]. We estimated the idealised extravascular and intravascular distances separately and compared predictions with constants and other models against the same reference.

Chien et al. reported simulated in-zone rates of 99.43% for a fixed distance of 47.5 cm in zone I and 82.1% for 25.5 cm in zone . Their rule combining the umbilical level with a ratio of surface distances yielded a predicted zone hit rate of 94.6% in development and 95.88% in 97 patients at another hospital [16]. We adapted the distance-ratio offset equation using a locally learned access anchor. In our 94 external patients, the adapted rule yielded right and left zone MAEs of 14.2 and 14.3 mm and hit counts of 81/94 and 80/94, compared with 11.6 and 12.2 mm and 88/94 and 90/94 for FAST (Supplementary Table S26). Changing the anchor precludes attributing these differences to xiphoid measurement error.

Van de Voort et al. reported a mean vertical skin depth to the left common femoral artery of 34.7 mm (SD 17.1 mm), correlated with body mass index (r = 0.654) [17]. Forward selection identified body thickness and width in our development cohort. Both dimensions describe body habitus: anterior-posterior and transverse abdominal diameters correlate with abdominal adiposity on CT [27].

A review of binary clinical prediction models found no discrimination advantage for machine learning over logistic regression in comparisons at low risk of bias [28]. A vascular surgery review found that 20 of 22 comparisons favoured machine learning, but 201 of 212 studies assessed were at high risk of bias [29]. Selecting a classifier on the basis of its observed error can make that estimate optimistic [30]; inadequate sample size can also compromise model development and evaluation [31]. In our learning curves, performance was measured for training sizes from 60 to 547 patients. Beyond that range, model rankings and crossing points depended on the assumed curve form.

Our development cohort of 547 was larger than those studied by Okada, MacTaggart, Pezy and Weng [8–11]. Other approaches used taped torso height [12], skeletal landmarks [13], or lengths specific to age, sex and access side [17]. The three-input model required neither age nor medical history and estimated ten total and component distances.

Interpretation of component results and model selection

A total bias close to zero concealed opposing extravascular and intravascular biases in zone III. Replacing either predicted component with its development mean increased total-distance MAE on both sides. Reporting the two components shows which errors offset each other and which part of the estimate benefits from individual measurements.

Expected squared and absolute losses are minimised by a conditional mean and median, respectively [22]. Maximising the probability of falling within an individual reference window is a different objective (Supplementary Note 1). Penalising predictions outside the window sometimes added a hit at the cost of greater midpoint error. With different penalties for total and component distances, the fitted total need no longer equal the sum of the fitted components.

With the same three inputs, SVR in the ten-family comparison and TabPFN v2 in the extended comparison had slightly lower mean MAE than ridge regression, but all paired simultaneous intervals included zero. These results do not establish equivalence. The lower mean error for SVR did not translate into more zone III hits. OLS and ridge regression had similar external mean errors. Ridge penalisation modestly reduced coefficient and prediction variation, and its explicit equations allow independent recalculation offline.

FAST required the fewest raw information items of the three fixed configurations and had mean error across the four total distances similar to PRECISE and FULL. Six or fifteen inputs reduced MAE point estimates for some components, but these differences were not confirmed after multiplicity correction.

Clinical and research implications

The FAST equations allow offline calculation from three measurements without fluoroscopy or a workstation. Future studies could test tape measurements in place of CT-derived dimensions in battlefield, prehospital and other resource-limited settings. In our external cohort of 94 patients from three centres, there were too few zone misses to identify which patients were most likely to have larger errors.

Future placement studies should report midpoint error, reference-window hits and component distances together, with the reference window and device geometry specified in advance. TRIPOD + AI distinguishes model development from performance evaluation [18], and methodological guidance emphasises study design when interpreting performance metrics [32].

Limitations

Both cohorts were drawn from routine CTA examinations. Patients were selected for anatomical eligibility, not traumatic haemorrhage. Restriction to ages 16 to 60 years excluded older patients, in whom vascular calcification and tortuosity are more common. Disease criteria and autoencoder exclusions may also have removed relevant variation: reconstruction-error outliers do not necessarily reflect measurement error. Resampling the 547 retained development patients cannot recover variation removed by screening.

The CT-derived dimensions were treated as fixed. Morrison et al. used tape-measured torso height [12]; here, the contribution of each dimension was assessed under ideal measurement conditions on CT. Tape repeatability and the feasibility of reproducing the CT definitions during resuscitation remain to be assessed.

The idealised extravascular segment is a geometric reference rather than a measurement along the catheter path. Insetting the CT boundaries by 20 mm at each end assumed a 40 mm balloon; longer assumed balloons reduced the zone hit rate even though the predictions were unchanged. For an assumed length of 60 mm, hit rates fell to 79.8% on the right and 72.3% on the left. Applying the estimates to a particular device requires its actual balloon length, catheter zero point and centre-to-tip offset.

This CTA study did not evaluate actual balloon placement, occlusion time, complications or survival.

Conclusions

Three CT-derived body dimensions improved continuous estimation of CT reference distances compared with constants while balancing midpoint error and the number of measurements. Balloon-centre MAE was 15.8 to 16.3 mm in zone I and 11.6 to 12.2 mm in zone , with simulated reference-window hit rates of 100% and 93.6% to 95.7%, respectively. These results may inform the development of fluoroscopy-free depth estimation based on bedside tape measurements.

Abbreviations

BT
body thickness
BW
body width
CI
confidence interval
CPU
central processing unit
CT
computed tomography
CTA

computed tomography angiography
EVL
idealised extravascular length
IVL
intravascular length
MAE
mean absolute error
NCTH
non-compressible torso haemorrhage
OLS
ordinary least squares
PROBAST + AI
Prediction model Risk Of Bias ASsessment Tool + AI
RBF
radial basis function
REBOA
resuscitative endovascular balloon occlusion of the aorta
RMSE
root mean squared error
SD
standard deviation
SMD
standardised mean difference
SNP
suprasternal notch-to-pubic symphysis distance
SVR
support vector regression
T
total reference distance
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + AI.
FAST,PRECISE and FULL label the three-,six- and fifteen-input configurations,respectively.

Declarations

Ethics approval and consent to participate

The Ethics Review Committee of the Chinese People's Liberation Army General Hospital approved the study (2026 – 784) and waived the requirement for informed consent. The approval framework covered all participating centres and the eligible age range. The study was conducted in accordance with the Declaration of Helsinki.

Consent for publication

No identifiable participant information is reported in this manuscript. The Ethics Review Committee of the Chinese People's Liberation Army General Hospital approved submission of the manuscript for publication.

Competing interests

The authors declare that they have no competing interests.

Additional files

Supplementary Materials (.docx): Supplementary information. Supplementary methods, notes, figures S1 to S11 and tables S1 to S32.

Funding

This study was supported by the National Key Research and Development Programme (2024YFC2419000) and the National Science and Technology Major Project (2023ZD0504302). The funders had no role in study design, data collection, analysis, interpretation or manuscript preparation.

Author Contribution

Conceptualisation: JY, FZ, SW, DR, WG; Data curation: all authors; Formal analysis: JY, DR; Funding acquisition: WG; Investigation: all authors; Methodology: JY, DR, WG; Project administration: JY, FZ, DR, WG; Resources: LC, HW, TM, WG; Software: JY, DR; Supervision: DR, WG; Validation: JY, YG, DR; Visualisation: JY, YG; Writing – original draft: JY, DR; Writing – review and editing: all authors. All authors reviewed and approved the manuscript for submission.

Acknowledgements

Not applicable.

Data Availability

The data were collected at military hospitals. The original CT images (DICOM files) cannot be shared. Participant-level measurements and analysis code are not publicly deposited. The authors may be contacted through the corresponding author (guoweiplagh@sina.com) for formal requests for derived

measurements, analysis code, or the research software, including requests with institutional approval and requests for humanitarian use. Access is subject to applicable institutional and regulatory approvals. Users may request help with software installation and use, as well as further improvements, through the same contact. The model parameters are reported in the Supplementary Information.

References

1. Eastridge BJ, Mabry RL, Seguin P, Cantrell J, Tops T, Uribe P, et al. Death on the battlefield (2001–2011): implications for the future of combat casualty care. J Trauma Acute Care Surg. 2012;73(6 Suppl 5):S431–7. 10.1097/ta.0b013e3182755dcc.
2. Morrison JJ, Stannard A, Rasmussen TE, Jansen JO, Tai NR, Midwinter MJ. Injury pattern and mortality of noncompressible torso hemorrhage in UK combat casualties. J Trauma Acute Care Surg. 2013;75(2 Suppl 2):S263–8. 10.1097/TA.0b013e318299da0a.
3. Johnson NL, Wade CE, Fox EE, Meyer DE, Fox CJ, Moore EE, et al. Determination of optimal deployment strategy for REBOA in patients with non-compressible hemorrhage below the diaphragm. Trauma Surg Acute Care Open. 2021;6(1):e000660. 10.1136/tsaco-2020-000660.
4. Northern DM, Manley JD, Lyon R, Farber D, Mitchell BJ, Filak KJ, et al. Recent advances in austere combat surgery: Use of aortic balloon occlusion as well as blood challenges by special operations medical forces in recent combat operations. J Trauma Acute Care Surg. 2018;85(1S Suppl 2):S98–103. 10.1097/TA.0000000000001966.
5. Campbell K, Naumann DN, Remick K, Wright C. Damage control resuscitation and surgery for indigenous combat casualties: a prospective observational study. BMJ Mil Health. 2021;167(1):18–22. 10.1136/jramc-2019-001228.
6. Jansen JO, Hudson J, Cochran C, MacLennan G, Lendrum R, Sadek S, et al. Emergency Department Resuscitative Endovascular Balloon Occlusion of the Aorta in Trauma Patients With Exsanguinating Hemorrhage: The UK-REBOA Randomized Clinical Trial. JAMA. 2023;330(19):1862–71. 10.1001/jama.2023.20850.
7. Harfouche MN, Bugaev N, Como JJ, Fraser DR, McNickle AG, Golani G, et al. Resuscitative Endovascular Balloon Occlusion of the Aorta in surgical and trauma patients: a systematic review, meta-analysis and practice management guideline from the Eastern Association for the Surgery of Trauma. Trauma Surg Acute Care Open. 2025;10(1):e001730. 10.1136/tsaco-2024-001730.
8. Okada Y, Narumiya H, Ishi W, Iiduka R. Anatomical landmarks for safely implementing resuscitative balloon occlusion of the aorta (REBOA) in zone 1 without fluoroscopy. Scand J Trauma Resusc Emerg Med. 2017;25(1):63. 10.1186/s13049-017-0411-z.
9. MacTaggart JN, Poulson WE, Akhter M, Seas A, Thorson K, Phillips NY, et al. Morphometric roadmaps to improve accurate device delivery for fluoroscopy-free resuscitative endovascular balloon occlusion of the aorta. J Trauma Acute Care Surg. 2016;80(6):941–6. 10.1097/ta.0000000000001043.

10. Pezy P, Flaris AN, Prat NJ, Cotton F, Lundberg PW, Caillot JL, et al. Fixed-Distance Model for Balloon Placement During Fluoroscopy-Free Resuscitative Endovascular Balloon Occlusion of the Aorta in a Civilian Population. JAMA Surg. 2017;152(4):351–8. 10.1001/jamasurg.2016.4757.
11. Weng D, Qian A, Zhou Q, Xu J, Xu S, Zhang M. A new method using surface landmarks to locate resuscitative endovascular balloon occlusion of the aorta based on a retrospective CTA study. Eur J Trauma Emerg Surg. 2022;48(3):1945–53. 10.1007/s00068-021-01686-0.
12. Morrison JJ, Stannard A, Midwinter MJ, Sharon DJ, Eliason JL, Rasmussen TE. Prospective evaluation of the correlation between torso height and aortic anatomy in respect of a fluoroscopy free aortic balloon occlusion system. Surgery. 2014;155(6):1044–51. 10.1016/j.surg.2013.12.036.
13. Eliason JL, Derstine BA, Horbal SR, Wang NC, Holcombe SA, Chiu CH, et al. Computed tomography correlation of skeletal landmarks and vascular anatomy in civilian adult trauma patients: Implications for resuscitative endovascular balloon occlusion of the aorta. J Trauma Acute Care Surg. 2019;87(1S Suppl 1):S138–45. 10.1097/ta.0000000000002247.
14. Li RF, Gong XF, Xu HB, Lin JT, Zhang HG, Suo ZJ, et al. Age affects vascular morphology and predictiveness of anatomical landmarks for aortic zones in trauma patients: implications for resuscitative endovascular balloon occlusion of the aorta. Eur J Trauma Emerg Surg. 2024;50(6):3099–108. 10.1007/s00068-024-02512-z.
15. Takata T, Yamada K, Yamamoto M, Kondo H. REBOA Zone Estimation from the Body Surface Using Semantic Segmentation. J Med Syst. 2023;47(1):42. 10.1007/s10916-023-01938-z.
16. Chien CY, Lee YL, Jeng MJ, Liu CJ. Using computed tomography to evaluate anatomic landmarks in Taiwanese trauma patients for insertion of resuscitative endovascular balloon occlusion of the aorta: a retrospective cohort study. Int J Surg. 2024;110(12):7900–8. 10.1097/js9.0000000000002079.
17. van de Voort JC, Verbeek BB, van der Burg BLSB, Hoencamp R. Exploring aortic morphology and determining variable-distance insertion lengths for fluoroscopy-free resuscitative endovascular balloon occlusion of the aorta (REBOA). World J Emerg Surg. 2024;19(1):29. 10.1186/s13017-024-00557-4.
18. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
19. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, 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. 10.1136/bmj-2024-082505.
20. Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. 1970;12(1):55–67. 10.1080/00401706.1970.10488634.
21. Pavlou M, Ambler G, Seaman S, De Iorio M, Omar RZ. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. Stat Med. 2016;35(7):1159–77. 10.1002/sim.6782.

22. Gneiting T. Making and Evaluating Point Forecasts. J Am Stat Assoc. 2011;106(494):746–62. 10.1198/jasa.2011.r10138.
23. Efron B. Nonparametric standard errors and confidence intervals. Can J Stat. 1981;9(2):139–58. 10.2307/3314608.
24. Clopper CJ, Pearson ES. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika. 1934;26(4):404–13. 10.1093/biomet/26.4.404.
25. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–7. 10.1007/BF02295996.
26. Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat. 1979;6:65–70.
27. Yim JY, Kim D, Lim SH, Park MJ, Choi SH, Lee CH, et al. Sagittal abdominal diameter is a strong anthropometric measure of visceral adipose tissue in the Asian general population. Diabetes Care. 2010;33(12):2665–70. 10.2337/dc10-0606.
28. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22. 10.1016/j.jclinepi.2019.02.004.
29. Li B, Feridooni T, Cuen-Ojeda C, Kishibe T, de Mestral C, Mamdani M, et al. Machine learning in vascular surgery: a systematic review and critical appraisal. NPJ Digit Med. 2022;5(1):7. 10.1038/s41746-021-00552-y.
30. Boulesteix AL, Strobl C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. BMC Med Res Methodol. 2009;9(1):85. 10.1186/1471-2288-9-85.
31. Riley RD, Ensor J, Snell KIE, Archer L, Whittle R, Dhiman P, et al. Importance of sample size on the quality and utility of AI-based prediction models for healthcare. Lancet Digit Health. 2025;7(6):100857. 10.1016/j.landig.2025.01.013.
32. Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific standards are a disservice to patients and society. J Clin Epidemiol. 2021;138:219–26. 10.1016/j.jclinepi.2021.05.018.

Figures

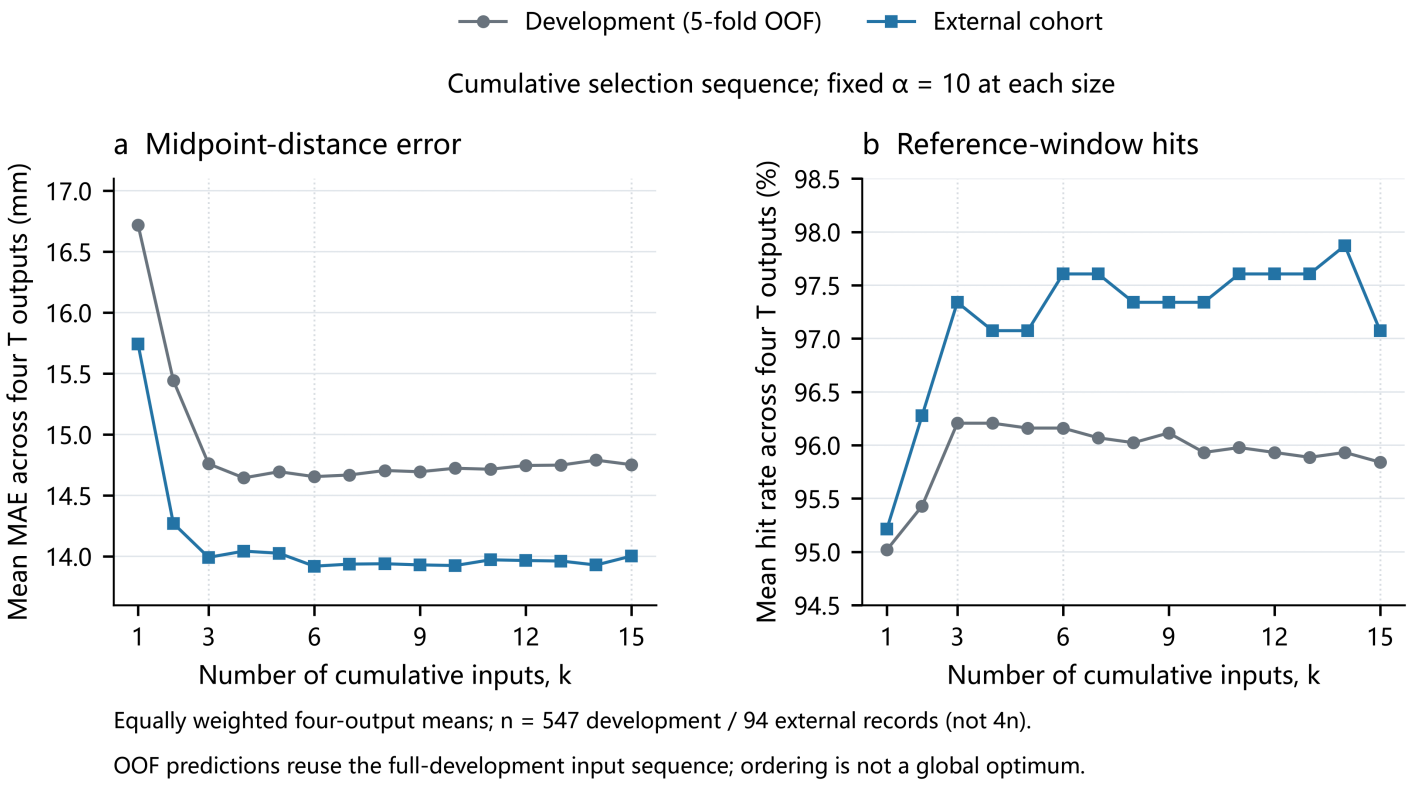

Figure 1

Mean midpoint error decreased from one to three measurements along the forward-selection path

a, Mean absolute error (MAE) across the four total distances (T); b, mean reference-window hit rate across the four T targets. The horizontal axis shows the number of inputs along one forward-selection path (k = 1 to 15), with α fixed at 10; alternative subsets of the same size are not shown. Development results are from five-fold out-of-fold evaluation, and external results are from 94 patients.

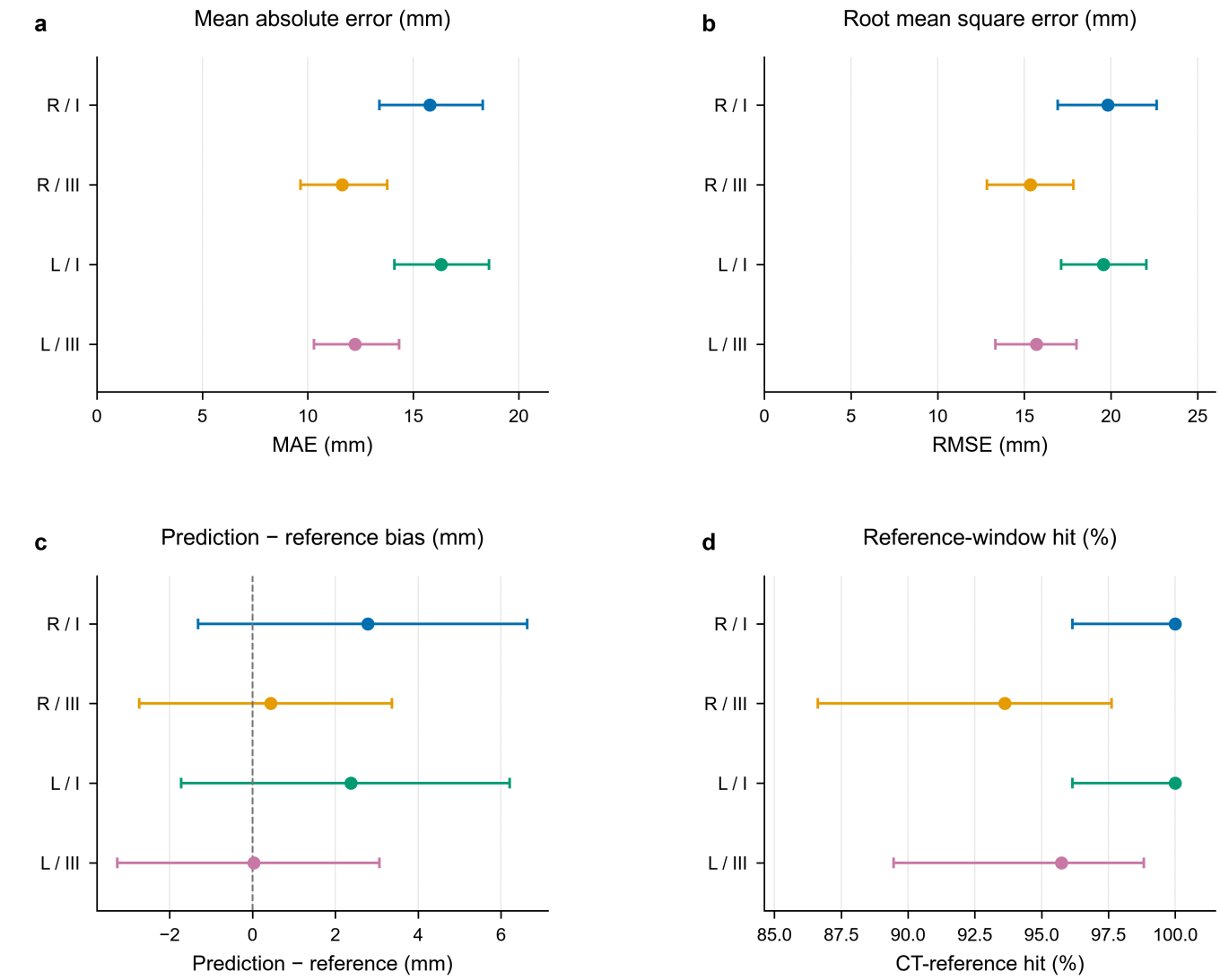

Figure 2

All zone I reference windows were hit despite non-zero midpoint errors

a, Mean absolute error; b, root mean squared error; c, bias; d, reference-window hit rate. Each panel shows results for right and left zones I and Ⅲ in the 94 external patients. Error bars represent 95% confidence intervals: a, b and c use pointwise percentile bootstrap intervals from complete records, and d uses Clopper–Pearson intervals.

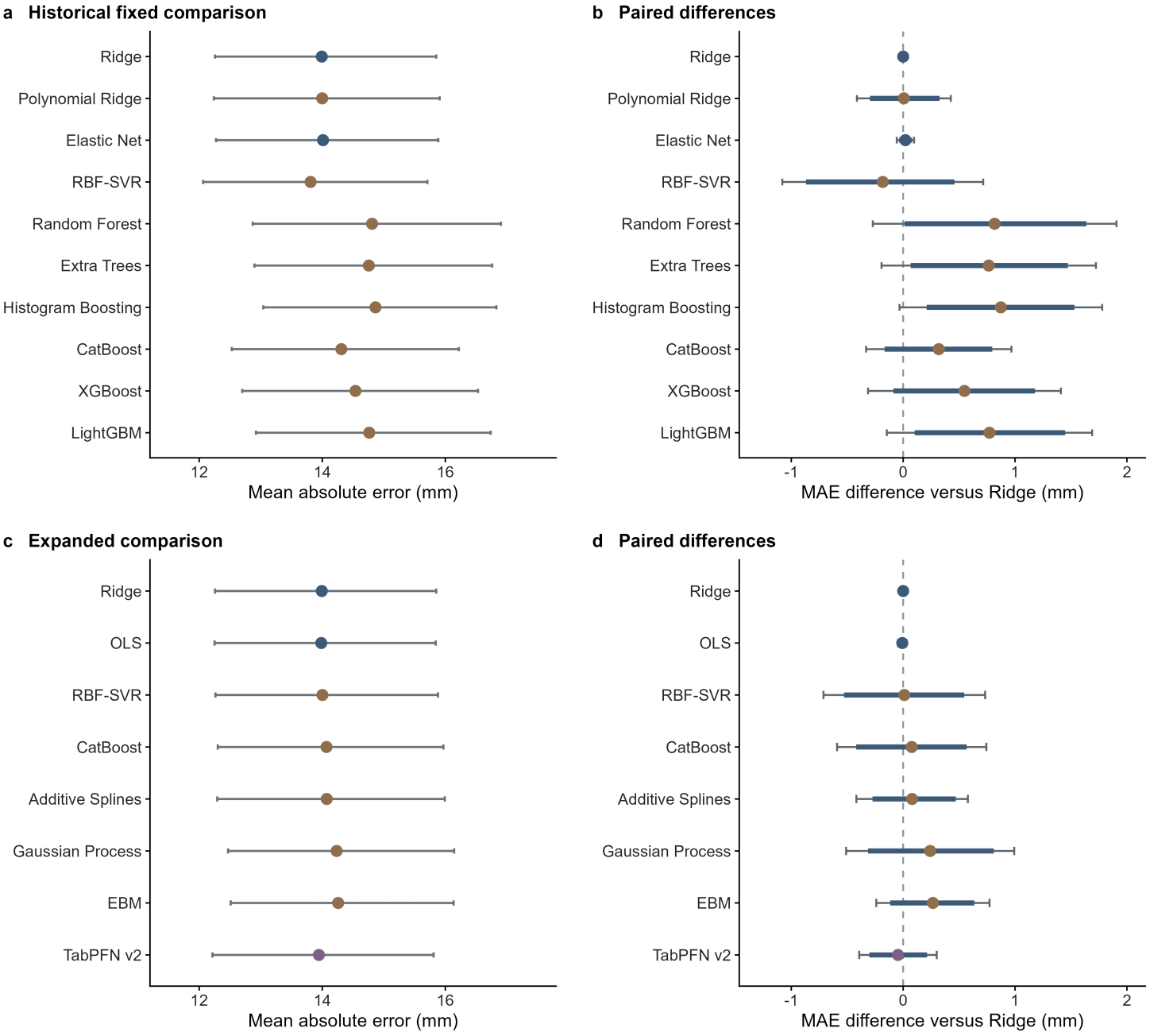

Figure 3

Simultaneous intervals for paired mean absolute error differences all included zero

a, Mean absolute error (MAE) across four total distances for ten model families.
b, Paired MAE differences versus ridge regression.
c, Mean MAE for eight models in the extended comparison.
d, Their paired MAE differences versus ridge regression.
All used the same three measurements in 94 external patients. Intervals are pointwise 95% in a and c; thick blue and thin grey lines in b and d show pointwise and simultaneous 95% intervals. Negative differences favour the comparator.

Supplementary Files

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

- SupplementaryMaterials.docx