scientific reports

Check for updates

OPEN

# Towards a more reliable assessment of aortic diameters using a Bayesian Z-score

Luca Bindini[1]✉, Laurence Campens[2], Jesse Davis[3], Laura Muiño-Mosquera[4,5], Simon D'hulst[4], Julie De Backer[5,6], Stefano Nistri[7,8] & Paolo Frasconi[1]

**The Z-score is a conceptually simple and widely adopted standard for assessing aortic dilatation from echocardiographic measurements. It is routinely used to monitor patient progression and schedule follow-up checks. However, several criticisms have been raised due to the intrinsic limitations of the typically homoscedastic and linear predictive models. In this paper, we reinterpret the Z-score as a quantitative measure of the *aleatoric uncertainty* affecting aortic diameters, after indexing by a limited number of predictive variables. This view reveals an additional, previously overlooked limitation: the presence of *epistemic uncertainty*, arising from limited or biased reference datasets. When epistemic uncertainty is high, the Z-score becomes unreliable, yet current tools fail to indicate this. We therefore adopt a Bayesian reformulation based on heteroscedastic Gaussian process regression, where diameters and their aleatoric uncertainties are modeled as random variables. In this framework, the Z-score itself is random, and clinicians receive both an expected value and a *high density interval* quantifying epistemic uncertainty. Trained on a merged dataset of 1,947 healthy subjects, our Bayesian Z-score detects more dilatations in at-risk patients, identifies uncertain cases, and offers a more reliable basis for clinical decision-making.**

**Keywords** Z-score, Aortic dilatation, Heteroscedastic Gaussian process regression

Thoracic aortic (TA) dilatation is a clinically relevant feature associated with adverse outcomes (e.g., aortic dissection or rupture and aortic regurgitation)[1]. Typically, it is evaluated using two-dimensional transthoracic echocardiography (TTE), computed tomography, or magnetic resonance. In daily practice, TTE remains the first-line modality thanks to its availability, safety, and cost. However, the definition of TA dilatation depends on the chosen approach, for example, absolute cut-offs, sex-adjusted thresholds, or body-size–indexed thresholds.

As the detection of TA dilatation should result in personalized follow-up strategies, both an accurate definition of normalcy and the availability of tools for assessing normalcy are key aspects for clinical and research purposes. Notably, TTE normal limits of thoracic aortic size are not immediately comparable across the vast literature on the subject[2]. Reasons include differences in demographic and anthropometric characteristics of the study populations, different TTE modes, and different interface (leading-to-leading vs. inner- to-inner) and timing (end-systole vs. end-diastole) strategies for measurements. Moreover, the currently available reference values for aortic dimensions have notable shortcomings that restrict their clinical applicability. First, some studies assess the aortic size at a single TA level (usually the sinuses of Valsalva). Second, only a few studies make software available for assessing normalcy in new patients (e.g., by means of a Z-score calculator). Third, the majority of available methods are only applicable to specific age-groups, or provide graphic nomograms separated for sex, body size and age groups, while only two studies provide calculators for each aortic level on a wide age range[2,3]. Thus, particularly in certain ages (e.g., patients in the age range 15–18) different calculators may result in potentially inconsistent results.

Among the available approaches to define normalcy, conventional Z-scores figure prominently, as they summarize how far a measured diameter $y$ deviates from its expected value given the patient context x. However, they do not indicate model uncertainty. In particular, they mix *aleatoric* uncertainty (irreducible variability

[1]AI Lab, Department of Information Engineering, University of Florence, Firenze 50139, Italy. [2]Department of Cardiology, Rigshospitalet, 2100 Copenhagen, Denmark. [3]DTAI Research Unit, Department of Computer Science & Leuven.AI, KU Leuven, 3001 Leuven, Belgium. [4]Division of Pediatric Cardiology, Department of Pediatrics, Ghent University Hospital, 9000 Ghent, Belgium. [5]Center for Medical Genetics, Ghent University Hospital, 9000 Ghent, Belgium. [6]Department of Cardiology, Ghent University Hospital, 9000 Ghent, Belgium. [7]Cardiology Service, CMSR Veneto Medica, 36077 Altavilla Vicentina, Italy. [8]Department of Cardio-Thoraco-Vascular Sciences and Public Health, University of Padua, 35128 Padua, Italy. ✉email: luca.bindini@unifi.it

across subjects with identical covariates) with *epistemic* uncertainty, which arises when the reference data are sparse around x. The latter is precisely when clinicians would benefit from an explicit warning, as management decisions based on borderline Z values may otherwise be misled by model extrapolation rather than data-supported estimates.

To make these ideas concrete, we first review the conventional Z-score used in echocardiography, then summarize its known limitations, and finally introduce a Bayesian formulation that separates aleatoric from epistemic uncertainty and returns, alongside the point estimate, a high-density interval that quantifies model reliability.

**Contributions**

This study makes three contributions:

- We recast the echocardiographic Z-score using a Bayesian framework[4] where $Z$ is a random variable whose posterior is derived from a heteroscedastic Gaussian-process model. In this formulation, patients are characterized by an expected value and a 95% high-density interval (HDI).
- We train on a merged reference population of $N = 1{,}947$ healthy subjects obtained from two previously published datasets: $D_1$ (Campens et al.)[3] and $D_2$ (Frasconi et al.)[2].
- We study the prevalence of dilatation in an out-of-sample at-risk cohorts and explicitly identify borderline cases whose HDI straddles the clinical threshold, situations in which the score should be interpreted with caution.

## The classic Z-score

For a given context vector x describing a patient (e.g., age, sex, height, weight), the aortic diameter $y$ measured at a specific anatomical level is modeled by the conditional distribution $p(y \mid \mathrm{x})$. Under the assumption that this conditional is Gaussian, the classic Z-score is defined as

$$Z_{classic} = \frac{y_* - m(\mathrm{x}_*)}{s(\mathrm{x}_*)} \tag{1}$$

where $\mathrm{x}_*$ and $y_*$ denote the measured context and diameter, respectively, for a given patient. The mean function, $m$, aiming to predict the expected diameter $\mathbb{E}[y|\mathrm{x}_*]$ and the standard deviation function, $s$, aiming to predict the expected standard deviation of the diameter, $\sqrt{\mathbb{V}\mathrm{ar}[y|\mathrm{x}_*]}$, are computed by a predictive model trained on the study population, i.e., a dataset of selected healthy subjects. The denominator in Equation 1 can be immediately interpreted as the predictive uncertainty due to the inherent randomness in the diameters given a fixed context. In statistics and machine learning this is called *aleatoric* uncertainty[5–8] (AU) and cannot be reduced by adding more subjects to the dataset. Intuitively, *aleatoric* uncertainty captures the natural between-subject variability that persists even with identical covariates. On the other hand, *epistemic* uncertainty (introduced in Section 1.3) reflects a lack of model knowledge due to limited or biased reference data, which can be reduced by collecting more representative data.

Note that if the predictive model is deterministic, after the patient is fully observed, $Z$ is not random. Indeed, in the whole literature on reference values for the aorta, the predictive model is a linear regression model trained by minimizing the mean squared error[3,9–14]. Additionally, the dependency of $s$ on the context is often dropped, using a single global residual instead. scale estimated on the training set, typically the root mean squared error (RMSE) of the fitted regression. This yields a homoscedastic model and the following commonly used plug-in Z-score

$$Z_{classic} = \frac{y_* - (w^\intercal \mathrm{x}_* + b)}{\mathrm{RMSE}} \tag{2}$$

where $w$ and $b$ are the model's parameters.

## Well-known limitations of the classic Z-score

Albeit straightforward to understand and implement, there are simplifying assumptions behind Equation 2 that may critically affect normalcy assessment. Several papers have argued against these simplifications and warned about their effects[15–17]. The major problems are summarized below.

*Nonlinearity*

Approximately, diameters grow proportionally with age in children, but then tend to saturate in adults, yielding an overall nonlinear behavior[15,18]. Similar nonlinear effects have been observed for BSA and BMI[3,19,20]. Diverse ad-hoc remedies have been proposed to circumvent the nonlinear dependence of aortic diameters with x. One is to limit the study to children, e.g., defined as age $< 18$ years[21], or adults, e.g., defined as age $\geq 15$ years[10]. Unfortunately, Z-scores from separate models differ near the cutoff age and deciding which of the two models should be used may be unclear and impractical, particularly when follow-up is needed. Additionally, the strategy leads to a proliferation of models if cutoffs are to be applied to several variables. This proliferation is undesirable because it increases maintenance burden, creates discontinuities around arbitrary cut-offs, and places additional cognitive load on clinicians who must choose among multiple, partially overlapping tools. The use of sex-specific models is also common[21], even when the goal is to have a single model for all ages[3]. Another common remedy is to apply Box&Cox[22] nonlinear transformations to some variables before linear regression. Sensible examples include logarithms, exponentials, and polynomials[3,18,20,21]. As an example, the model in Campens et al.[3] is structured as

$$\begin{aligned}\log m(\mathrm{x}_*) = {} & \mathbb{1}\{\mathrm{Sex} = \mathrm{m}\}\,(w_{a,m}\log \mathrm{Age} + w_{b,m}\mathrm{BSA} + b_m) \\ & + \mathbb{1}\{\mathrm{Sex} = \mathrm{f}\}\,(w_{a,f}\log \mathrm{Age} + w_{b,f}\mathrm{BSA} + b_f)\end{aligned} \tag{3}$$

with body surface area computed as[23] $\mathrm{BSA} = 0.007184 \times \mathrm{Height}^{0.725} \times \mathrm{Weight}^{0.425}$. However, these transformations are somewhat arbitrary, and, in general, a nonlinear function of several variables cannot be reduced to a linear function applied to individually transformed variables.

*Heteroscedasticity*

By *heteroscedasticity* we mean that the variance of aortic diameters, $\mathbb{V}\mathrm{ar}[y \mid \mathrm{x}]$, changes with the patient context x. In other words, the spread of normal diameters is not constant across age, sex, height, weight or BSA. Not only mean values of diameters change nonlinearly with age and height, but also their variances. This means that the classic Z-score (Equation 2) does not take into account differences in population variability for a given context vector x. Averaging variance over all subjects unavoidably leads to overestimates in contexts with low variability and underestimates in contexts with high variability[15,19]. Log-transformations applied to context variables, such as age and height, can somewhat mitigate heteroscedasticity, but do not completely eliminate it. Heteroscedasticity has been addressed in various forms. Early approaches attempted to reduce it by splitting data into separate groups[24]. Later, Altman[25] proposed a method where variance is estimated by training a second regression model on the absolute residuals of the first model, a technique that may be seen as a simplified, linear, and non-Bayesian version of the approach we proposed in this paper (Section 5). Altman's method has been used to assess aortic dilatation in some previous works[20,26].

## A lesser-known limitation: epistemic uncertainty

It is a standard fact that, in addition to the aleatoric uncertainty discussed above, supervised models are also affected by *epistemic* uncertainty (EU), which is inherent to the model and due to the lack of training data[7,27]. Intuitively, when a test point arrives that is too different from the training points, the model does not possess sufficient knowledge to make an accurate prediction. Unlike AU, EU may be reduced by increasing the dataset size *and* its representativeness. When studying aortic reference values, it is in principle possible (with some effort) to collect more normal subjects. However, it is not equally possible to ensure that the probability distribution from which the training set is sampled matches the distribution from which test points (patients) will be drawn at assessment time. This is because many constraints that define exclusion criteria need to be enforced when producing the dataset[28–30]. For example, obesity is a typical exclusion criterion in aortic normalcy studies, but when an obese person arrives as a test patient, the predicted diameter may be affected by epistemic uncertainty, if no subjects with similar age and weight were included in the training set (see an illustration in Figure 1, where obese individuals fall into regions of low data density).

As a result, the classic Z-score may fail to reflect true dilatation, since it could be based on an inaccurate diameter prediction. While this scenario is somehow unavoidable, it is a problem within the classic framework that the presence of EU is not exposed in any way to the clinicians as a warning that the Z-score is unreliable.

## The Bayesian Z-Score

Bayesian learning is an elegant and principled way of disentangling AU and EU[7]. We propose here to use this approach for echocardiographic assessment of aortic diameters.

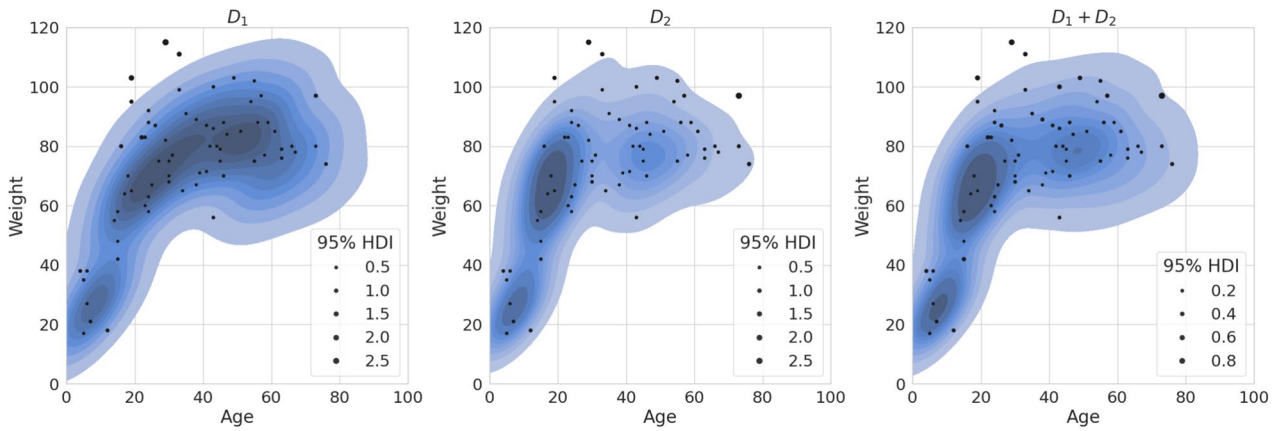


**Fig. 1**. Density plots for the male subjects in Campens et al.'s data ($D_1$), Frasconi et al.'s data ($D_2$), and the union of these two datasets ($D_1 \cup D_2$). Shaded areas cover 95% of the probability mass. Black markers correspond to patients with a Marfan syndrome diagnosis. It can be expected that predicted diameters (and therefore Z-scores) for patients falling in low-density regions will be affected by epistemic uncertainty because predictive models have been trained on a limited number of examples from these regions. This is indeed what we observe: the size of the black markers is proportional to the width of the 95% HDI. When the two datasets are merged, intervals shrink but remain larger for patients in low-density regions.

In the classic approach, as noted above, functions $m$ and $s$ in Equation (1), are fixed and thus $Z_{classic}$ is a point value. From the Bayesian perspective, $m$ and $s$ are instead sampled from a posterior distribution over functions. As a result, $Z$ becomes itself a random variable, rather than a point value. This is significant since the conditional density $p(z|\mathrm{x}_*, y_*)$ derived from the posterior provides us with information about EU, which can be revealed by inspecting its concentration. While distributional assessments of $Z$ are not exclusive to Bayesian inference in principle, the GP posterior predictive distribution provides a coherent and practically convenient way to account for uncertainty in patient-specific scores. In Figure 2, we show two examples of conditional densities leading to different patient assessments.

The expectation, $\mathrm{BZ} \doteq \mathbb{E}[Z]$, is a single number that may be treated in the same way as the classic Z-score. Epistemic uncertainty is quantified via the high-density interval (HDI), denoted as $\mathbb{I}[Z]$ and defined as the smallest interval of values for $Z$ such that $p(z|\mathrm{x}_*, y_*)$ integrates to a given threshold[31] (usually 95%). While in general the highest-density regions can be disjoint for multimodal densities, we found that with our data, the density of $Z$ is always unimodal; thus, the reported HDIs are single contiguous intervals.

Several tools can be employed to perform Bayesian regression and obtain the posteriors $m$ and $s$, and thus a Bayesian Z-score. In practice, we propose heteroscedastic Gaussian process regression[32] (HGPR) because of its elegance and because it is a nonlinear solution, thus helping not only to disentangle AU and EU, but also to address nonlinearity (Section 1.2.1) and heteroscedasticity (Section 1.2.2). Technical details are provided in Section 5.

## Dataset

### Healthy reference population

Healthy population was derived from two different previously published studies, both conducted according to the guidelines of the Declaration of Helsinki. The first one[3] recruited only healthy individuals aged $\geq$1 year from 2008 to 2013 (pediatric or adult cardiology department, Ghent University Hospital - Ghent University Hospital ethical committee study reference: B67020138099). The second one[2] collected healthy subjects aged $\geq$5 years (Department of Pediatric Cardiology, Meyer Hospital, Florence; Cardiovascular Center, ASU Trieste; and Cardiology Service, CMSR Veneto Medica, Altavilla Vicentina, Italy) between 2013 and 2017 Ethical review and approval were waived for this study since it used (i) only anonymized data, (ii) from databases produced during routine clinical activities, and (iii) after informed consent. Recruitment followed the previously published criteria[2,3].

### Echocardiographic protocol

Comprehensive transthoracic Doppler echocardiography was performed with commercially available phased-array systems, following a predefined protocol for acquisition, storage, and measurement. Parasternal long-axis views were optimized at four aortic levels: annulus, sinuses of Valsalva (SoV), sinotubular junction, and proximal ascending aorta (AA). The largest diameters were measured at end-diastole, perpendicular to the aortic axis, using the leading-edge to leading-edge technique (except the annulus). For the present study, only SoV and AA measurements were considered.

Height and weight were measured at the time of TTE. Body mass index (BMI) was computed as weight/$\mathrm{height}^2$. Body-surface area (BSA) was calculated according to the Du Bois formula.

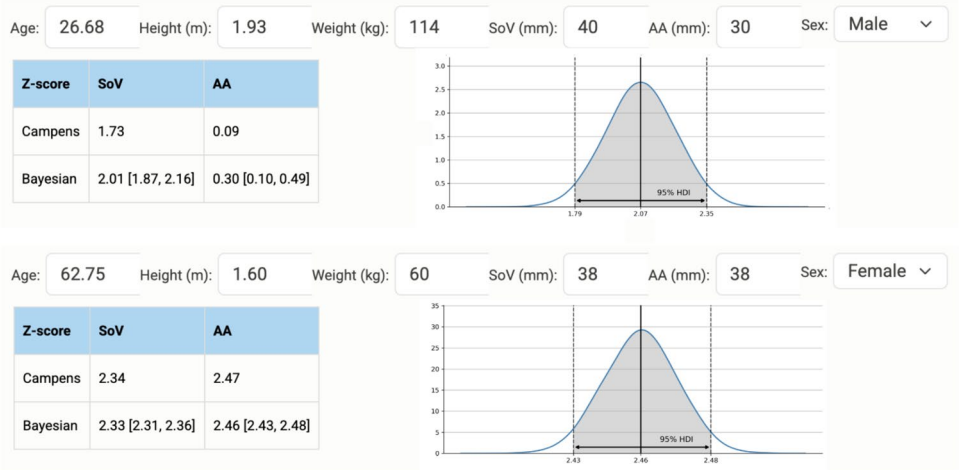


**Fig. 2**. Conditional densities for the Bayesian Z-score. Top: SoV for a 26-year-old male patient with 1.93m height, 114 kg weight, and 40 mm SoV diameter; the expected $Z$ in this case is near 2 but the density is not concentrated and its interval is large, i.e., there is a significant probability that $Z$ is actually above or below 2. Bottom: AA for a 62-year-old female patient with 1.60m height, 60 kg weight, and 38 mm AA diameter; in this case the density is concentrated and the whole interval is well above 2. Therefore, the patient can be safely assumed to be dilatated. Both patients in these examples were diagnosed with Marfan syndrome. Note that conditional densities of $Z$ are not Gaussian and thus not necessarily symmetric around the mean.

| Variable | Total | Male | Female |
|---|---|---|---|
| Age [years] | 37 [17–53] | 24 [16–48] | 43 [24–56] |
| Height [cm] | 167 [158–174] | 173 [164–180] | 163 [157–168] |
| Weight [kg] | 64 [53–75] | 70 [55–80] | 60 [52–69] |
| BSA [$m^2$] | 1.7 [1.5–1.9] | 1.9 [1.6–2.0] | 1.6 [1.5–1.8] |
| BMI [$kg/m^2$] | 22.6 [19.7–25.7] | 23.0 [19.6–25.9] | 22.3 [19.8–25.5] |
| SoV diameter [mm] | 29 [26–32] | 30 [27–33] | 28 [25–31] |
| AA diameter [mm] | 28 [25–31] | 28 [25–31] | 28 [25–32] |

**Table 1**. Baseline characteristics of the healthy reference population ($N = 1{,}947$). Values are median [interquartile range].

| | SoV | | | AA | | |
|---|---|---|---|---|---|---|
| Cohort | Campens | Bayesian | straddle | Campens | Bayesian | straddle |
| BAV | 33.6% | 35.6% | 1.7% | 64.7% | 68.9% | 0.9% |
| Marfan | 53.0% | 63.2% | 7.7% | 15.9% | 16.8% | 3.5% |

**Table 2**. Prevalence of aortic dilatation ($Z > 2$) in Marfan and BAV patients according to Campens' Z-score and the proposed Bayesian Z-score, for two anatomical levels: sinuses of Valsalva (SoV) and ascending aorta (AA). We also report the percentage of "straddle" cases whose $\mathbb{I}[Z]$ crosses the clinical threshold.

The two source cohorts used compatible acquisition protocols and measurement conventions; after harmonizing units, measurement sites, and exclusion criteria, we treated them as a single reference population for training.

The studies in Italy and Belgium were approved by the local Independent Ethics Committees and the Institutional Review Board (IRB) of the respective hospital. All subjects participating in the study gave written informed consent.

### Sample size and demographics

The merged reference population comprised $N = 1{,}947$ healthy subjects aged 1–89 years. Age distribution was: 384 (19.7%) $\leq$ 15 years, 551 (28.3%) between 16–35, 594 (30.5%) between 36–55, and 418 (21.5%) $\geq$ 56 years. As shown in Table 1, females were on average older, shorter, and lighter, with smaller BSA than males. Aortic diameters were generally larger in men at the sinuses of Valsalva (SoV), while ascending aortic (AA) diameters tended to be similar.

### At-risk cohorts

For independent validation, we evaluated two patient cohorts deemed at risk of thoracic aortic dilatation and not having undergone aortic surgery: (i) 117 subjects with Marfan syndrome (MFS) and related disorders, diagnosed according to the revised Ghent criteria and confirmed with the identification of a (likely) pathogenic FBN1, LOX, TGFBR1, TGFBR2, TGFB2, TGFB3, SMAD2, SMAD3, MYH11, or ACTA2 variant. All patients were prospectively evaluated at the Pediatric or Adult Cardiology department of Ghent University Hospital between March 2024 and September 2025. These patients were part of a study approved by the Ethics Committee of the Ghent University Hospital (study reference: ONZ-2023-0552) and conducted in accordance with the principles of the Declaration of Helsinki. All participants provided written informed consent prior to participation. (ii) 351 subjects with bicuspid aortic valve (BAV), previously described in Frasconi et al.[2] and recruited based on the above mentioned process. Both studies were conducted in accordance with the principles of the Declaration of Helsinki and all participants provided written informed consent prior to participation.

## Results

### Prevalence of dilatation in at-risk patients is larger with Bayesian Z-score compared to Campens Z-score

Table 2 summarizes the percentages of patients with ($Z > 2$) in two pathological cohorts according to Campens Z-score and the Bayesian Z-score, where both models were trained on the same reference population of 1947 healthy subjects. We also report the percentage of "straddle" cases, i.e. patients whose $\mathbb{I}[Z]$ crosses the clinical threshold ($Z_{\max} > 2$ and $Z_{\min} < 2$).

To provide a more intuitive picture of how the two definitions of positivity overlap, Figure 3 reports Venn diagrams for the same cohorts and aortic levels. Each diagram contrasts patients identified as dilatated by the Campens Z-score and by the Bayesian Z-score. The intersection highlights concordant cases, whereas the exclusive regions capture patients considered positive by one method only. The Bayesian-specific counts also report in parentheses the number of "straddle" cases, i.e. patients whose $\mathbb{I}[Z]$ crosses the clinical threshold. Patients in the external region (outside both circles) are those with neither Campens nor BZ above the threshold.

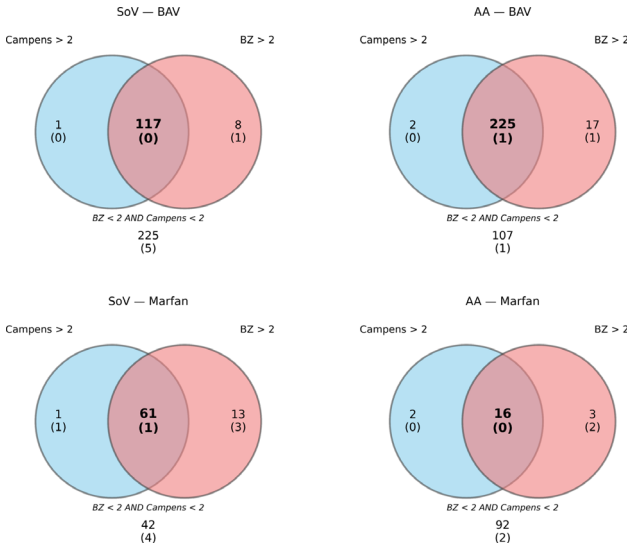


**Fig. 3**. Venn diagrams of patient distributions across cohorts and anatomical levels (SoV and AA). Blue circles = Campens$> 2$; red circles = $BZ > 2$; purple = intersection. Numbers inside circles indicate patient counts; numbers in parentheses indicate "straddle" cases detected by Bayesian Z-score ($Z_{\max} > 2$, $Z_{\min} < 2$). Patients outside both sets (gray area) satisfy Campens$< 2$ and $BZ < 2$. Note that circle areas are schematic and not proportional to patient counts.

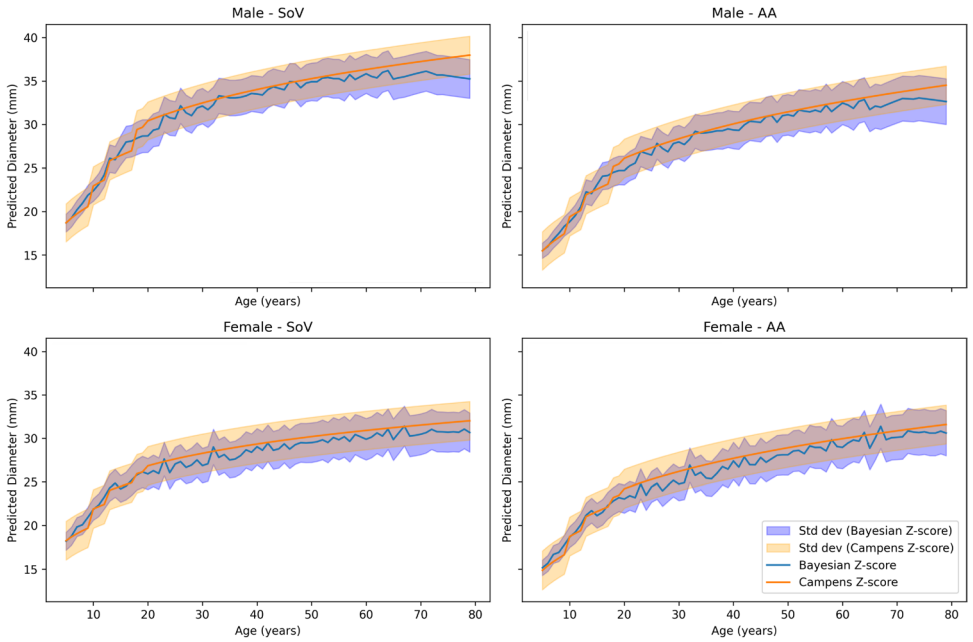


**Fig. 4**. Predicted mean aortic diameter versus age in the healthy reference population ($N = 1\,947$), stratified by sex. Orange = Campens model; blue = heteroscedastic Gaussian-process regressor (foundation of the Bayesian Z-score). Shaded areas depict one predictive standard deviation of the Bayesian Z-score. Left panels refer to the sinuses of Valsalva (SoV); right panels to the ascending aorta (AA).

The diagrams confirm the results of Table 2: Bayesian Z-score detects more positives overall, but importantly it also makes explicit which of these fall in the epistemic–uncertain zone (straddle cases).

## The log-based transformation largely captures the nonlinear effect

Figure 4 simultaneously addresses the two methodological issues raised in Section 1, namely the *non-linearity* of the age–diameter relationship and the *heteroscedasticity* of its residual variance.

Although the heteroscedastic Gaussian-process regressor (HGPR) is a fully non-parametric model, its learned mean curves (blue) track those of the classical Campens regression (orange) with striking fidelity. Both exhibit

| Model | SoV | | AA | |
|---|---|---|---|---|
| | MAE [mm] | RMSE [mm] | MAE [mm] | RMSE [mm] |
| Campens (BSA) | 2.15 | 2.75 | 2.16 | 2.81 |
| Bayesian GP (BSA) | 2.14 | 2.75 | 2.15 | 2.80 |
| Bayesian GP (Height + Weight) | 2.13 | 2.73 | 2.14 | 2.79 |

**Table 3**. Cross-validated prediction error on the healthy reference set for three different models: Campens' Z-score which uses BSA as a feature, the GP which uses BSA as a feature, and the GP that uses height and weight instead of BSA. Results are reported separately for two anatomical levels: SoV and AA. MAE = mean absolute error; RMSE = root-mean-square error (both in mm).

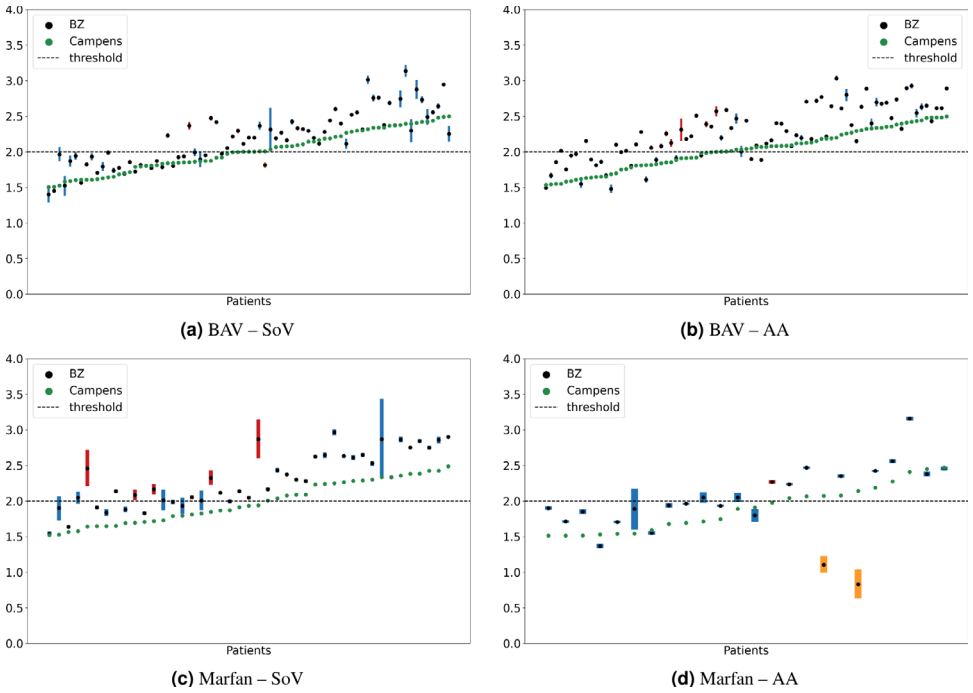


**Fig. 5**. BZ (black dots; $\mathbb{I}[Z]$ in blue) versus Campens Z-score (green) across the two pathological cohorts and two anatomical levels. Patients in each panel are sorted by increasing Campens Z-score. Blue $\mathbb{I}[Z]$ bars that cross the clinical threshold at 2 indicate cases where epistemic uncertainty prevents a definitive classification. Bars colored in red correspond to patients with $BZ > 2$ and Campens$< 2$, whereas orange bars indicate the opposite case ($BZ < 2$ and Campens$> 2$).

the characteristic logarithmic shape, rapid growth during childhood and adolescence followed by a plateau in adulthood, thus empirically confirming the long-standing assumption that a log-type relation is an adequate first-order description of normal aortic growth. This is the first empirical validation of such an assumption on a dataset larger than those used in previous studies.

The HGPR also reveals substantial input-dependent noise: the predictive standard deviation (shaded band) is lower in childhood and increases with age. This pattern, invisible to a homoscedastic linear model, quantitatively supports the need for a context-dependent dispersion term when computing Z-scores.

Predictive diameters were estimated via 5-fold cross-validation on the healthy reference set. For each fold, models were trained on four folds and evaluated on the held-out fold; MAE and RMSE were computed on held-out predictions and averaged across folds.

Table 3 shows that the HGPR not only models heteroscedastic noise explicitly, but also achieves a marginal improvement in mean absolute and root-mean-square error over the Campens score. A further advantage of the Bayesian framework is the ability to include *height* and *weight* as *separate* inputs, rather than compressing them into body-surface area (BSA). This extra degree of freedom lets the Gaussian process (GP) adapt to subjects who share the same BSA yet differ in physique (e.g., tall-lean vs. short-obese) yielding (marginal) improvements over diameter prediction accuracy.

## Epistemic uncertainty and its relevance

Figure 5 illustrates how epistemic uncertainty manifests in practice. In the Marfan cohort (upper panels), several HDIs straddle the diagnostic cutoff, signaling that the true Z-score could plausibly lie on either side of the critical

value; these ambiguous points correspond to extreme anthropometric profiles, such as very tall or under-/overweight individuals who are under-represented in the reference dataset. In the BAV cohort (lower panels), the HDIs are generally narrower, consistent with better coverage of the corresponding contexts in the reference data. Across all panels, the green Campens curve offers no indication of this lack of confidence, whereas the blue HDI bars provide an immediate visual cue that the Bayesian estimate should be interpreted with caution.

Another key point illustrated by Figure 5 is that two patients may exhibit the same expected Z-score but very different HDI widths. In such cases, the GP disentangles *aleatoric* from *epistemic* uncertainty: the expected Z reflects the intrinsic variability of diameters (AU), while the interval length explicitly quantifies the lack of knowledge due to sparse training data (EU). This information is absent from conventional scores, yet it is crucial for tailoring the level of clinical confidence in each prediction.

To further explore how epistemic uncertainty manifests in different regions of the context space, we computed two-dimensional heatmaps of the HDI interval width $|\mathbb{I}[Z]|$ as a function of *age* (5–80 years) and *BSA* (1–2.5 $\mathrm{m}^2$). For each anatomical site (sinuses of Valsalva – SoV, and ascending aorta – AA), we considered both sexes and fixed three reference diameters (28, 32, and 36 mm).

Plots in Figure 6 can be interpreted as "uncertainty maps": the color scale encodes the width of $\mathbb{I}[Z]$, i.e., the degree of epistemic uncertainty. Blue areas mark regions where the posterior is narrow and predictions are reliable, while red areas highlight zones where the reference data are sparse and predictions are less trustworthy. White corresponds to $|\mathbb{I}[Z]| = 1$, which we adopt as a practical threshold beyond which the uncertainty becomes clinically relevant.

Epistemic uncertainty is lowest in the central part of the distribution (adolescents and adults with average body sizes) where both $D_1$ and $D_2$ provided abundant training data. In contrast, uncertainty grows toward the boundaries of the age–BSA plane (young children, elderly, or very small/large body sizes), illustrating how the lack of reference subjects directly inflates the HDI. Importantly, the dependence of these maps on the diameter is practically negligible; the observed variability does not exhibit any systematic trend with respect to it.

Clinically, these maps provide an immediate visual tool to judge the reliability of the Bayesian Z-score. A patient located in the blue region can be reliably classified, whereas patients falling within the red areas should not be straightforwardly considered abnormal, as the apparent deviation may partly reflect insufficient reference data rather than a true pathological dilatation.

## Methodological details

Normalcy can be established via anomaly detection[33] (AD), a well established AI setting where the goal is to identify data points (patients in our case) that deviate from other observations (healthy subjects). While AD can be effectively applied to the study of aortic reference values[2], an anomaly in the classic setting is sought on the whole set of available variables so that patients that have rare joint values for sex, age, weight and height may be flagged as anomalous regardless of their aortic size. Although infrequent, these cases are undesired. Therefore, we adopt here the more recent *contextual* anomaly detection[34,35] (CAD) framework, a conditional setting where variables are split into two groups: contextual variables (in our case sex, age, weight and height) and behavioral variables (in our case aortic diameters). In CAD, a data point is deemed to be anomalous if the relationship between the two groups of variables deviates from the norm. Although never formally framed in this way, the classic Z-score can be immediately recognized as a CAD algorithm because the underlying regression problem aims to model the conditional distribution of diameters given contextual variables. To instantiate CAD, we adopt the Bayesian nonparametric heteroscedastic GP regression (HGPR) framework proposed by Bindini et al.[4], which yields a posterior predictive distribution for the behavioral variable conditional on the context. In our application, we fit *two* coupled Gaussian processes as in the original framework: one for the conditional mean of the aortic diameter and one for the logarithm of its standard deviation. This construction separates aleatoric variability (through an input-dependent scale) from epistemic uncertainty captured by the posterior over functions, addressing the limitations of point-estimate Z-scores discussed in Section 1.

**Dual–GP formulation** We assume

$$y \mid f_1, f_2, \mathrm{x} \;\sim\; \mathcal{N}\big(f_1(\mathrm{x}),\, \exp\{2\, f_2(\mathrm{x})\}\big), \tag{4}$$

where $f_1$ models the conditional mean and $f_2$ the *log-standard-deviation*. Independent GP priors are assigned,

$$f_1 \sim \mathcal{GP}(m_1, k_1), \qquad f_2 \sim \mathcal{GP}(m_2, k_2), \tag{5}$$

with constant mean functions $(m_1, m_2)$, rational–quadratic kernels $(k_1, k_2)$ and ARD (Automatic Relevant Determination) length scales[36].

In Gaussian process regression, the kernel $k(\mathrm{x}, \mathrm{x}')$specifies how function values at two inputs are correlated a priori, thereby encoding assumptions about smoothness and characteristic length-scales across the covariate space[36]. Commonly used choices include the squared-exponential (RBF), Matérn, and rational-quadratic (RQ) kernels. As shown in Bindini et al.[4], switching among these kernels yields only marginal differences in the resulting scores and uncertainty profiles in comparable HGPR settings. For this reason, we use the RQ kernel with ARD length-scales as a flexible and robust default, since it can accommodate multiple effective length-scales through a single parametric form. In particular, the RQ kernel can be written as

$$k_{\mathrm{RQ}}(\mathrm{x}, \mathrm{x}') = \sigma^2 \left(1 + \frac{1}{2\alpha} \sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{\ell_d^2}\right)^{-\alpha}, \tag{6}$$

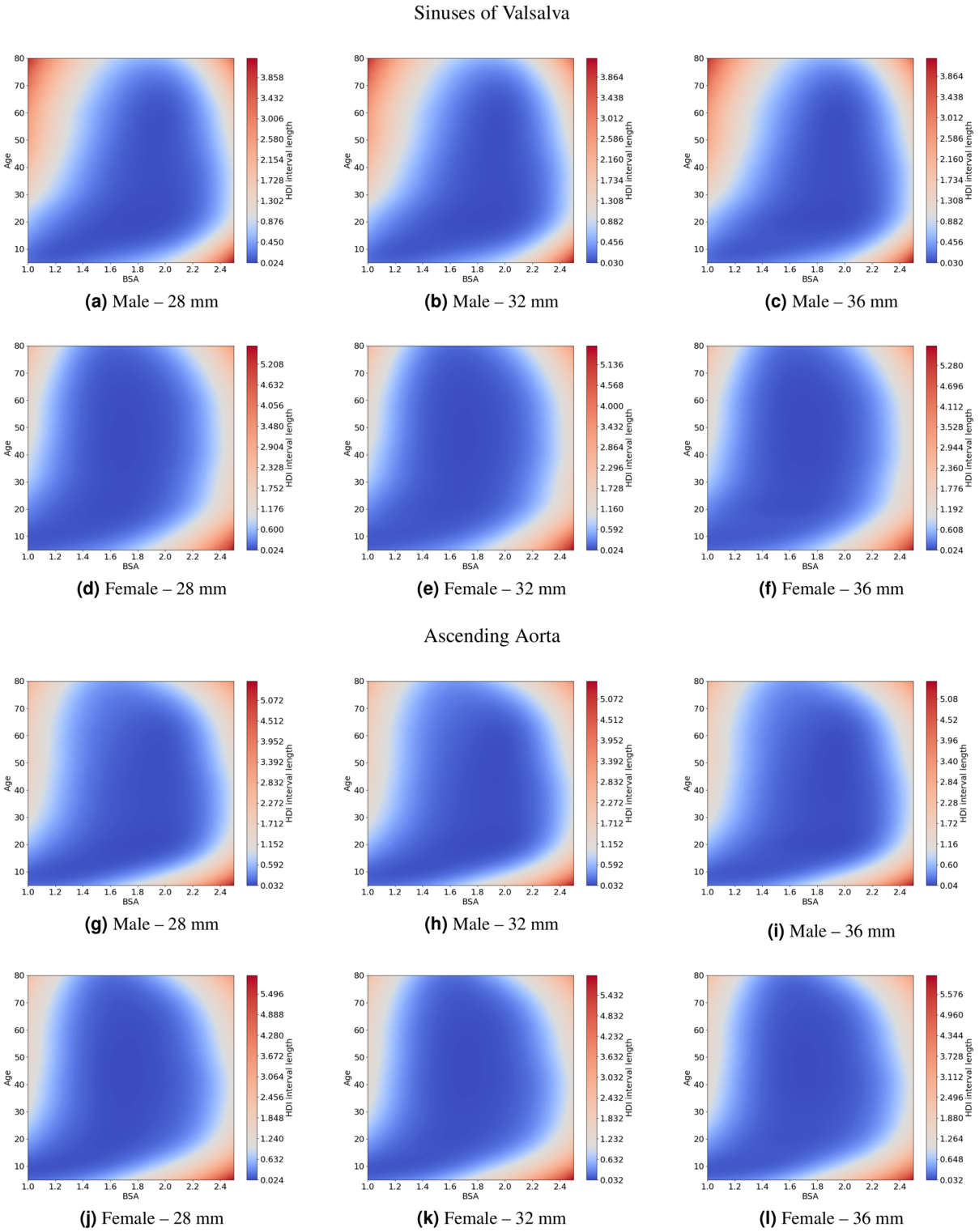


**Fig. 6**. Uncertainty maps. Each panel corresponds to a fixed reference diameter (28mm, 32 mm, or 36 mm). Color is indexed by the length of the 95% HDI so that blue areas indicate regions with reliable predictions (because of lower epistemic uncertainty). White corresponds to an HDI length of 1, which we adopt as a practical threshold beyond which uncertainty becomes clinically relevant.

where $\sigma^2$ is the signal variance, $\alpha > 0$ controls the relative weighting of different length-scales, and $\{\ell_d\}_{d=1}^{D}$ are the ARD length-scales (one per input dimension). Notably, as $\alpha \to \infty$ the RQ kernel approaches the RBF kernel, which can therefore be interpreted as a limiting special case of this prior.

The Bayesian Z-score is therefore

$$Z(\mathrm{x}, y) = \frac{y - f_1(\mathrm{x})}{\exp\{f_2(\mathrm{x})\}} \tag{7}$$

Because $f_1$ is Gaussian and $\exp[-f_2]$ is log-normal, we can derive the expectation in closed-form[4]:

$$\mathrm{BZ} \doteq \mathbb{E}[Z] = (y - m_1(\mathrm{x}))e^{-m_2(\mathrm{x})+\sigma_2^2(\mathrm{x})/2} \tag{8}$$

The $95\,\%$ highest-density interval $\mathbb{I}[Z] = [Z_{\min}, Z_{\max}]$ is defined as the shortest interval containing $95\,\%$ of this density (obtained with Monte Carlo sampling), and its width $|\mathbb{I}[Z]|$ quantifies the epistemic component of uncertainty.

**Sparse variational inference** Exact GP inference scales cubically with the number of training points *N*. We employ a sparse variational approximation[37], representing each process through *M* inducing points initialized with $5\,\%$ randomly chosen points from the training data. While it is technically possible to increase *M* (or relax the approximation) to improve fidelity, larger *M* can substantially increase runtime while yielding only marginal changes in the resulting scores[4]. Optimization alternates natural-gradient steps ($\gamma = 0.02$) on the variational parameters with Adam updates ($\eta = 10^{-2}$) on kernel hyperparameters, for up to $4 \times 10^4$ epochs. Unless stated otherwise, we retain the default HGPR implementation settings of Bindini et al.[4] to facilitate reproducibility. We use the HGPR implementation available in GPflow[38] v. 2.12. Alternative scalable GP strategies include Nearest Neighbor Gaussian Processes (NNGP), which exploit sparse conditional independence structure to reduce computational cost[39]. All experiments were run on an NVIDIA RTX 2080 GPU.

## Discussion

This study introduces a Bayesian formulation of the Z-score for echocardiographic assessment of aortic diameters. By explicitly disentangling *aleatoric* from *epistemic* uncertainty, the Bayesian Z-score addresses three major limitations of conventional approaches: nonlinearity, heteroscedasticity of residual variance, and lack of an explicit representation of model uncertainty.

From a clinical standpoint, the Bayesian Z-score provides not only a point estimate of abnormality, but also an interval that quantifies the reliability of the prediction. Our experiments in Marfan and bicuspid-valve cohorts demonstrate that it detects a larger prevalence of dilatation compared to the Campens Z-score, while simultaneously identifying borderline cases where epistemic uncertainty prevents a definitive classification. This increase in prevalence is expected: by modeling nonlinear age and body-size dependencies and allowing for heteroscedastic residuals, the Bayesian GP tends to yield slightly smaller expected diameters in extreme anthropometric profiles. As a result, Z-scores in these patients are shifted upwards, uncovering cases of clinically relevant dilatation that the Campens model may underestimate. Cases with a wide HDI interval are of high clinical importance: current calculators may mislead clinicians into over- or under-estimating disease severity, whereas Bayesian Z-score explicitly flags these instances and may inform closer monitoring.

The heteroscedastic Gaussian-process regressor (HGPR) captures both the nonlinear age dependence of aortic growth and the input-dependent variability of diameters. The resulting model closely replicates the empirically observed logarithmic trend across childhood and adulthood, while revealing lower variance in early childhood. The cross-validated prediction errors show that the Bayesian approach achieves at least comparable, and in some cases marginally better, accuracy than classical regression, but with the practical advantage of explicit uncertainty quantification via highest-density intervals.

The two-dimensional heatmaps of HDI width across age and BSA provide a novel visualization tool for assessing where reference models are less reliable. We observed that epistemic uncertainty concentrates at the boundaries of the demographic distribution (children, elderly, very small or large body sizes). These results are consistent with the limited availability of healthy reference subjects in such regions and highlight the need for targeted data collection to improve model robustness.

Overall, the Bayesian Z-score represents a conceptually simple yet powerful extension of current practice. By coupling accurate diameter prediction with an explicit quantification of uncertainty, it offers clinicians not only a measurement of abnormality but also a measure of trust. This dual information may ultimately lead to more personalized management of patients at risk for aortic disease.

## Data availability

The datasets used in this study are potentially available upon reasonable request.

## Materials availability

An interactive online calculator is available at https://aorta-normalcy.unifi.it/to facilitate the use of the method by researchers and clinicians.

Received: 17 October 2025; Accepted: 23 March 2026
Published online: 27 March 2026

## References

1. Kim, J. B. et al. Risk of aortic dissection in the moderately dilated ascending aorta. *Journal of the American College of Cardiology* **68**, 1209–1219 (2016).
2. Frasconi, P. et al. Two-Dimensional Aortic Size Normalcy: A Novelty Detection Approach. *Diagnostics* **11**, 220. https://doi.org/10.3390/diagnostics11020220 (2021).

3. Campens, L. et al. Reference values for echocardiographic assessment of the diameter of the aortic root and ascending aorta spanning all age categories. *The American Journal of Cardiology* **114**, 914–920. https://doi.org/10.1016/j.amjcard.2014.06.024 (2014).
4. Bindini, L., Perini, L., Nistri, S., Davis, J. & Frasconi, P. Dealing with uncertainty in contextual anomaly detection. *Transactions on Machine Learning Research* (2026).
5. Kiureghian, A. D. & Ditlevsen, O. Aleatory or epistemic? Does it matter?. *Structural Safety* **31**, 105–112. https://doi.org/10.1016/j.strusafe.2008.06.020 (2009).
6. Kendall, A. & Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, vol. 30 (2017).
7. Hüllermeier, E. & Waegeman, W. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning* **110**, 457–506. https://doi.org/10.1007/s10994-021-05946-3 (2021) (**1910.09457.**).
8. Zhou, X., Liu, H., Pourpanah, F., Zeng, T. & Wang, X. A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing* **489**, 449–465. https://doi.org/10.1016/j.neucom.2021.10.119 (2022).
9. Roman, M. J., Devereux, R. B., Kramer-Fox, R. & O'Loughlin, J. Two-dimensional echocardiographic aortic root dimensions in normal children and adults. *The American journal of cardiology* **64**(507–512), 00846 (1989).
10. Devereux, R. B. et al. Normal limits in relation to age, body size and gender of two-dimensional echocardiographic aortic root dimensions in persons ≥15 years of age. *The American Journal of Cardiology* **110**, 1189–1194. https://doi.org/10.1016/j.amjcard.2012.05.063 (2012).
11. Lopez, L. et al. Relationship of echocardiographic z scores adjusted for body surface area to age, sex, race, and ethnicity. *Circulation: Cardiovascular Imaging* **10**, e006979. https://doi.org/10.1161/CIRCIMAGING.117.006979 (2017).
12. Cantinotti, M. et al. Nomograms for two-dimensional echocardiography derived valvular and arterial dimensions in caucasian children. *J. Cardiol.* **69**, 208–215. https://doi.org/10.1016/j.jjcc.2016.03.010 (2017).
13. Martinez-Millana, A. et al. Optimisation of children z-score calculation based on new statistical techniques. *PLoS One* **13**, e0208362. https://doi.org/10.1371/journal.pone.0208362 (2018).
14. Patel, H. N. et al. Normal values of aortic root size according to age, sex, and race: Results of the World Alliance of Societies of Echocardiography study. *J. Am. Soc. Echocardiogr.* **35**, 267–274. https://doi.org/10.1016/j.echo.2021.09.011 (2022).
15. Mawad, W., Drolet, C., Dahdah, N. & Dallaire, F. A review and critique of the statistical methods used to generate reference values in pediatric echocardiography. *J. Am. Soc. Echocardiogr.* **26**, 29–37. https://doi.org/10.1016/j.echo.2012.09.021 (2013).
16. Colan, S. D. The why and how of z scores. *J. Am. Soc. Echocardiogr.* **26**, 38–40. https://doi.org/10.1016/j.echo.2012.11.005 (2013).
17. Curtis, A., Smith, T., Ziganshin, B. & Elefteriades, J. The mystery of the z-score. *AORTA* **04**, 124–130. https://doi.org/10.12945/j.aorta.2016.16.014 (2016).
18. Daubeney, P. E. F. et al. Relationship of the dimension of cardiac structures to body size: An echocardiographic study in normal infants and children. *Cardiol. Young* **9**, 402–410. https://doi.org/10.1017/S1047951100005217 (1999).
19. Chubb, H. & Simpson, J. M. The use of z-scores in paediatric cardiology. *Ann. Pediatr. Cardiol.* **5**179. https://doi.org/10.4103/0974-2069.99622 (2012).
20. Sluysmans, T. & Colan, S. D. Structural measurements and adjustments for growth. *In Echocardiography in Pediatric and Congenital Heart Disease, chap.* **5**, 61–72. https://doi.org/10.1002/9781118742440.ch5 (John Wiley & Sons Ltd (2016).
21. Gautier, M. et al. Nomograms for aortic root diameters in children using two-dimensional echocardiography. *Am. J. Cardiol.* **105**, 888–894. https://doi.org/10.1016/j.amjcard.2009.11.040 (2010).
22. Box, G. E. P. & Cox, D. R. An analysis of transformations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **26**, 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x (1964).
23. Du Bois, D. & Du Bois, E. F. A Formula to Estimate the Approximate Surface Area If Height and Weight Be Known. *Archives of Internal Medicine* **XVII**, 863–871. https://doi.org/10.1001/archinte.1916.00080130010002 (1916).
24. Goldstein, H. The construction of standards for measurements subject to growth. *Hum. Biol.* **44**, 255–261 (1972).
25. Altman, D. G. Construction of age-related reference centiles using absolute residuals. *Stat. Med.* **12**, 917–924. https://doi.org/10.1002/sim.4780121003 (1993).
26. Fernandes, S. et al. Bicuspid aortic valve and associated aortic dilation in the young. *Heart* **98**, 1014–1019. https://doi.org/10.1136/heartjnl-2012-301773 (2012).
27. Senge, R. et al. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Inf. Sci.* **255**, 16–29. https://doi.org/10.1016/j.ins.2013.07.030 (2014).
28. Ricci, F. et al. Cardiovascular magnetic resonance reference values of mitral and tricuspid annular dimensions: The uk biobank cohort. *Journal of Cardiovascular Magnetic Resonance* **23**, 5. https://doi.org/10.1186/s12968-020-00688-y (2021).
29. Asch, F. M. et al. Need for a global definition of normative echo values–rationale and design of the World Alliance of Societies of Echocardiography normal values study (wase). *J. Am. Soc. Echocardiogr.* **32**, 157-162.e2. https://doi.org/10.1016/j.echo.2018.10.006 (2019).
30. Lancellotti, P. et al. Normal reference ranges for echocardiography: Rationale, study design, and methodology (norre study). *European Heart Journal - Cardiovascular Imaging* **14**, 303–308. https://doi.org/10.1093/ehjci/jet008 (2013).
31. Kruschke, J. K. *Doing Bayesian Data Analysis* 2nd edn. (Academic Press, 2015).
32. Goldberg, P., Williams, C. & Bishop, C. Regression with input-dependent noise: A gaussian process treatment. In *Advances in Neural Information Processing Systems*, vol. 10, 493–499 (MIT Press, 1997).
33. Chandola, V., Banerjee, A. & Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys* **41**, 15:1–15:58. https://doi.org/10.1145/1541880.1541882 (2009).
34. Song, X., Wu, M., Jermaine, C. & Ranka, S. Conditional anomaly detection. *IEEE Trans. Knowl. Data Eng.* **19**, 631–645. https://doi.org/10.1109/TKDE.2007.1009 (2007).
35. Li, Z. & van Leeuwen, M. Explainable contextual anomaly detection using quantile regression forests. *Data Min. Knowl. Discov.* **37**, 2517–2563. https://doi.org/10.1007/s10618-023-00967-z (2023).
36. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning* (Adaptive Computation and Machine Learning (MIT Press, Cambridge, 2006).
37. Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial intelligence and statistics*, 567–574 (2009).
38. Matthews, A. G. d. G. et al. Gpflow: A gaussian process library using tensorflow. *Journal of Machine Learning Research* **18**, 1–6 (2017).
39. Datta, A., Banerjee, S., Finley, A. O. & Gelfand, A. E. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *J. Am. Stat. Assoc.* **111**, 800–812 (2016).


## Author contributions

L.B. developed the method, implemented the models and pipelines, designed and ran all experiments, analyzed the data, prepared figures and tables, and wrote the original draft of the manuscript. L.C., J.D.B., and S.N. performed echocardiographies and aortic measurements on the reference population, provided data for analysis, and revised the final manuscript. J.D. provided feedback on the method and helped revise the manuscript. J.D.B.,

L.M., and S.D. performed echocardiographies and aortic measurements on the cohort of patients with Marfan syndrome and related disorders, provided data for analysis, and revised the final manuscript. S.N. performed echocardiographies and aortic measurements on the cohort of patients with bicuspid valve, provided data for analysis, and revised the final manuscript. P.F. conceived the method, supervised the computational work, and contributed to writing.

## Funding
The work of P.F. was partially supported by the European Union NextGenerationEU and the Italian National Recovery and Resilience Plan through the Ministry of University and Research (MUR), under Project CAI4DSA (CUP B13C23005640006).

## Declarations

## Competing interests
The authors declare no competing interests.

## Additional information
**Correspondence** and requests for materials should be addressed to L.B.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.