university of groningen

faculty of science and engineering

# Enhancing Aortic Valve Diameter Prediction: Accounting for Demographic Variability and Measurement Techniques

Master Project Mathematics

July 2014

Student: L. Dijkshoorn

First supervisor: Prof.dr. M.A. Grzegorczyk

Second supervisor: Prof.dr. G.A. Lunter

# Enhancing Aortic Valve Diameter Prediction: Accounting for Demographic Variability and Measurement Techniques

Leah Dijkshoorn

July 2024

**Abstract**

This thesis aims to develop an enhanced model for aortic valve (AV) diameter prediction that accounts for demographic variability and discrepancies in measurement techniques. Leveraging an extensive and diverse donor dataset from Cryolife Inc., previously unobservable patterns are analysed. This donor dataset is made up of physical AV diameter measurements, whereas diagnosis for aortic stenosis is done via echocardiographic AV diameter estimates. Thus, to address potential discrepancies between physical and echocardiographic measurements, a supplementary dataset from the University Medical Center Groningen (UMCG) was used for adults and models from existing literature were used for those 18 and under. The recent acquisition of the Lopez *et al.* dataset – which encompasses only those 18 and under – allowed for retroactive validation and a robust analysis of measurement biases. Following the data exploration of the donor dataset, an unexplainable trend was observed in the AV diameter measurements over time. Consequently, a bespoke segment neighbourhood algorithm was developed to objectively identify changepoints in the residuals. The trends between these changepoints were then corrected in the AV diameter measurements. It was found that the most suitable model to predict AV diameter was a generalised additive model (GAM) including tensor product smooth interaction terms. To account for heteroscedasticity due to the large demographic variability, a GAM was created to model the conditional standard deviation with respect to the demographic attributes/variables. Combined, these models can be used for Z-score computation, as is standard in the cardiology field. This research enhances predictive accuracy and uncertainty quantification for aortic valve diameters, contributing to more reliable assessments for diagnosis and aortic valve replacement surgery.

# Contents

# 1 Introduction

The primary objective of this study is to develop a refined model of aortic valve (AV) diameter measurements that accounts for both demographic variability and measurement technique discrepancies. By comparing the final model against previous models and applying it to extreme cases – such as individuals who are obese or significantly smaller than average – we aim to enhance the predictive accuracy and uncertainty quantification of AV diameter measurement techniques.

Traditional datasets used in such studies have been limited by size and demographic diversity, restricting the applicability of research findings across different populations. The acquisition of an extensive and diverse donor dataset from Cryolife Inc., a company specialising in aortic health, presents an unprecedented opportunity to fill this gap in cardiovascular research. Encompassing a wide range of AV diameter measurements across a diverse population allows one to explore patterns and discrepancies that were previously unobservable.

To ensure the accuracy of our findings, potential discrepancies between physical measurements of AV diameter and those obtained via echocardiography – a common non-invasive technique used to assess heart valves – were addressed. To this end, a supplementary dataset consisting of echocardiographic measurements was compiled at the University Medical Center Groningen (UMCG). Given the challenges in collecting echocardiographic data, only about 60 measurements from a less diverse adult population were available. The findings could not be extrapolated to younger demographics. Therefore, regression equations from existing literature concerning individuals under 18 were to be utilised to assess systematic biases in younger populations. Towards the end of the research conducted for this thesis, a new echocardiographic dataset became available. Specifically, the Lopez *et al.* dataset which only encompasses individuals 18 and under. This allowed for retroactively validating the previously applied method, as well as a more robust analysis of the bias between the physical and echocardiographic measurements for those 18 and under. These two echo datasets were employed to create a bridging model between the physical and echo AV diameter measurements.

The research begins with the data description, exploration, and preprocessing of all three datasets: the Cryolife donor dataset, the UMCG data, and the Lopez *et al.* dataset. This is followed by a visual exploration of the potential differences between the physical and echocardiographic AV diameter measurements. During the donor data and model exploration, a trend was observed in the AV diameter measurements over time and the residuals over time. With the aim of better generalisability, this was corrected via the implementation of a bespoke changepoint algorithm.

The subsequent sections – Theory, Methodology, and Results – are each divided into subsections that cover: modeling and evaluation, using body surface area (BSA) as a variable rather than height and weight, correcting trends in the AV diameter measurements, and addressing the potential echocardiographic biases. The Theory section introduces the modelling frameworks as well as the research questions. Moreover it covers the rationale behind the choices made, based on theoretical considerations and expected outcomes. The Methodology section details the practical implementations and parameter choices. The Results section presents the findings, discussing where these findings match and do not match initial expectations, as well as the implications.

The significance of this research lies in its potential to provide a more accurate, validated model of aortic valve measurements that can be utilised in diverse clinical settings. This could lead to improved diagnostic and surgical outcomes for patients with aortic valve anomalies, particularly in cases where existing models fall short. Moreover, a detailed and robust analysis of the uncertainty, not just accuracy, accompanying healthy AV diameter predictions has not yet been researched. This thesis also includes a flexible implementation of the segment neighbourhood algorithm, allowing for practically any cost function and penalty choice.

By bridging the gap in knowledge with a comprehensive analysis of a uniquely large and diverse dataset, this thesis contributes to the broader field of cardiovascular research, offering insights that could inform future studies and medical practices.

## 2 Motivation

Aortic stenosis (AS) is a severe cardiovascular condition that affects a substantial number of individuals. For example, in the United States up to 1.5 million people suffer from AS and approximately 500,000 are experiencing severe AS [1][2][3]. Without timely intervention in the form of aortic valve replacement (AVR), a significant portion of patients with severe AS face a bleak prognosis, with a survival rate of less than two years after the onset of symptoms[4]. Aortic stenosis is commonly caused by aortic valve disease, the prevalence of which has been steadily increasing[5]. In 2019, its prevalence was as high as 1160 cases per million persons resulting in 411 valve procedures per million for aortic valve disease in the United States in 2019[6]. Notably, the majority of these cases present in patients at older ages i.e. well into adulthood.

The choice of the appropriate aortic valve replacement is paramount in ensuring successful outcomes for these patients. Incorrect expectations about a patient's prosthetic valve can lead to patient prosthesis mismatch (PPM)[7]. However, selecting the correct valve size is complex. The two primary challenges in aortic valve replacement surgery involve avoiding undersizing, which can lead to insufficient capacity, valve leakage or dislodgement, and preventing oversizing, which may result in improper expansion, reduced valve longevity, potential complications with the heart's electrical system, and damage to the surrounding area[8]. Moreover, improper valve sizing can result in aortic regurgitation, a condition where the aortic valve fails to close properly, allowing blood to flow back into the left ventricle[9].

Pediatric cardiologists customarily use nomograms and Z-scores of the aortic valve diameter, for which many models are available online[5]. However, the large number of models - based on relatively small numbers of echocardiographs - cannot be extrapolated to adults, as they were not part of the datasets analysed. There is also a lack of standardisation in this area of research[10]. Instead, adult cardiologists and cardiac surgeons use simplifications that have evolved over decades. Nomograms and Z-scores are most commonly based on BSA. However, models based on BSA are potentially less accurate for individuals in extreme cases e.g. obese people[11]. Considering the increasing obesity epidemic, this is becoming a more prevalent issue. It has been proposed to use a separate model with a different cut-off value for those with BMIs over 30, but this dichotomy only partially solves the problem as the problem is not of a dichotomous nature. Therefore, it would lead to improved insight into the relative size of aortic valves if we would have access to reliable expected AV diameters for adult patients with a wider range of BMI values, and taking the universally available variables sex, age, height, and weight into account.

To address these critical challenges and enhance the accuracy and precision of aortic valve size predictions, this project will leverage three distinct datasets. The first of which is the Cryolife dataset made up of measurements from aortic heart valves obtained from deceased donors. These valves have been measured physically, and health attributes of the donors have also been collected. Notably, this dataset is the largest of its kind and covers a wide range of characteristics, including groups that have not been previously studied. Furthermore, with such a large dataset, it becomes possible to accurately quantify the precision of the expected AV diameters. It will also be possible to investigate how the standard deviation varies with regards to the predictors. However, it is important to note that this dataset relies on physical measurements, while echocardiograms are the primary diagnostic imaging tool in clinical practice. The second and third datasets, made up of echocardiographic measurements and the same predictors, will be used to investigate whether there is a systematic, quantifiable difference between the physical and echo measurements for AV diameters in adults and those 18 and under. This ensures the models based on the donor data are applicable in a clinical setting.

# 3 Literature Review of the Current Models

From four recent papers, this literature review summarises the final models found to estimate the size of the aortic annulus (also known as the aortic valve diameter). It evaluates the strengths and limitations of each paper, and underscores the necessity for an inclusive model that reduces uncertainty and extends across a broader demographic spectrum, including a wider age and weight range. Some quick characteristics can be summarised in a table as follows:

| **Model** | **Published** | **Sample size** | **Age (years)** | **Predictor** |
|---|---|---|---|---|
| Mahgerefteh *et al.* | 2021 | 3,215 | 2 – 18 | BSA or Height |
| Lopez *et al.* | 2017 | 3,566 | 0 – 17.99 | BSA |
| Cantinotti *et al.* | 2016 | 1,151 | 0 – 17 | BSA |
| Pettersen *et al.* | 2008 | 782 | 0 – 18 | BSA |

Table 1: Summary of papers

One immediate limitation of all of these papers is that the data used only spans individuals under 18. While the most significant change in valve sizes occur between 0 and 18, it has not been well-studied how AV diameter potentially changes in adulthood. Furthermore, all the papers have rather small sample sizes. Obtaining a large number of echocardiographic measurements from healthy people over a broad demographic is challenging, thus the papers have been limited on the amount of data they can base their models on. The papers also only consider one parameter: an estimate of body surface are (BSA). With the exception of Mahgerefteh *et al.* who also include a model based solely on height. This is not without reason, Lopez *et al.* found that age, sex, and race were statistically significant but not clinically significant. While on average these variables may not lead to a clinically significant difference in predictions, what is not considered is the improvement in the *certainty* of the predictions. This becomes especially relevant for individuals in extreme subgroups.

In this field, the standard for diagnostic purposes is using Z-scores. Explicitly, these are computed as follows:

$$\text{Z-score} = \frac{f(\text{AV}_0) - \mathbb{E}[f(\text{AV})]}{\tilde{\sigma}}, \tag{1}$$

where $\text{AV}_0$ is the observed aortic valve (AV) diameter, $\tilde{\sigma}$ is the estimated standard deviation, $f$ is a transformation designed to make the residuals normally distributed, and $\mathbb{E}[f(\text{AV})]$ is the model used to predict the average, healthy AV diameter. Below are the models from each paper.

The Mahgerefteh *et al.* model is from a 2021 paper which builds on the work done by Lopez *et al.* in 2017. The same model structure is considered with slightly different results when using BSA as a parameter, and Mahgerefteh *et al.* also introduces and compares the validity of using only height as a parameter.

**Model 1** (Mahgerefteh *et al.*[12])**.**

$$\mathbb{E}\left[\frac{\text{AV}}{\text{BSA}^{0.5}}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{AV}_i}{\text{BSA}_i^{0.5}}\right) = 1.50 \quad \Longleftrightarrow \quad \mathbb{E}\left[\text{AV}\right] \approx 1.50 \cdot \text{BSA}^{0.5} \tag{2}$$

$$\mathbb{E}\left[\frac{\text{AV}}{\text{Height}_i}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{AV}_i}{\text{Height}}\right) = 1.17 \quad \Longleftrightarrow \quad \mathbb{E}\left[\text{AV}\right] \approx 1.17 \cdot \text{Height} \tag{3}$$

**Model 2** (Lopez *et al.*[13])**.**

$$\mathbb{E}\left[\frac{\text{AV}}{\text{BSA}^{0.5}}\right] \approx \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{AV}_i}{\text{BSA}_i^{0.5}}\right) = 1.48 \quad \Longleftrightarrow \quad \mathbb{E}\left[\text{AV}\right] \approx 1.48 \cdot \text{BSA}^{0.5} \tag{4}$$

These model structures are somewhat unusual as they require you to first transform the outcome variable using the independent variable (e.g., dividing by $\text{BSA}^{0.5}$). This was done as a normalising transformation for the Z-score computation. Then, the mean of what is referred to as the "indexed

parameter" is the expected value. For later work, we would like to obtain a model with AV as the outcome variable. Thus, it has been rearranged for $\mathbb{E}[\text{AV}]$, where the AV diameter is given in centimetres. However, it is important to note that while it does resemble a simple linear regression without an intercept, these models are not equivalent to that. In this case the coefficient estimate for $\sqrt{\text{BSA}}$ is approximated via the mean of the AV diameter measurements divided by the respective transformed BSA measurement. This is not equivalent to estimating the same coefficient via a linear regression, though it is a close approximation. For both Lopez and Mahgerefteh, the standard deviation is simply estimated from the indexed parameter.

As previously mentioned, Lopez *et al.* investigated whether including only BSA as a parameter was sufficient. They report no clinically significant difference when including age, sex, and race. Mahgerefteh *et al.* built on this by investigating whether the further simplification of only including height would lead to a sufficiently accurate model. However, since both papers have datasets that include healthy, non-obese children it makes intuitive sense that only height could provide sufficiently accurate models when compared to BSA.

**Model 3** (Cantinotti *et al.*[14])**.**

$$\mathbb{E}[ln(\text{AV})] = 2.750 + 0.515 \cdot ln(\text{BSA}) \tag{5}$$

Cantinotti *et al.* uses a different approach for the model structure. Both the measurements and the predictor BSA are transformed using a log transformation, and then a simple linear regression is used. The AV diameter predictions – after transforming them back to the original scale – are given in millimetres. Unlike the ethnically diverse datasets used in Lopez *et al.* and Mahgerefteh *et al.*, this dataset only includes Caucasian Italians. Considering that Lopez *et al.* found race to be statistically significant, this could have led to some bias in the final model. To estimate the standard deviation to be used in the Z-score, the standard estimate of the error (SEE) is used. This is computed using the root mean squared error (RMSE) and they obtained a value of 0.088.

**Model 4** (Pettersen *et al.*[15])**.**

$$\mathbb{E}[ln(\text{AV})] = -0.874 + 2.708 \cdot \text{BSA} - 1.841 \cdot \text{BSA}^2 + 0.452 \cdot \text{BSA}^3 \tag{6}$$

Lastly, Pettersen *et al.* also log transformed the measurements, however they used a non-linear modeling approach. In contrast to a simple linear model, they have applied a generalised additive model (GAM). The AV diameter predictions are given in centimetres. A very good fit was achieved, with an R squared of 0.94. However, this is specific to their data which is made up of quite a small sample size comparatively. To compute the Z-score, the RMSE is also used – as was done by Cantiontti *et al.* – and they obtained a value of 0.214.

Despite the limitations of their datasets, these models allow for a great starting point and can be used to explore whether there may be a systematic difference in physical measurements and echocardiographic measurements for those under 18. On average, their models have accurate results for the demographics studied. Thus, it is not unreasonable to consider using these models to investigate a potential bias between echocardiographic aortic valve diameter measurements and the physical measurements from the donor dataset.

# 4 Data Description, Exploration, and Preprocessing

Each subsection covers one dataset and follows the same structure:

1. A **Data description**
2. Using **histograms and bar charts** to visualise how the raw data is distributed
3. **Density plots** split by the binary variable of interest
4. Investigating **Outliers**
5. **Correlation matrices** for a concise overview
6. **Simple linear plots** of the outcome variable versus the independent variables alongside linearisation transformations
7. **LOESS (locally estimated scatterplot smoothing) fits** to explore whether nonlinear relationships may be present
8. Other notable finds

## 4.1 Donor Data

The dataset originates from the database of Cryolife Inc, a US company, that measured, processed, and sold aortic homografts for decades, containing observations of 75,142 subjects and recorded nine features. The outcome variable of interest is the aortic valve (AV) diameter and since it has 2,366 missing values, this brings the dataset to 72,776 observations. The dataset also includes 'PV diameter', this is the diameter of the pulmonary valve and is not used. The donor characteristics recorded are age, sex, weight, and height. The body surface area (BSA) has been computed according to the Haycock formula (eq. 12), this feature is discussed in Section 5.3.2. Age has been recorded in years and covers a large range, between 0 and 59. Notice the large range in weight and height as well.

| Feature | Units | Number of NAs | Mean | Median | Range |
|---|---|---|---|---|---|
| AV diameter | mm | 2,366 | 21.47 | 22.50 | 3.50 – 36.50 |
| PV diameter | mm | 24,156 | 23.86 | 25.50 | 3.50 – 35.50 |
| Age | Years | 0 | 34.28 | 39.00 | 0 – 59 |
| Sex | Boolean | 0 | – | – | 0,1 |
| Weight | kg | 0 | 78.98 | 79.55 | 1.34 – 352.27 |
| Height | cm | 0 | 164.17 | 172.72 | 30.48 – 236.22 |
| BSA (Haycock) | $m^2$ | 0 | 1.88 | 1.98 | 0.12 – 4.43 |
| Dissection Date | Date | 0 | – | – | 1985/06/05 – 2016/09/27 |

Table 2: Summary of dataset features.

To begin the data exploration, histograms and bar charts were created to visualise how the data is distributed, which are included in the appendix in Figure 41. The AV measurements, height, weight, and BSA seem to be bimodally distributed, with one large peak and a smaller peak at the lower values, this can be explained since one can also see a large spike in the age histogram at age zero i.e. this is related to distinct subgroups within the data (adults and children). The age histogram itself shows three peaks: a pronounced one at zero years, with other notable peaks around ages 20 and 50. From the bar chart, it can be seen that there is approximately double the number of male observations compared to female (32% female and 68% male). Given the large size of the dataset, this imbalance will not be a problem. As can be seen in the density plots split by sex and overlayed, the data is similarly representative for both males and females.

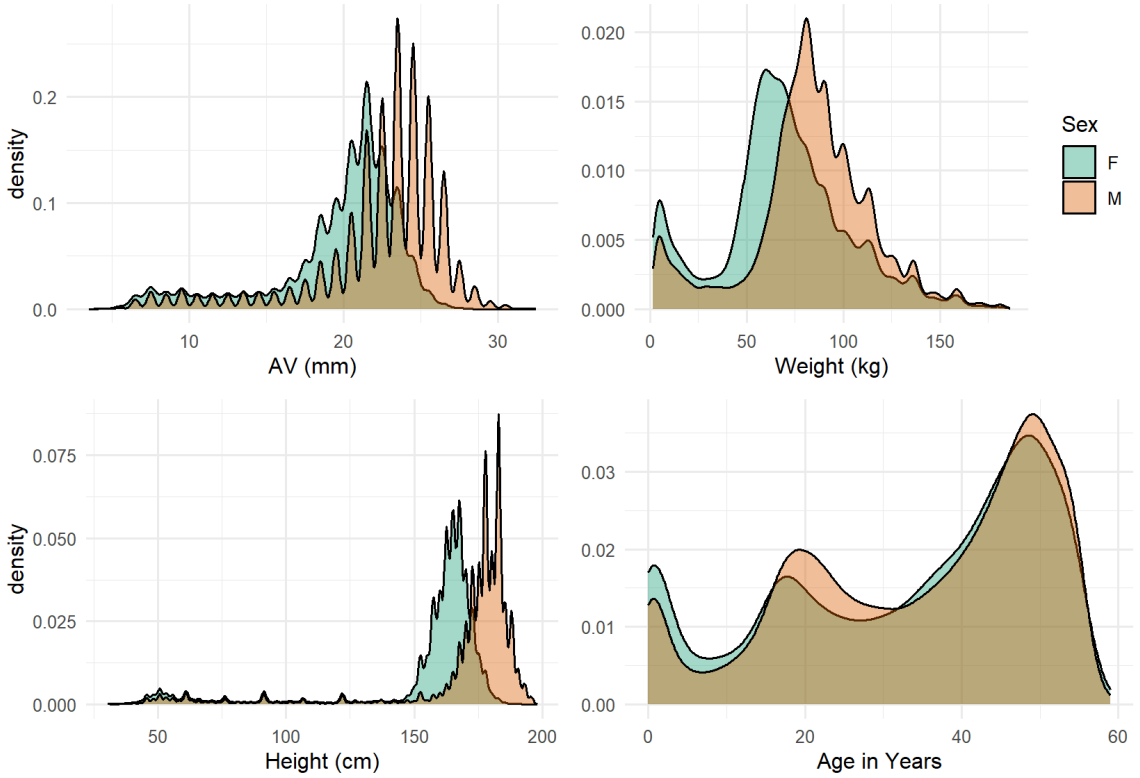



Figure 1: Density plots split by sex for the donor data.

Note that the densities in Figure 1 have evenly spaced spikes, these are most evident in the male AV diameter measurements. These spikes are due to the fact that the original measurements were taken in imperial units and converted to metric units i.e. from feet, inches, and pounds to centimetres, millimetres, and kilograms.

The independent variables not being normally distributed is not an assumption or problem for generalised linear models (GLMs) or generalised additive models (GAMs). However, there is an assumption that the error component is from the exponential family e.g. normally distributed. This assumption is typically addressed by ensuring that the response variable follows an appropriate distribution from the exponential family, possibly by selecting an appropriate link function. This is one method to more closely adhere to the normality assumption of the residuals. In our case, the response variable is not very normally distributed, nor does it seem to follow any other distribution from the exponential family. After a log transformation, there is an improvement in the bimodality, potentially leading to residuals that better meet the normality assumption. This can be seen in Figure 2 below. Furthermore, the bimodality in the outcome variable looks to be reflected and may well be captured via by the - possibly nonlinear - patterns and interactions in the data. This may also help result in an error component that satisfies the necessary assumption for the residuals. Ultimately, after fitting the models, we will assess this assumption and the overall model fit through residual analysis to confirm. Moreover, while this assumption aids in robust statistical analysis, it is most important when inference is the goal. This assumption supports the validity of standard errors, confidence intervals, t-tests, etc. derived from the model. Since prediction with high accuracy and certainty is the aim, this assumption is of less concern. In this case, generalisability to unseen data is more critical.

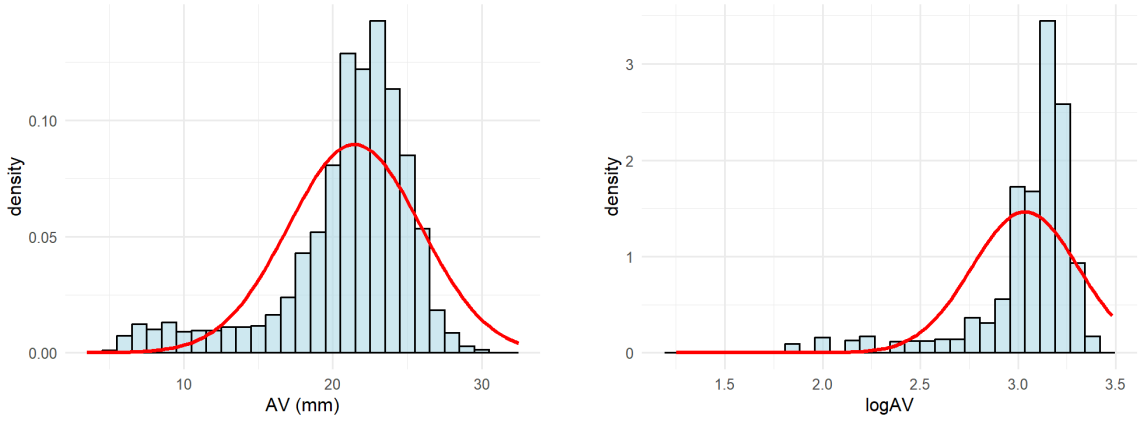



Figure 2: Histograms showing before and after applying a log transformation on the outcome variable AV. The normal curve is given in red.

With regards to outliers, the approach used was to only dispose of theoretically implausible observations, those that were likely inaccurate measurements. Notice that in the height and weight histograms, there are some observations of extremely high values. To further investigate possible outliers, age versus AV, height, and weight were plotted. For a more in-depth look, these plots were also faceted by sex.

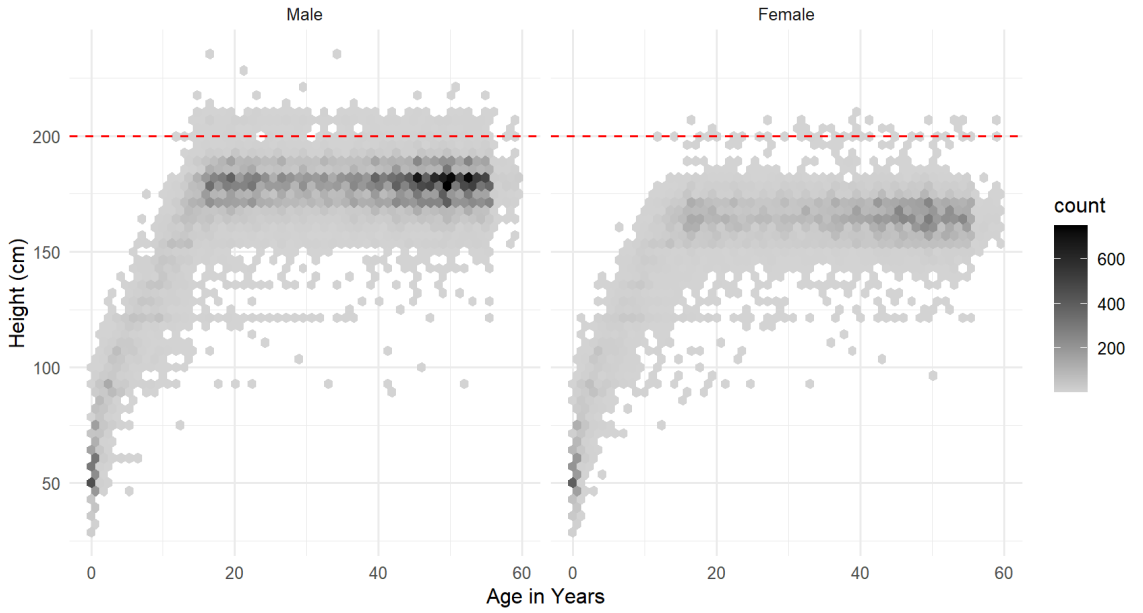



Figure 3: **Raw data:** Height versus Age faceted by sex with a horizontal, dashed red line at 200cm. Hex bins are used due to the huge amount of data.

In Figure 3, notice that for both males and females there are observations that clump around a horizontal line at a height of 200cm. These look disconnected from how the rest of the data is distributed. Moreover, according to the CDC's 2015-16 NHANES survey, heights at 198cm are in the 99.84th percentile for males and 100th percentile for females[16]. Whereas for our data they are in the 98.89 and 99.66 percentiles for males and females respectively, this is far too many.

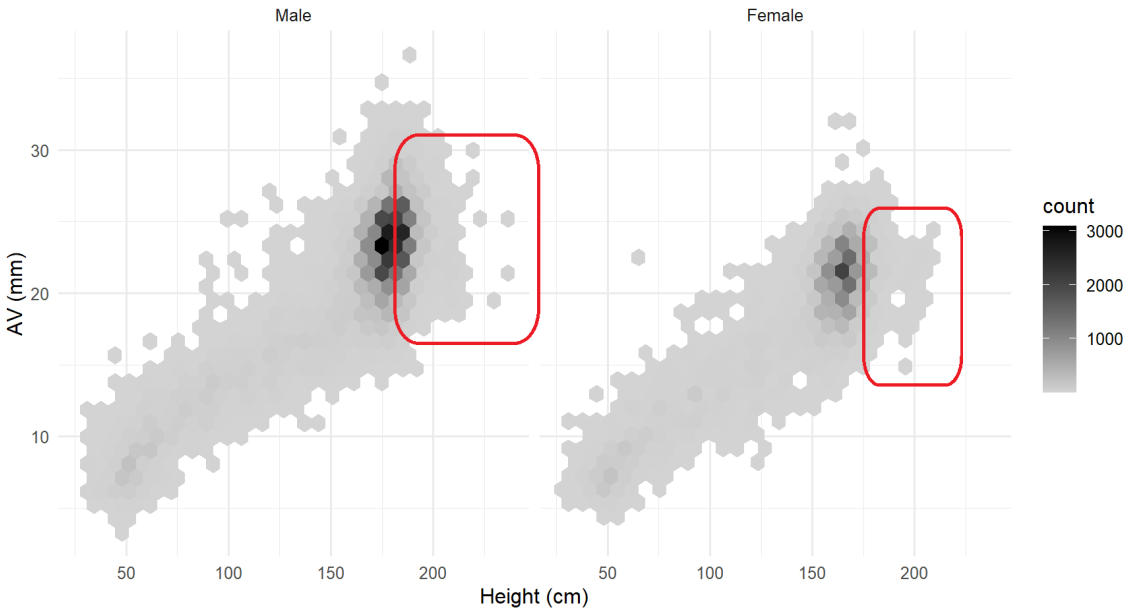



Figure 4: **Raw data:** Height versus AV faceted by sex. In the red box are the points that demonstrate strange behaviour.

Furthermore, in Figure 4, when plotting the AV diameter versus height it can be seen that these heights at 200cm exhibit a counter-intuitive behaviour with respect to the AV diameter. At these larger heights, the observations are clumping at smaller AV diameter values (highlighted in the Figure using a red box). This goes against the general shape of the data and would imply that, at a height of around 200cm and beyond, the AV diameter decreases. From this exploration and the CDC survey, it was decided that heights above 198cm were to be removed as these observations - at least the vast majority of them - are likely inaccurate. Using the height-for-age charts from the WHO[17], the implausible observations for under 5s were also discarded e.g. a 2 year old with a height above 125cm. Since one of the goals is to obtain a model applicable for a diverse population, only the observations with very extreme weights (outside of the 99th percentile) were dropped.

To explore which predictors are promising and to get a concise overview, correlation matrices can be useful. The correlation matrices split by sex can be found in the appendix. For this data all the predictors have high correlation coefficients with the response variable AV. Between the predictors the correlation coefficients are also very high, however multicollinearity is also not a concern for prediction. To take a closer look at the relationships between the predictors and the outcome variable, simple linear plots can be used as is demonstrated in Figure 5. Additionally, for GLMs, there is an assumption that the relationships between the outcome variable and the predictors are linear. These plots can also be used to check whether that assumption is satisfied.

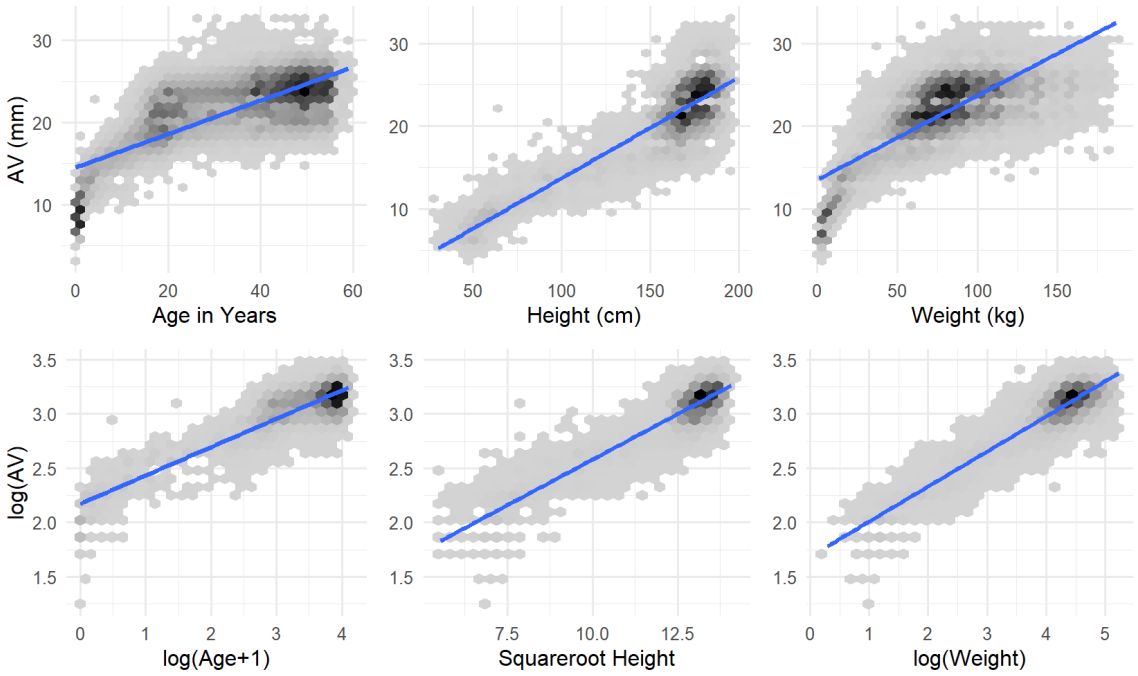



Figure 5: Linearity transformations for the donor data. The plots on the top are before the chosen transformations are applied and the plots on the bottom are after.

There is certainly an improvement in the linearity of the relationships, this was also checked via fitting simple linear models and checking the goodness-of-fit metrics. Moreover, simplifying the relationship between the predictors and outcome variable can also help reduce the complexity for nonlinear modelling approaches. After applying these transformations, LOESS fits can be used to see if there is any evidence to support whether nonlinear relationships may still be better suited.

When applying separate LOESS models by sex, as seen in Figure 6, a divergence for larger values is captured for all three predictors. For small values, the fit is also seemingly less skewed in all three plots. LOESS allows for very flexible fits, and despite this we get very smooth results. This is a positive sign as it shows that they are not overfitting. There is definitely evidence to support a nonlinear modeling approach as, compared to the linear models, the LOESS results appear to be more appropriate and seem to capture nuances that the linear models do not.

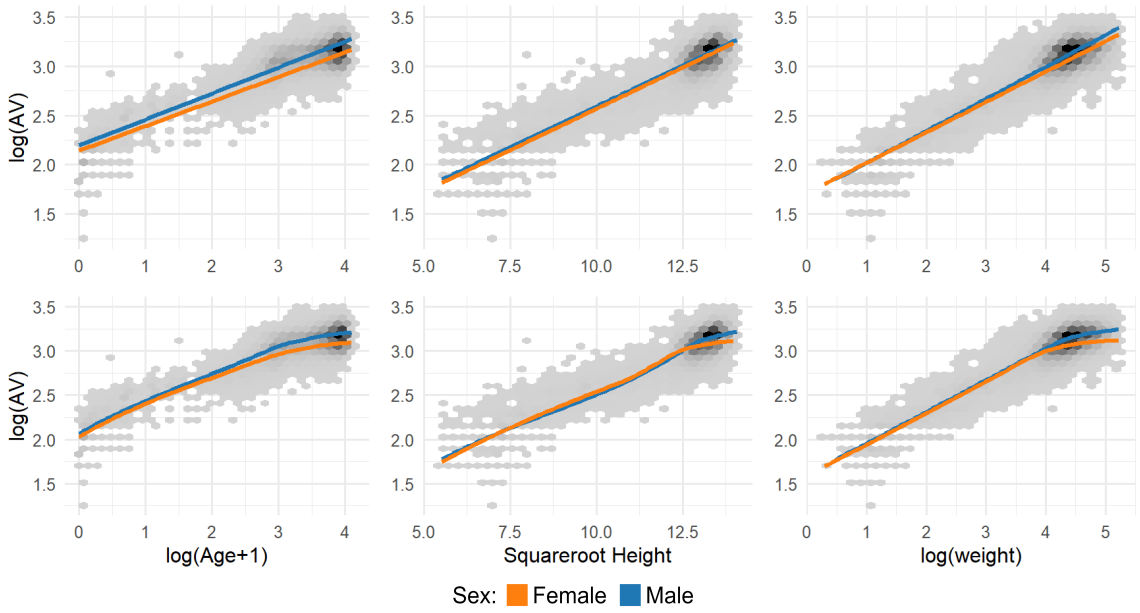



Figure 6: **Donor data:** separate models are applied based on sex, overlayed on the entire dataset. On the top are linear fits and on the bottom are LOESS fits.

The final notable find during the exploration is a trend in the AV diameter measurements over time, shown in Figure 7. Most obvious is a seemingly negative linear trend around the years 1995 and 2004. However, further trends and changepoints could be identified via an objective analysis. When visualising the predictors over time, there were no indications of similar trends (see Figure 43 in the appendix). For further analysis, once models have been applied, residuals can be used to better investigate this finding. Why changepoints are identified in the residuals and not the AV diameter measurements themselves is explained in Section 5.3.3. It is also important to note that there is very limited data before the year 1995, which gets more and more sparse further back in time.

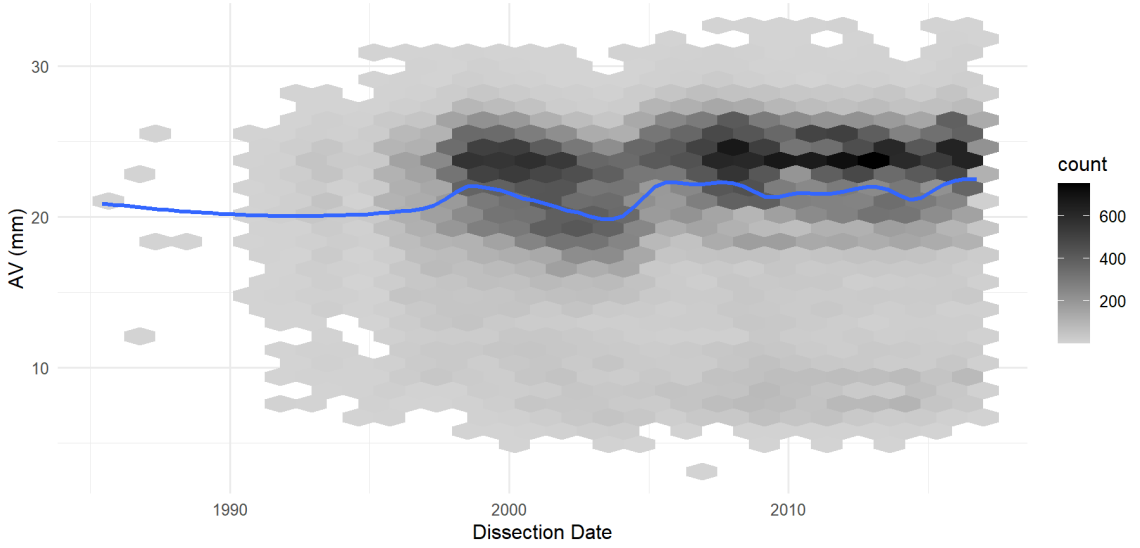



Figure 7: A LOESS model applied to the AV measurements over time.

## 4.2 Echocardiographic Data

### 4.2.1 UMCG data: Over 18s

The Echocardiographic data has been collected at the UMCG with the intention to investigate the difference between the physical AV diameter measurements obtained from the donor data and echocardiographic estimations. Therefore, the same clinical features are present. There are 85 observations, however once NAs are removed there are 68 observations.

| Feature | Units | Number of NAs | Mean | Median | Range |
|---|---|---|---|---|---|
| Mean AV diameter | mm | 4 | 21.00 | 21.18 | 13.80 – 25.57 |
| Age | Years | 0 | 44.57 | 42.50 | 18 – 80 |
| Sex | Boolean | 0 | – | – | M, F |
| Weight | kg | 1 | 77.86 | 76.00 | 53 – 130 |
| Height | cm | 12 | 176.7 | 176.0 | 159 – 199 |
| BSA (Haycock) | $m^2$ | 12 | 1.95 | 1.93 | 1.54 – 2.64 |

Table 3: Summary of dataset features.

The mean AV diameter is the average of three measurements. Four echocardiograms were found to be inadequate and not suitable for measurements of the aortic annulus diameter. The age range only covers adults, thus this dataset can only be used to compare the difference in measurements for adults. As was done for the donor data, histograms were created as a starting point. The histograms are seen in the appendix, Figure 44. For this data, the variables are somewhat normally distributed. Compared to the donor data, these histograms are less skewed and they do not seem to be obviously bimodal. However, since there is not a large amount of data, the overall shape of the histograms are more susceptible to fluctuations caused by the random nature of sample collection. This can result in a histogram that does not clearly resemble a normal distribution, even if the data are from a normally distributed population. For the outcome variable AV, the Shapiro-Wilks test gives a p-value of 0.133. Thus, there is not enough evidence to say that AV is not normally distributed. Lastly, it can be seen that the number of observations for males and females are approximately equal. The density plots disaggregated by sex are not very informative, due to the small number of observations overall. They are included in the appendix in Figure 45 and show nothing of concern. Throughout the exploration, no outliers were found i.e. no seemingly inaccurate measurements.

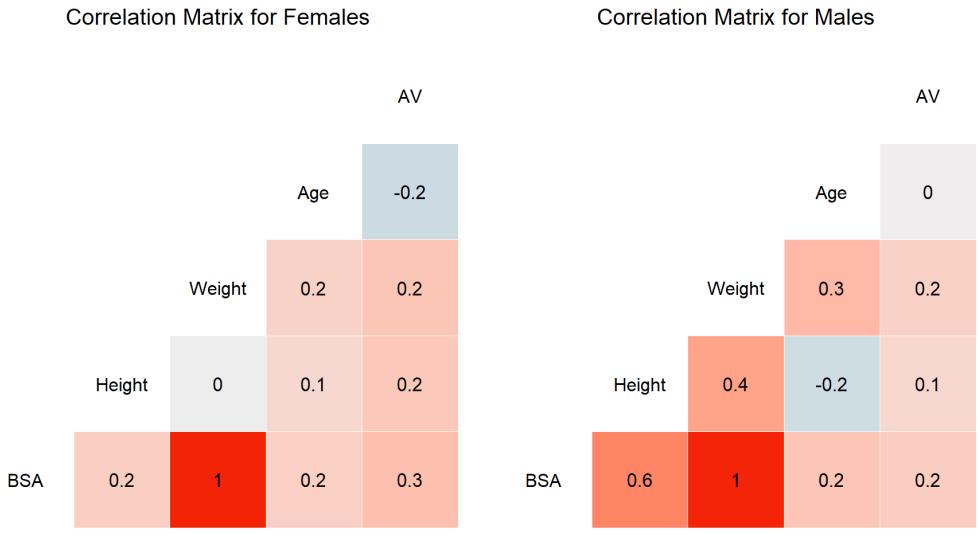



Figure 8: Correlation matrices for the continuous variables for the adult echocardiographic measurements split by sex.

When looking at the relationships between AV and the predictors via correlation matrices in Figure 8, there was an interesting correlation found. Notably, exclusively for females there is a negative correlation between AV diameter and age. This can also be seen visually in Figure 9, when looking at the linear models in the AV versus age plot. For males, the linear fit looks practically flat whereas for females there is a negative trend. According to expert opinion, when visibility is low, more conservative estimates of the AV diameter are taken. Thus, this negative correlation could, for instance, be due to the breast tissue around the chest making it more challenging to get clear echocardiograms.

Various transformations were applied to the variables and no improvement (or worsening) in linearity was found, Figure 46 demonstrating this can be seen in the appendix. Goodness-of-fit metrics also agree with what can be seen visually.

The LOESS fit plots in Figure 9 look overly sensitive, which suggests that there is overfitting. These irregular patterns indicate that a flexible, nonlinear approach is fitting noise rather than capturing true relationships, leading to potentially misleading interpretations. This is likely due to the fact that this is a small dataset. In such cases, linear models are often more effective. Linear models tend to generalise better in scenarios with limited data, providing a more robust and reliable analysis. Furthermore, the linear models in Figure 9 show that, when split by sex, there is more than a simple shift in difference in the AV diameter measurements between males and females. Only when looking at AV versus weight does it resemble a simple change in the intercept.

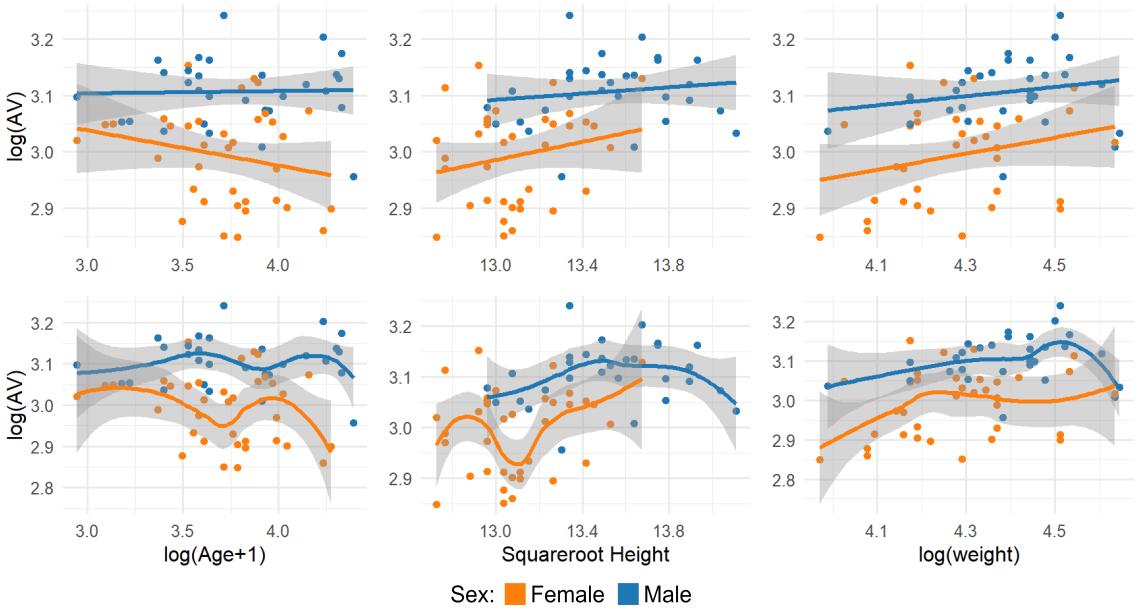



Figure 9: **UMCG echo data:** separate models are applied based on sex, overlayed on the entire dataset. The observations are also colour coded. On the top are linear fits and on the bottom are LOESS fits.

### 4.2.2 Lopez *et al.* Data: Under 18s

Towards the end of my project, the Lopez *et al.* dataset was acquired, which recently became publicly available. This dataset comprises 3,566 observations, offering a valuable resource for an analysis of a potential bias between physical and echocardiographic measurements. Below is Table 4, giving an overview of the variables included in the dataset.

| Feature | Units | Number of NAs | Mean | Median | Range |
|---|---|---|---|---|---|
| AV | mm | 336 | 1.66 | 1.70 | 0.00 – 3.40 |
| Age | Years | 0 | 7.60 | 6.36 | 0.00 – 17.99 |
| Sex | Boolean | 0 | – | – | 0, 1 |
| Weight | kg | 1 | 29.76 | 22.00 | 2.00 – 94.00 |
| Height | cm | 0 | 116.3 | 118.6 | 43.2 – 207.2 |
| BSA (Haycock) | $m^2$ | 1 | 0.95 | 0.85 | 0.16 – 2.21 |

Table 4: Summary of Lopez dataset. Only the relevant features from their dataset(s) have been included

In Figure 47, the histogram for each variable has three peaks except for the AV diameter. The AV looks somewhat bimodal, also after various transformations. Similarly as for the donor data, this is not a big concern. In the bar chart, it can be seen that the number of observations for males and females are approximately equal.

When looking at the density plots disaggregated by sex in Figure 10, notice that there is much more overlap compared to the donor data and the adult echo data. This makes sense since the data only spans those under 18, a more significant difference in characteristics would only be expected if the data would cover individuals into adulthood.

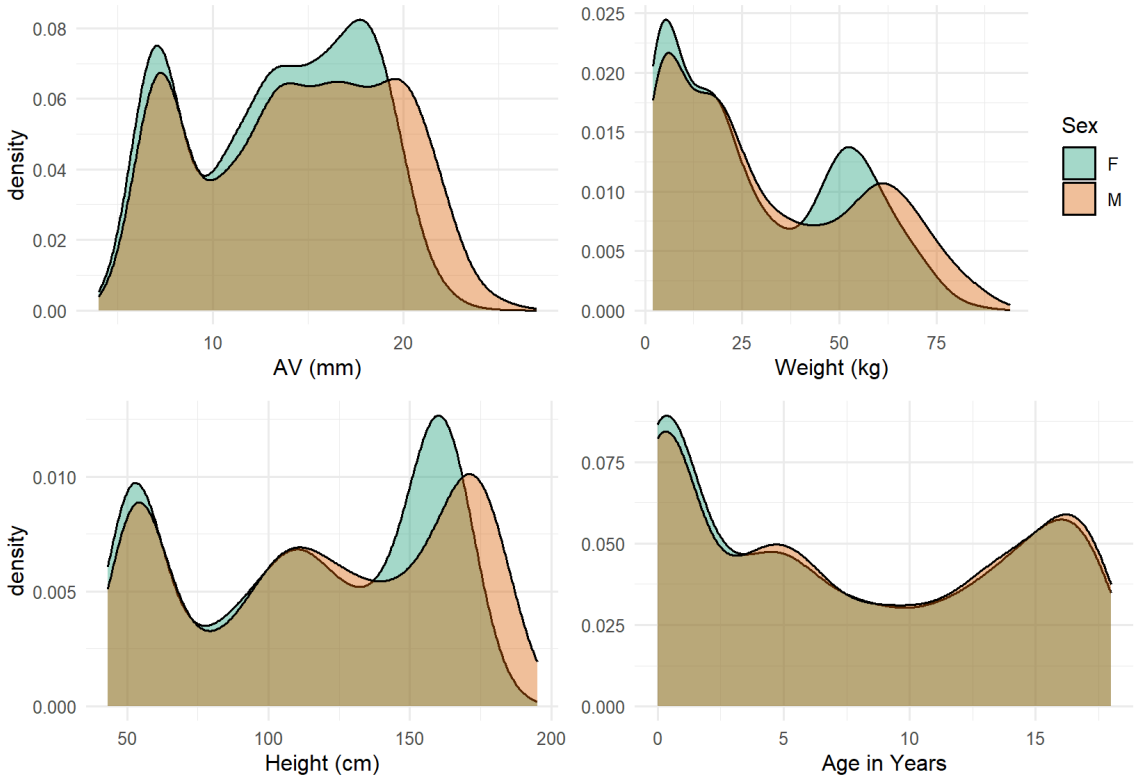



Figure 10: Density plots split by sex for the Lopez *et al.* echo data.

The data was examined for outliers through the same methods described previously. There were no observations that deviated significantly from the overall pattern of the dataset, nor did any observation have any indication of being an implausible or inaccurate measurement. Thus, no observations had to be discarded.

The correlation matrices in Figure 48 (in the appendix) show very high correlation for all the variables, both with the outcome variable AV and between each other. There are no notable differences in the correlation coefficients when split by sex.

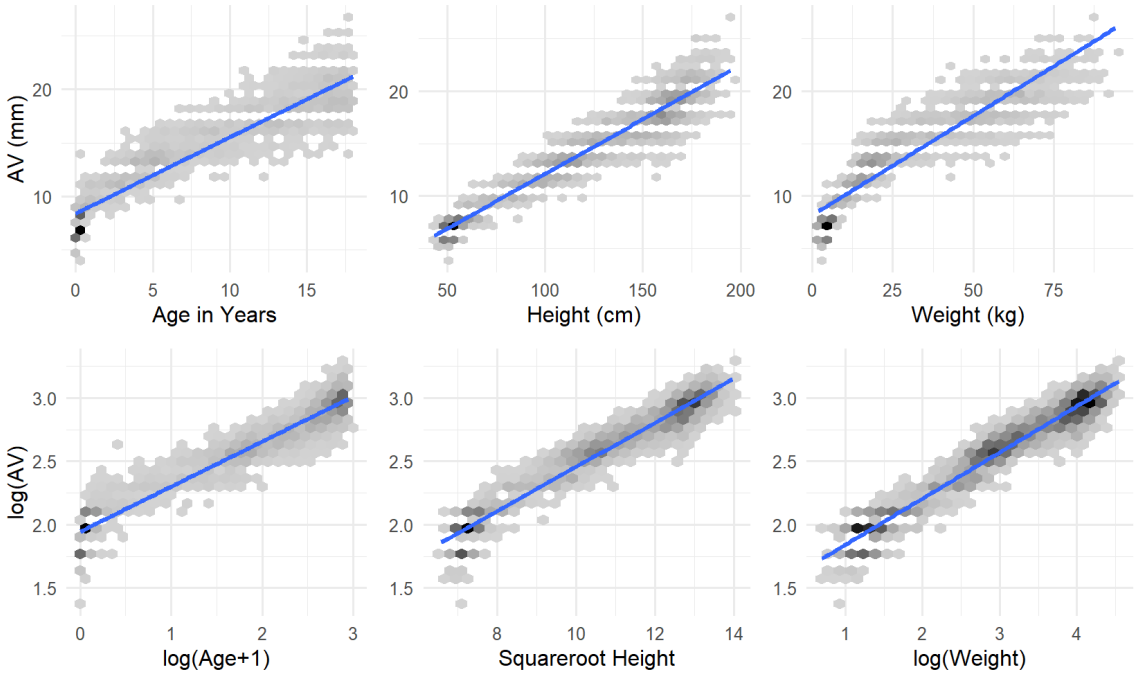



Figure 11: Linearity transformations for the Lopez *et al.* data. The plots on the top are before the chosen transformations are applied and the plots on the bottom are after.

When applying linearity transformations, seen in Figure 11, the most obvious improvement in linearity is seen for the weight variable. However, there is also an improvement in the relationship with age. Specifically, the downward curve seen for lower ages is not well-captured before the transformations. For height, there is a slight upwards curve in the pattern of the data that is not captured. After the transformations, the larger heights are better fit with a linear model but there seems to be a downward curve introduced that it is not fitting so well. Overall, the transformations applied have resulted in relationships closer to linear between the predictors and the outcome variable.

As before, to check whether a nonlinear approach may still be more appropriate LOESS fits were applied in Figure 12. Similarly to the donor data, the dataset is sufficiently large, enabling the LOESS models to produce very smooth curves that do not appear to overfit. Despite being slight, there are subtleties in the data that seem to be better fit with a nonlinear approach. Take, for example, the downward curve at lower heights mentioned previously. For the weight predictor, it can be seen that the same divergence between to the two LOESS models seen in the donor data also starts to appear here. However, this divergence is less noticeable since this data does not include as large a range in weight. Given the exploratory evidence seen here, pursuing a nonlinear method (GAMs) could be justified as it poses little risk of overfitting and may offer improvements in prediction accuracy compared to a linear approach.

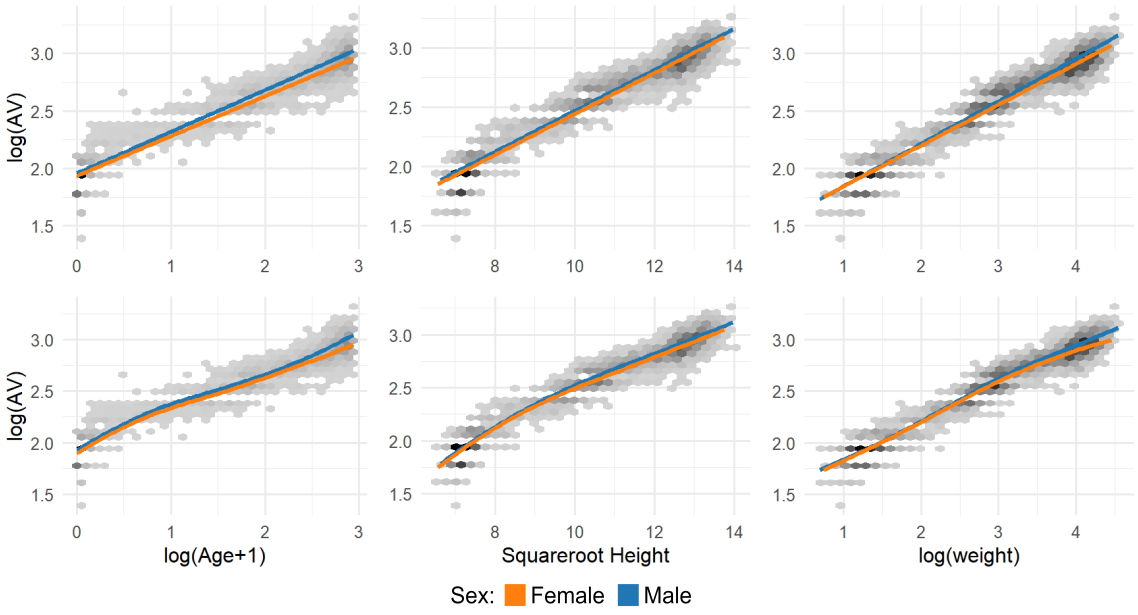



Figure 12: **Lopez** *et al.* **data:** separate models are applied based on sex, overlayed on the entire dataset. The observations are also colour coded. On the top are linear fits and on the bottom are LOESS fits.

## 4.3 The Difference Between the Donor Data and the Echo Data

To begin investigating the potential systematic differences between physical and echocardiographic measurements, a visual exploration is first conducted. It is crucial to acknowledge the following:

- The UMCG echocardiographic measurements are confined to adults
- The Lopez *et al.* echocardiographic measurements are confined to under 18s
- Previous models exclusively encompass individuals under 18 years of age

Access to the Lopez *et al.* data was only achieved close to the end of writing my thesis. Consequently, the UMCG echocardiographic data was employed to explore differences in adults, while previous models were utilised to assess differences in individuals below 18 years of age. Now, the Lopez *et al.* data has also been included in assessing the difference. This will allow for an analysis of how informative the methods and conclusions would have been based on using a model of the echo AV diameter measurements alone.

### 4.3.1 Using All the Echocardiographic data to Explore the Differences

Looking at the density plots in Figure 13, there is an indication that on average the UMCG echocardiographic measurements are smaller than the physical donor measurements. This is further supported by the similarly distributed weight, height but without the smaller peak around the smaller measurements (since the UMCG data only spans adults). Explicitly, despite the echocardiographic data including observations that are generally older, larger people, the AV measurements are smaller when compared to the donor data. However, the measurements do not seem to be smaller by a huge margin.

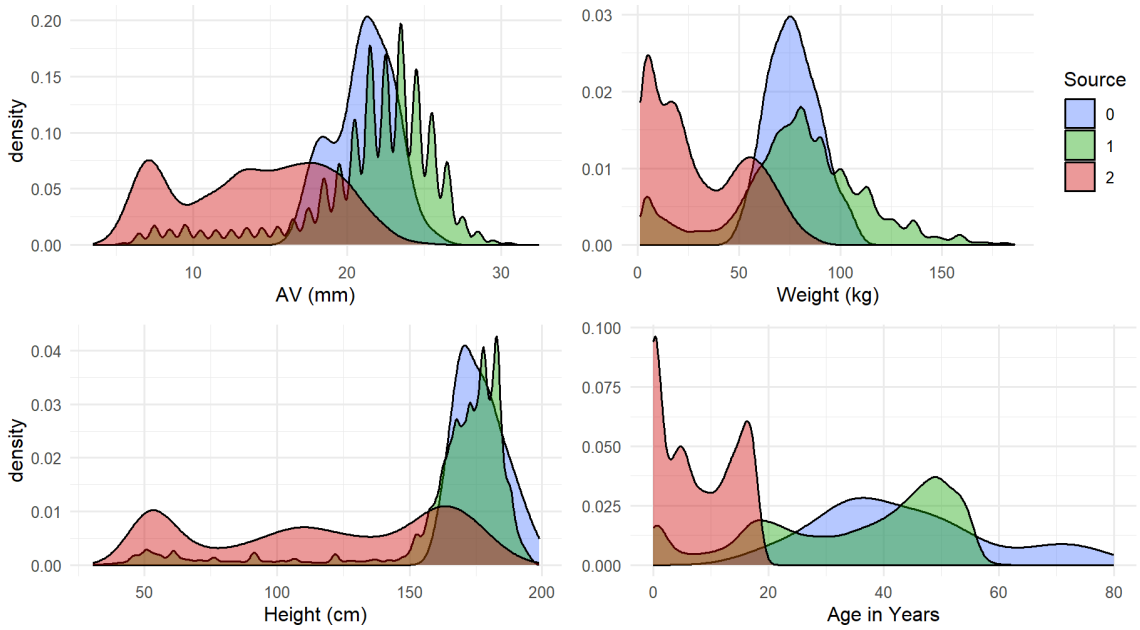



Figure 13: Density plots with the **UMCG echo data** is denoted by **0**, the **donor data** is denoted by **1**, and the **Lopez echo data** is denoted by **2**.

To further explore the distributions of the echocardiographic datasets compare to the donor dataset, boxplots are used. Subsets of the donor data are used such that the age demographics of the respective echo datasets are comparable. The boxplots in Figure 14 support that on average, the adult UMCG echocardiographic AV diameter measurements are slightly smaller than the physical donor measurements. This is despite also showing that the height, weight, and age are similarly distributed.

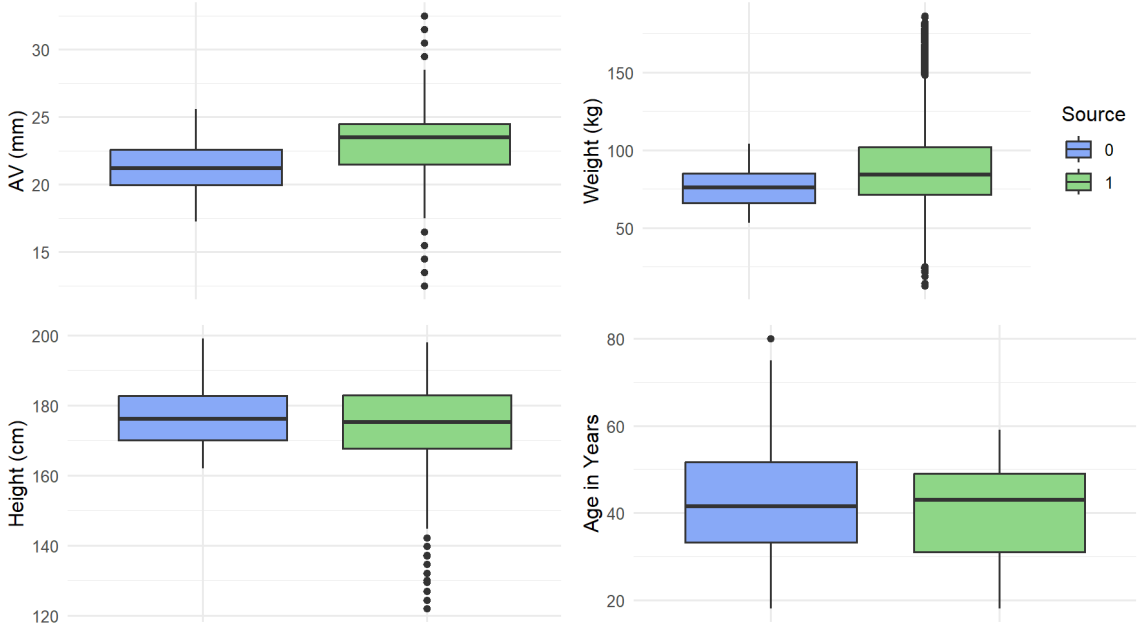



Figure 14: Boxplots comparing the variables for the **UMCG echo data** denoted by **0** and an adult subset of the **donor data** denoted by **1**.

In Figure 15, comparing box plots for the Lopez *et al.* echocardiographic measurements and the donor measurements, the median AV look approximately the same. This is along with the median and range for every predictor being similar. The only noticeable difference is that the donor data spans slightly more varied demographics, especially for weight.

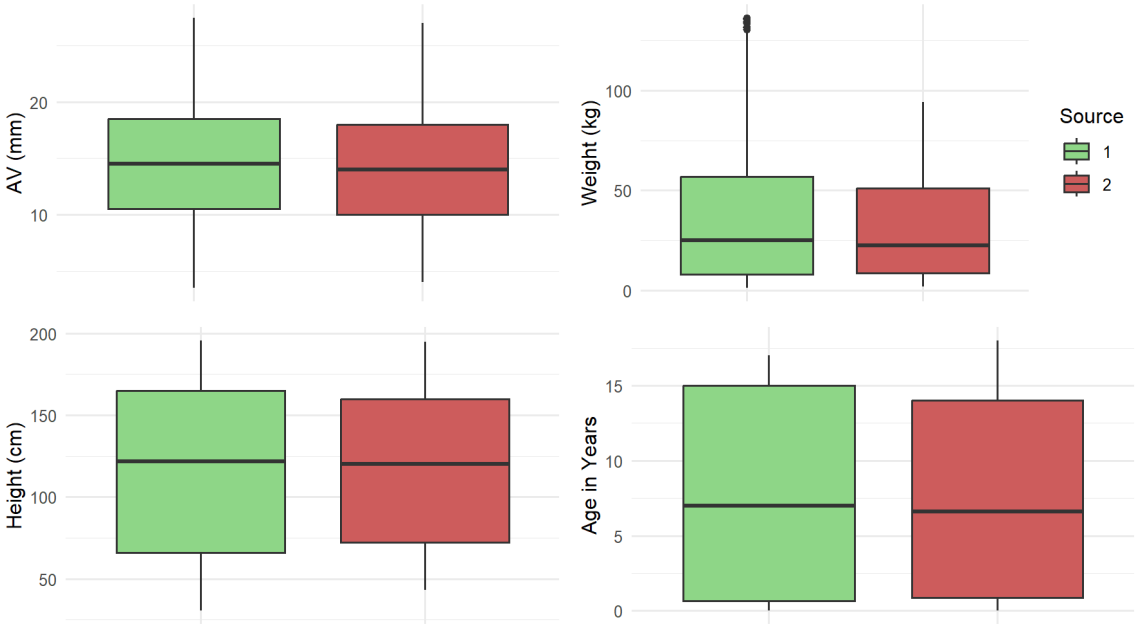



Figure 15: Boxplots comparing the variables for the **Lopez echo data** denoted by **2** and a subset of individuals 18 and under for the **donor data** denoted by **1**.

To check whether the AV diameter measurements vary similarly with the predictors, the three different datasets was plotted together and separate LOESS fits were applied. This also allows one to see whether there is evidence for a noticeable discontinuity between the UMCG and Lopez *et al.* echo datasets.

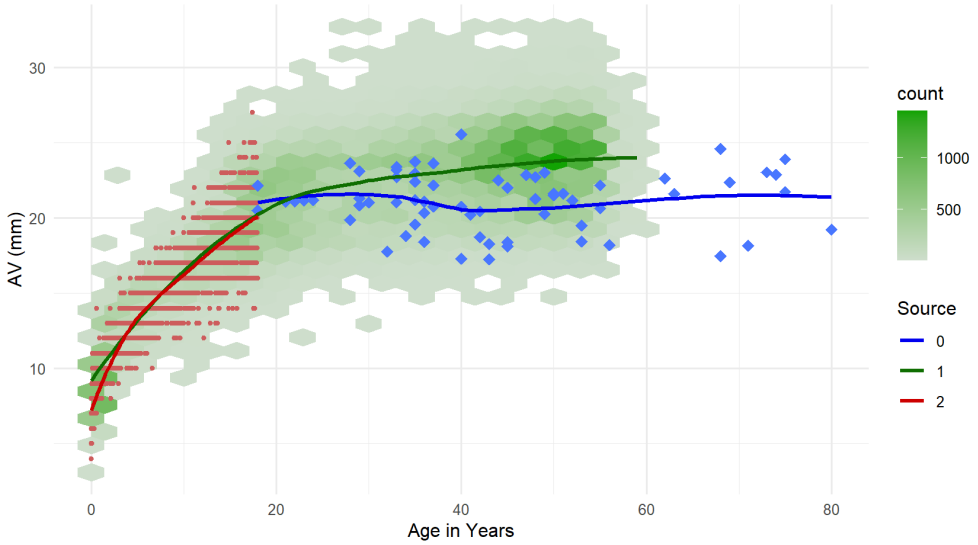



Figure 16: LOESS fits and the data for AV versus Age with: the **UMCG echo data** is denoted by **0**, the **donor data** is denoted by **1**, and the **Lopez echo data** is denoted by **2**.

Figure 16 shows the AV versus age, and this shows some very interesting results. This plot is of particular interest since we know that the three datasets span different age ranges. It can be seen that the Lopez *et al.* data and LOESS fit overlap heavily with the donor data and fit. This would be evidence pointing towards no systematic difference between the physical and echocardiographic AV diameter measurements for those 18 and under. With regards to the UMCG data, there does seem to be an indication that the AV diameter measurements are systematically lower than those of the donor data. However, the difference in the quantity of data is notable. Drawing solid conclusions based on the 68 UMCG echo observations will be somewhat dubious.

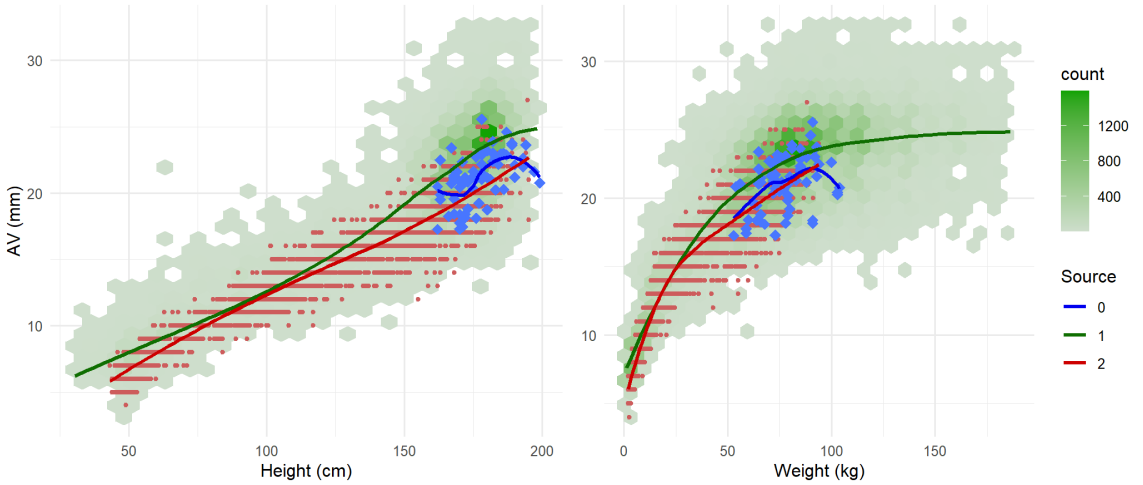
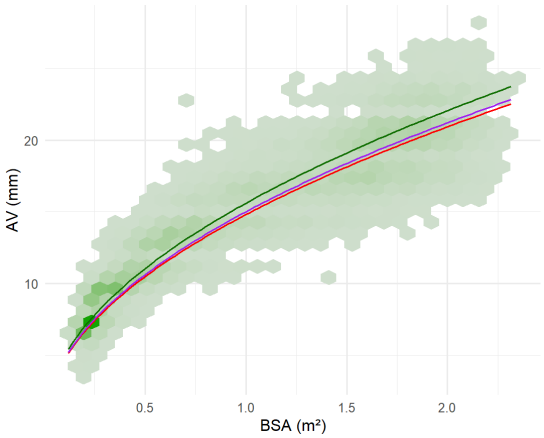



Figure 17: LOESS fits and the data for AV versus Age with: the **UMCG echo data** is denoted by **0**, the **donor data** is denoted by **1**, and the **Lopez echo data** is denoted by **2**.

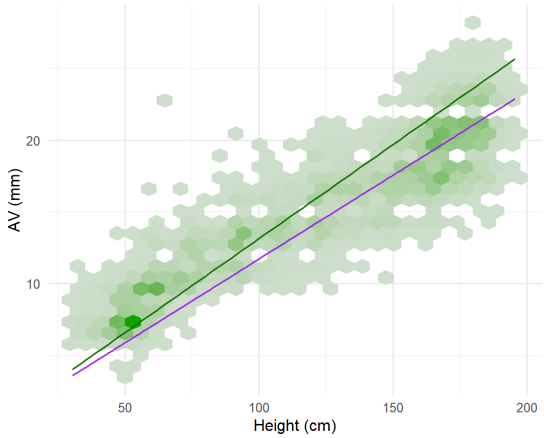
For the AV versus height in Figure 17, both the Lopez and UMCG LOESS fits are consistently below the donor data fits. LOESS fits for the UMCG echo data are not very appropriate due to the limited data, leading to curves that are not very smooth. However, the UMCG data still looks consistently below the donor data when compared to the same regions. In these two plots it can be seen that both echo datasets overlap heavily and the LOESS fit for the Lopez data looks reasonable for the UMCG data as well. Notice that the Lopez LOESS fits deviate more from the donor fits for large heights and weights, as well as very low heights and weights. The fact that their data spans a smaller demographic could be contributing towards this. Thus, when quantifying and correcting the echocardiographic bias it makes sense to ensure a subsample of the donor data is taken such that the demographics are similarly distributed.

### 4.3.2 Using Existing Models to Explore Differences in Individuals 18 and Under

The trained models, with their determined coefficients, from the literature overview are plotted along with the same model structures trained on a subset of the donor data. This gives an indication of how their chosen model structures would have differed if trained on physical measurements with similar demographics.



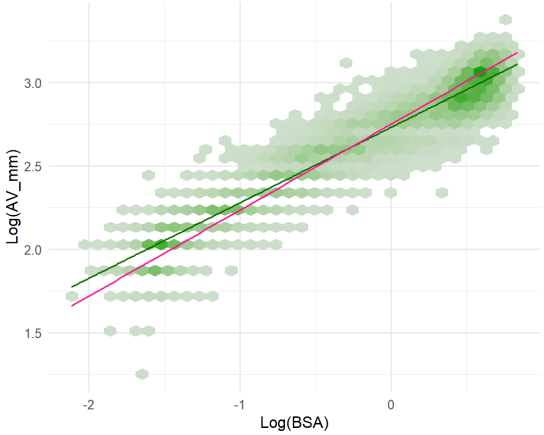

(a) The Lopez and Mahgerefteh models based on BSA (equations 4 and 2 respectively).



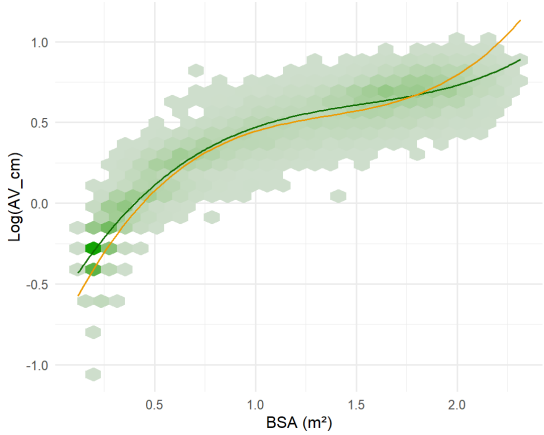

(b) The Mahgerefteh *et al.* model based on height (eq. 3).

Figure 18: In **green** are the same model structures trained on a subset of the **donor data**, in **red** is the **Lopez *et al.* model**, and in **purple** are the **Mahgerefteh *et al.*** models.

In Figure 18, it can be seen that the Lopez and Mahgerefteh models are consistently below the same model structures fit using the donor subset. This implies that the echo AV diameter measurements may be systematically smaller than the physical AV diameter measurements from the donor data.




(a) The Cantinotti *et al.* model (eq. 5).




(b) The Pettersen *et al.* model (eq. 6).

Figure 19: In **green** are the same model structures trained on a subset of the **donor data**, in **pink** is the **Cantinotti *et al.* model**, and in **orange** is the **Pettersen *et al.*** model.

Looking at the donor fits compared to the Cantinotti and Pettersen models (Figure 19), there is a crossover apparent in both plots. For the most part, both models are also below the donor fits. However, these plots are indicating that extrapolation for large BSA values is inappropriate for both of these models.

# 5 Theory

The Theory section explains the rationale behind the research by outlining and introducing the models and frameworks that underpin the study. The section also describes the research hypotheses and questions. Specifically, it covers the reasoning behind the choices made and what results could be expected based on theoretical justifications and previous findings.

## 5.1 Introduction to Generalised Linear Models and Generalised Additive Models

Throughout this entire section the following two books are referenced: "Introduction To Generalized Linear Models" by Dobson and Barnett[18] and "Generalized Additive Models: an introduction with R" by Simon N. Wood[19]. For generalised linear models (GLMs), the estimation of the regression coefficients $\boldsymbol{\beta}$ typically uses an iterative algorithm such as Iteratively Reweighted Least Squares (IRLS). To understand the update equation used to estimate $\boldsymbol{\beta}$, a few things need to be introduced.

The general form of a GLM is expressed as

$$g(\mathbb{E}[y_i]) = g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}, \qquad y_i \sim EF(\mu_i, \phi) \tag{7}$$

where $g$ is the link function, EF denotes the exponential family, $\phi$ is a scale parameter, $\mu_i$ is the expected value of the response variable $y_i$, $\eta_i$ is the linear predictor, and $\mathbf{x}_i$ is the vector of predictor variables for the $i$-th observation. In the IRLS update equation, the design matrix $\mathbf{X}$ is used. This is an $n \times p$ matrix of predictors, where $n$ is the number of observations and $p$ is the number of predictors. The working dependent variable $z_i$ is used to linearise the model and is defined as

$$z_i = \eta_i + g'(\mu_i)(y_i - \mu_i),$$

where $g'(\mu_i)$ is the derivative of the link function. Next, the weight matrix $\mathbf{W}$ is used to adjust for variability and also linearise the model fitting process. It is a diagonal matrix with elements

$$w_i = \frac{1}{g'(\mu_i)^2 \mathrm{Var}(y_i)},$$

Finally, the coefficient vector $\boldsymbol{\beta}$ is updated using the IRLS update equation:

$$\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z},$$

where $\boldsymbol{\beta}^{(k+1)}$ represents the updated estimate of the regression coefficients after the $k$-th iteration.

Generalised additive models (GAMs) extend GLMs by allowing the linear predictor to be represented as the sum of smooth functions of the predictors, rather than just a linear combination. The general form of a GAM is given by:

$$g(\mathbb{E}[y_i]) = g(\mu_i) = \eta_i = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots, \tag{8}$$

where $\eta_i$ is the additive predictor, and $f_j(\cdot)$ are smooth functions of the predictors $x_k$. The estimation of GAMs looks very similar to what was described for GLMs with a few key differences. First, each smooth function $f_j(x_j)$ is represented using basis functions. For instance, cubic splines with $K$ basis functions can be used, where:

$$f_j(x_j) = \sum_{k=1}^{K} \beta_{jk} B_{jk}(x_j),$$

with $B_{jk}(x_j)$ being the spline basis functions and $\beta_{jk}$ the coefficients. Next, a design matrix $\mathbf{X}$ is constructed for the model, made up of design matrices for each smooth term $\mathbf{X}_j$. The coefficients are estimated using a penalised version of the Iteratively Reweighted Least Squares (IRLS) algorithm, known as Penalised Iteratively Reweighted Least Squares (P-IRLS). At each iteration $k$, the coefficients are updated according to:

$$\boldsymbol{\beta}^{(k+1)} = \left( \mathbf{X}^T \mathbf{W} \mathbf{X} + \sum_j \boldsymbol{\lambda}_j \mathbf{S}_j \right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z},$$

where $\mathbf{S}_j$ is the penalty matrix for each smooth term, and $\boldsymbol{\lambda}_j$ represents the smoothing parameters. This iterative process incorporates both the fitting of the model and the penalisation of the smooth functions to prevent overfitting.

## 5.2 Tensor Product Smooth Interactions in Generalised Additive Models

When looking at the general form of a generalised additive model (GAM), equation 8, notice that there is a multivariate smooth function included. These can be used to capture complex interactions and, in this paper, are used when considering highly complex, "full" GAMs. Using the term "full" with respect to GAMs is somewhat an abuse of terminology, since "full" (or "complete") is typically used when referring to linear models. Regardless, it is a convenient way to refer to a GAM that includes the smooth main effects and all the possible complex smooth interaction terms.

It quickly became apparent that choosing how to include multivariate smooths was not as straightforward as expected. There are several ways to include complex interactions, with the most prominent methods being isotropic smooths, and two ways to include tensor product smooth interactions e.g. $\mathrm{s}(x_1, x_2)$, $\mathrm{te}(x_1, x_2)$ and $\mathrm{ti}(x_1, x_2)$ respectively, where $x_1, x_2$ are covariates. For this section, Simon Wood's book[19] was once again extremely useful.

Isotropic smooths are functions that produce identical predictions of the response variable under any rotation or reflection of the covariates. This implies that the smooth function is invariant to geometric transformations of the covariate space, meaning it treats all directions equally and does not favour any particular orientation or axis. Consequently, to maintain the desired invariance assumption of the isotropic smooth, the covariates should be on the same scale or unit. If the covariates are not on the same scale, rotations or reflections would distort the relative contributions of different covariates, leading to different predictions. In simple terms, isotropic smooths assume uniform smoothness in all directions, using the same smoothing parameter for each predictor. In our context, isotropic smooths are not appropriate for including multivariate smooths since the variables are not on the same scale or unit.

Tensor product smooths address the issue of modeling responses when variables have different units or scales, allowing for different levels of smoothness in different directions. These smooths combine marginal smooths of each predictor to construct an interaction term by taking the tensor product of basis functions for each predictor. For instance, using cubic splines, the tensor product of cubic splines in each direction forms the smooth. For three variables, the basis functions are represented as:

$$f_{x_1x_2x_3}(x_1, x_2, x_3) = \sum_{i=1}^{I}\sum_{l=1}^{L}\sum_{k=1}^{K} \beta_{ilk} b_k(x_3) d_l(x_2) a_i(x_1).$$

If $\odot$ is the row-wise Kronecker product, given an appropriate ordering of the $\beta_{ilk}$ into a vector $\boldsymbol{\beta}$, the design matrix $\mathbf{X}$ can then be evaluated as follows:

$$\mathbf{X} = \mathbf{X}_{x_1} \odot \mathbf{X}_{x_2} \odot \mathbf{X}_{x_3}.$$

The marginal design matrices $\mathbf{X}_{x_1}$, $\mathbf{X}_{x_2}$, and $\mathbf{X}_{x_3}$ are the design matrices that evaluate the marginal smooths for each variable. Using the row-wise Kronecker product then combines these marginal design matrices into a single design matrix and ensures all interactions between the variables are included. The accompanying penalty, as derived by Wood, would be:

$$J(f_{x_1x_2x_3}) \approx J^*(f_{x_1x_2x_3}) = \lambda_{x_1} J^*_{x_1}(f_{x_1x_2x_3}) + \lambda_{x_2} J^*_{x_2}(f_{x_1x_2x_3}) + \lambda_{x_3} J^*_{x_3}(f_{x_1x_2x_3}), \tag{9}$$

where

$$\begin{aligned}
J^*_{x_1}(f_{x_1x_2x_3}) &= \boldsymbol{\beta}^\top \tilde{\mathbf{S}}_{x_1} \boldsymbol{\beta}, \quad \tilde{\mathbf{S}}_{x_1} = \mathbf{S}'_{x_1} \otimes I_L \otimes I_K, \\
J^*_{x_2}(f_{x_1x_2x_3}) &= \boldsymbol{\beta}^\top \tilde{\mathbf{S}}_{x_2} \boldsymbol{\beta}, \quad \tilde{\mathbf{S}}_{x_2} = I_I \otimes \mathbf{S}'_{x_2} \otimes I_K, \\
J^*_{x_3}(f_{x_1x_2x_3}) &= \boldsymbol{\beta}^\top \tilde{\mathbf{S}}_{x_3} \boldsymbol{\beta}, \quad \tilde{\mathbf{S}}_{x_3} = I_I \otimes I_L \otimes \mathbf{S}'_{x_3}.
\end{aligned}$$

Each $\mathbf{S}_i$ corresponds to a penalty matrix for the basis functions of one variable, penalising wiggliness. The prime and star, i.e. $\mathbf{S'}_i$ and $J_i^*$, represent the re-parametrisations that they have undergone (explained in detail in Wood's book[19]). The $\otimes$ symbol denotes the Kronecker product and $\lambda_i$ is the smoothing parameter controlling the degree of smoothness for each variable independently. To avoid further complexity and unnecessary depth, further details and derivations are omitted. The most important thing to notice here is that there is a penalty term for each marginal smooth.

The tensor product smooths include both the marginal smooths of each variable and their interactions e.g. the smooth main effects are included within $\text{te}(x_1, x_2)$. For inference, this can be problematic because it makes it challenging to test whether the main effects and lower-order interactions would be sufficient (or whether the higher-order interaction is statistically significant). Since our main goal is prediction, there would presumably be no question about using this method to include multivariate smooths. However, it isn't quite so simple.

There is another formulation of tensor product smooth interactions, $\text{ti}(\cdot)$, that is used when one wants to exclude the marginal smooths and lower order interactions. These are referred to as "ANOVA decompositions of smooths". According to Wood, "if we subject the marginal smooths of a tensor product to sum-to-zero identifiability constraints before constructing the tensor product basis, then the resulting interaction smooths do not include the corresponding main effects"[19]. This constraint centres the smooth functions around zero, effectively removing the constant term from the basis functions, isolating interaction effects from main effects. This is achieved by column-centering the marginal design matrices $\mathbf{X}_i$ before constructing the final design matrix $\mathbf{X}$. This then removes the unit function from the span of the marginals, with the result that the tensor product basis will not include the smooth main effects that result from the product of a marginal basis with the unit functions in the other marginal bases. The resulting penalty structure for this formulation is slightly different. Instead of a sum of penalty terms for each marginal smooth, there is one combined penalty term for the interaction.

In practice, it was found that using $\text{te}(\cdot)$ alone resulted in models with worse evaluation metrics. When including the main effects and lower order interactions there was an improvement in performance, but further improvement was seen when using $\text{ti}(\cdot)$. Consider the following two models:

$$\hat{y} = \beta_0 + \text{s}(x_1) + \text{s}(x_2) + \text{ti}(x_1, x_2) \tag{10}$$

$$\hat{y} = \beta_0 + \text{te}(x_1, x_2) \tag{11}$$

These two formulations should be practically identical, given that they have the same smoothing basis and are only partitioned differently. However, there is also a difference in their penalty structure. The $\text{ti}(\cdot)$ model, equation 10, has one penalty per univariate smooth term and a combined penalty for the interaction term. In contrast, the $\text{te}(\cdot)$ model, equation 11, is made up of two penalty terms i.e. one per marginal smooth as demonstrated in equation 9. Why exactly this difference results in the $\text{ti}(\cdot)$ model arriving at better results is not entirely clear. More penalisation could result in reduced complexity which could possibly limit overfitting, however this is somewhat speculative. Interestingly, this same scenario and outcome was seen by Wood in an example implementation in chapter 7.2.3[19]. In any case, due to the better results, when implementing the full GAM, $\text{ti}(\cdot)$ was used.

## 5.3 Research Questions

The Research Questions section outlines the specific questions and goals that this thesis aims to address. These are derived from the motivation, theoretical framework, and literature review, reflecting the gaps identified and the objectives of the research.

### 5.3.1 Modelling: Theoretical Choices and Rationale

This paper considers both generalised linear models (GLMs) and generalised additive models (GAMs) as modelling frameworks. These were introduced in Section 5.1. Beginning with GLMs, a conventional approach favoured for its simplicity and effectiveness, provides a preliminary analysis of relationships within data and also allows us to investigate whether a complex, nonlinear approach is necessary or not.

The preprocessing stage involved transformations applied to both the outcome variable and independent variables. These transformations helped to both improve the linearity between the outcome variable and the predictors as well as bring the distribution of the outcome variable closer to normal. Such transformations ensure that the assumptions underlying linear regression — linearity, homoscedasticity, and normality — are more closely met, thus providing a more robust analysis. No link function was found to be necessary, thus the models applied are simply linear models (LMs). Although GAMs are designed to handle non-linearity, starting with variables that exhibit simpler, near-linear relationships can reduce the complexity of the smooth functions required. This simplification can lead to more stable models. Essentially, less complex splines might be required to capture the dynamics in the data, leading to a more straightforward computation and potentially more robust model outcomes. Thus, the same transformations of the variables are used when applying GAMs.

Since the donor data spans a very diverse population, subjects that are obese, very short, and well into adulthood are included. Considering these cases, the potential limitations of GLMs – which were indicated to in the exploration – become apparent. In these instances, the relationship between variables may not be adequately captured by a linear approach, even after using a link function and transforming the dependent variables. While interactions can, and should be, included to better capture the complex relationships between the predictors and the outcome variable, they do not necessarily capture the possible nonlinear relationships. Furthermore, including an excessive number of interactions could possibly lead to overfitting. This can be particularly true if the model includes interactions that are not supported by a theoretical understanding of the data and variables i.e. simply including all variables and all possible combinations as is done in a full model.

Generalised additive models extend the linear model by allowing for nonlinear relationships between the dependent variable and the predictors via smooth functions. With the donor dataset exceeding 70,000 observations, concerns regarding overfitting with GAMs are mitigated. Moreover, splines incorporate regularisation techniques directly into their formulation i.e. a penalised version of the usual iteratively reweighted least squares (P-IRLS) is used. This regularisation in splines directly counters the increase in complexity, leading to a model that is more robust to overfitting. In contrast to GLMs with numerous interactions, GAMs could provide a more parsimonious approach to modelling complex, nonlinear relationships without excessive parameterisation. However, it would still be worthwhile to also consider a "full" GAM which includes tensor products. Tensor products are a way to include, in essence, n-dimensional interaction surfaces when the variables are not on the same scale/unit. Including such terms reduces interpretability further, thankfully that is not a concern since accurate prediction is the goal. However, the added complexity does also further increase risks of overfitting. To investigate whether overfitting is occurring or not, the performance on out-of-sample evaluation metrics of the full LM and GAM are compared to a reduced LM and GAM with comparable goodness-of-fit metrics. If generalisability is not compromised, there is no reason not to use the full model specification.

Accurate prediction of the healthy aortic valve (AV) diameter is not the only goal, to compute Z-scores it is also necessary to obtain an estimate of the standard deviation that accompanies the prediction. Since inference is not the goal, if the homoscedasticity assumption is violated it is of more concern to accurately predict the AV diameter size and associated standard deviation, not to ensure the assumption holds. To account for the heteroscedasticity, the residuals squared can be modelled with respect to the independent variables. The formula for this is given in the Methodology section (eq. 18). This is a model of the conditional variance and when taking the square root gives the conditional standard deviation. Additionally, it's important to note that the residuals also reflect the natural variability in AV diameters among healthy individuals e.g. for very young individuals, before puberty, less variation in the AV diameter would be expected. Given the data and predictors included, for even the best model this natural variation may not be possible to capture. Thus, it will likely be necessary to model the conditional standard deviation.

When it comes to model evaluation, there are a few considerations. For model selection, the R squared, AIC, and BIC were looked at. When the primary goal is prediction, these performance metrics, while useful, are not always the most critical metrics to consider. They are based on in-sample error estimates. This means they might not fully capture the model's performance on new data, which is a more critical aspect of predictive modeling. However, during the early stages of model selection, to obtain reduced models with similar performance metrics, the R squared, AIC and BIC are utilised. Direct measures of predictive accuracy, i.e. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), are more important in this case. To assess uncertainty, the conditional standard deviation with respect to the variables is plotted. Using the residuals, a model for the conditional variance can be made (eq. 18). Taking the square root of the conditional variance, the conditional standard deviations are plotted against the variables. The impact of the other variables are then isolated from each other as these have been accounted for in the model. Otherwise, the plots look very similar to each other since age, height, and weight interact heavily. These are used to help find the best model as well as compare the final model to established models from the literature. Lastly, they can also show whether the previous models were right to assume the homoscedasticity assumption was not violated and to use a static value for their estimates of the standard deviation to be used in the Z-score formula.

### 5.3.2 Should BSA be used?

By far the most popular BSA formula used in the established models is the Haycock formula[20]. This formula is as follows:

$$BSA_{\text{Haycock}}\ (\text{m}^2) = 0.024265 \cdot \text{Weight}\ (\text{kg})^{0.5378} \cdot \text{Height}\ (\text{cm})^{0.3964} \tag{12}$$

This is formula originates from a 1978 paper[20]. The coefficient and exponents in the Haycock formula are derived from fitting the model to a dataset consisting of 81 patients that range from infants to adults and it was aimed to have a normal distribution of body types. Individuals who are significantly above or below average body proportions may not be well-represented by this model. The formula assumes a homogeneous relationship between height and weight across all individuals. For a larger, more diverse dataset including atypical proportions — such as in extremely obese or very short individuals — the relationship between height and weight and how they contribute to body surface area likely do not follow the same pattern. This is exemplified by the variation seen in BSA values depending on the formula used. In Figure 20, at first glance, on the entire donor population, the different formulas are surprisingly consistent. This is despite their varying formulations and that the coefficients have been estimated on different datasets.

However, when looking at extreme subsets of the data a different picture can be seen. The average of applying the different formulas can be taken for each observation and the distribution of the respective differences for each formula can be visualised via boxplots. Two example subsets are demonstrated in Figure 21.

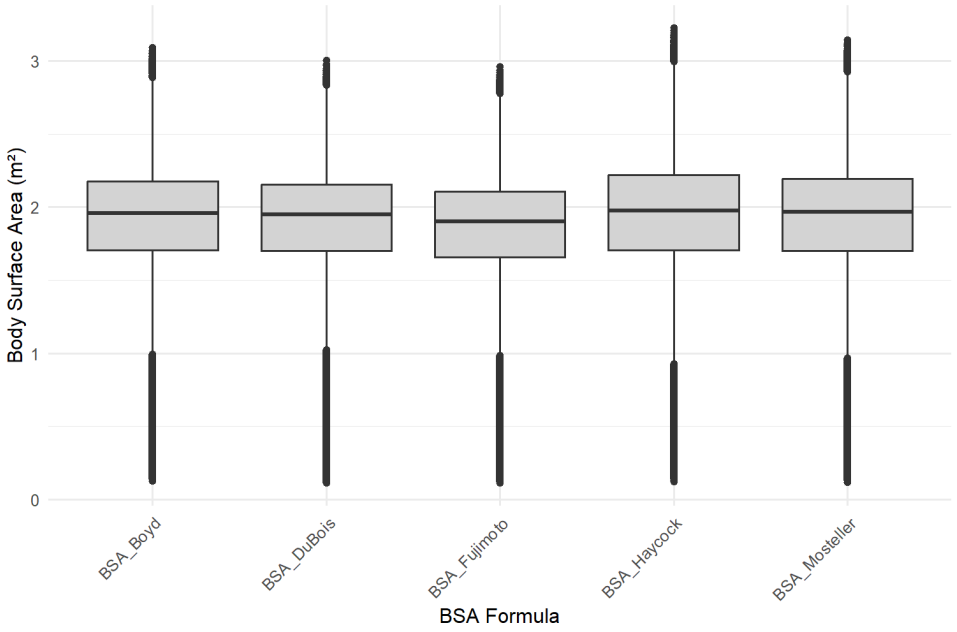
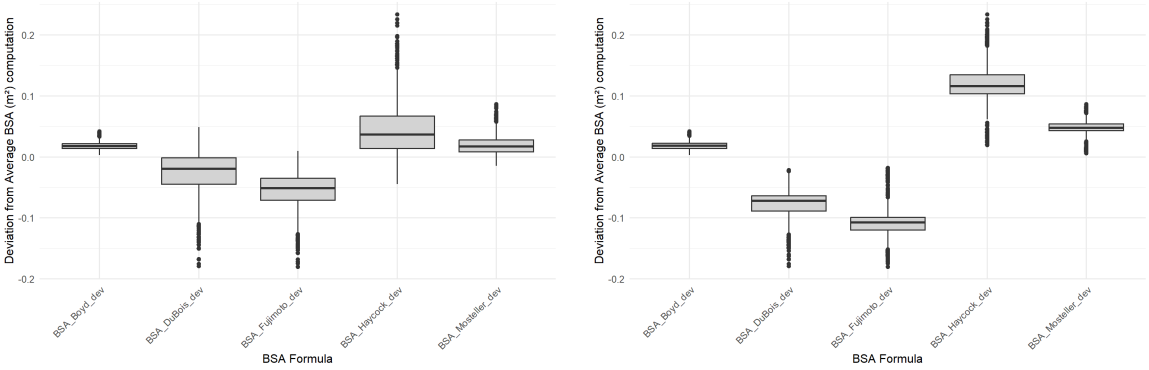



Figure 20: Various BSA formulas[21][22][23][24] applied to the full donor dataset.

In Figure 21, it can be seen that there is variation in the BSA values for very short adults, especially in the ranges. For obese individuals, there is more variation evident in the medians. When looking at the y-axis, the deviations do look slight. However, BSA values only range between around 0.1 and 3.5. Whether this disparity is indicative of issues with implementing models using BSA can be investigated. While the Lopez and donor datasets do not include body surface area measurements, BSA as a predictive variable compared to using height and weight (along with interactions) can be analysed. That analysis will focus on evaluating how models, which vary based on the included variables, perform in predicting AV diameter. Special attention will be given to how these models perform on individuals in extreme cases.




(a) **Very short adults.** A subset of 1160 adults below the 3rd percentile in height.

(b) **Very obese individuals.** A subset of 5390 individuals with a BMI greater than 40.

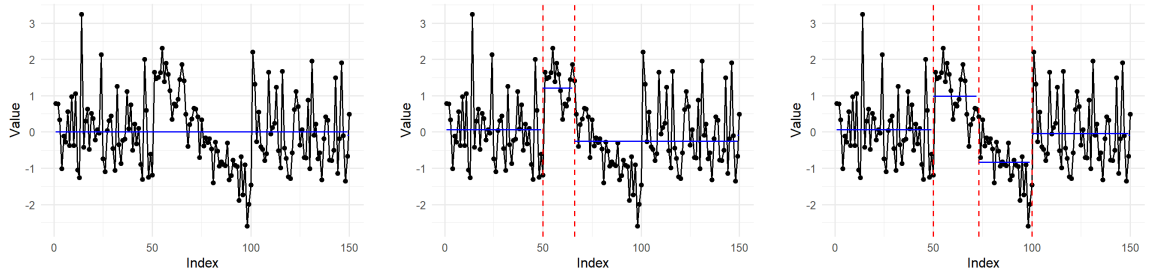
Figure 21: Boxplots comparing five different BSA formulas applied to subsets of the donor data.

### 5.3.3 Trend Correction: A Bespoke Changepoint Algorithm

This section explains why a bespoke implementation of the segment neighbourbood (SN) algorithm for changepoint analysis was necessary for trend correction in the donor data. The limitations of existing R packages are discussed and demonstrated via a simulated example.

During the data exploration phase, what looks to be a trend was observed between two time points in the AV diameter measurements over time. This trend was also evident in the residuals following the fitting of an initial predictive model. Notably, this trend is not documented in existing literature, nor could it be explained by expert opinions, suggesting an anomaly unique to the donor dataset.

The standard implementations of changepoint models available in various R packages assume a flat gradient between changepoints, which did not align with the observation of a decreasing trend rather than a mere shift in mean values. Moreover, the observed negative trend seems to be rather symmetric around the mean. Therefore, using algorithms that rely on minimisation of the MSE based on fitting means (a flat gradient) would likely miss these changepoints. An example of such a package is the 'changepoint' package in R. This has been applied to a simulated example:



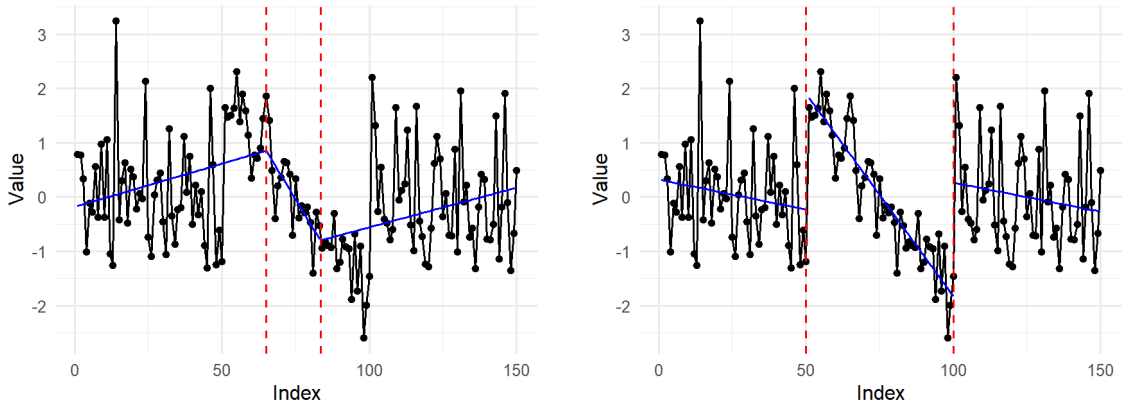

(a) **Binary Segmentation.** Zero changepoints were found.

(b) **PELT.** Two changepoints were identified, one is correct.

(c) **Segment Neighbourhood with SIC penalty** Three changepoints were identified, two are correct.

Figure 22: Applying all possible methods supported by the 'changepoint' package in R on simulated data that mimics the behaviour seen visually in the donor data.

A different R package allows for more flexibility than simply fitting means. 'Segmented' fits segmented (piecewise linear) regression models to the data, the sum of squared residuals is minimised to find the best fit. The package allows the slope of the regression line to change at designated changepoints, it does not support scenarios where these segments are disjointed. The end point of one linear segment must meet the starting point of the next, ensuring piecewise continuity. When applying this to the simulated data (Figure 23a), it can be seen that this would not be suitable for the donor data.




(a) **Segmented R package.** Two incorrectly identified changepoints.

(b) **Strucchange R package.** Both changepoints correctly identified.

Figure 23: Comparing two R packages that allow for linear fits.

The final, and most suitable, package found was 'strucchange'. This package allows for disjointed linear fits between segments and can be seen in Figure 23b. It works to find the correct changepoints in the simulated data and it outputs the coefficients for the linear models fit within each segment. This would likely be sufficient for our use case, however there are two limitations. The first is that there is no choice in the penalty used. For computing the optimal changepoints, their algorithm is based on the SN algorithm and by default it minimises the total residual sum of squares (RSS) for a specified maximum number of changepoints. Since the RSS gets smaller endlessly with more changepoints, once the maximum number of changepoints has been reached, BIC is used to determine the optimal number of changepoints (complexity is then penalised). Every number of changepoints and their respective RSS and BIC is printed. There is no way to manually tune how heavily the complexity is penalised, and this is adequate in the simulated example. On the one hand, the fixed structure ensures that model comparison is objective and consistent. On the other hand, it can be useful to allow for different penalty choices. For specific applications, especially in complex models or when domain-specific knowledge suggests a different approach, the fixed penalty choice might not be ideal. The second drawback is that there is also no flexibility in the cost function, only linear regression can be used. For the trend seen in the donor data, this is likely sufficient but it does limit easier and further exploration i.e. different cost functions may achieve better results.

Consequently, a segment neighbourhood algorithm with a different approach was developed. This SN algorithm accommodates different cost functions and penalty choices, and allows for discontinuity between segments. Additionally, this implementation outputs the optimal number of changepoints and their locations directly.

Bement and Waterman showed that the segment neighbourhood algorithm converges to the optimal number of changepoints[25]. The complexity of this algorithm has been well-defined by Auger and Lawrence[26]. Its complexity and runtime grows very fast with increasing amounts of data and changepoints. Explicitly, the time complexity of the algorithm is $\mathcal{O}(Qn^2)$, where $Q$ is the maximum number of changepoints under consideration and $n$ is the number of data points, i.e. the algorithm's performance degrades quadratically with the number of data points and linearly with the number of changepoints. To combat this, a minimum window (segment) size $\omega$ can be introduced to speed up the aglorithm. The overall complexity of each iteration will be reduced, as the number of feasible segmentations is limited. The only concern is that this constraint might lead the algorithm to not identify trends in segments smaller than $\omega$. If the underlying true changepoints all satisfy the minimum window size $\omega$, then the algorithm should still converge to the optimal segmentation. Furthermore, by skipping over short-term fluctuations, the segmentation might become more robust against noise. With an appropriate choice for $\omega$, there is not much concern that genuine trends are missed. Moreover, it would not be wise to fit linear models in very small segments, and with the

penalty term it would likely not achieve a minimised cost function anyway.

While the trend correction will be applied to the AV diameter measurements, the changepoints and the model of the trends will be identified via the residuals over time. Since the trend(s) in the AV measurements might be influenced by the independent variables included in the model, identifying the changepoints and trend(s) this way can lead to confounding effects where any identified trend is a mix of the genuine trend and the effects of the predictors. By applying a correction based on this, one may end up removing information that would have been provided by the predictors. Thus, it is a more appropriate approach to use the residuals. The residuals isolate the portion of the AV measurements not explained by the model i.e. the impact of the independent variables has already been accounted for. Therefore, trend(s) in the residuals reflect a purer signal of unmodeled trends. The goal is to ensure that the adjusted data results in a model that closer adheres to the assumptions and may generalise better.

### 5.3.4 Why Correcting a Possible Echocardiographic Bias Matters

The Cryolife donor dataset is made up of physical measurements of the AV diameter, whereas the previous models have been based on echocardiographic measurements of the AV diameter. Since echo measurements are an estimate of the true AV diameter, it seems counter intuitive to include a correction based on a possible echo bias. However, it is important to note that there are two scenarios in which predictions of the AV diameter based on the donor data will be used:

1. **For diagnostic purposes** to decide whether a patient's health complaints are due to an inappropriately sized aortic annulus
2. **In preparation for surgery** to predict the AV diameter a healthy person of the patient's specified characteristics should have had

When considering the first use case: if there is a quantifiable, systematic difference between the estimated AV diameter via echocardiographs it would be extremely important to incorporate this into the model. Since diagnosis is done via echo estimates, it could lead to an over/under-diagnosis of individuals. In the second use case, it makes no sense to include an echocardiographic bias correction. When predicting the most appropriate AV diameter for a patient, the measurements closest to physical reality should be used. This implies that there should be two separate models, one for diagnosis via Z-scores and one for precise AV diameter prediction.

Once pseudo-echo AV diameter measurements have been obtained for the donor data, it will not suffice to only create a bridging model between the physical and pseudo-echo measurements. The predictive model for diagnosis will need to be retrained with the pseudo-echo measurements so that a model for the conditional variance based on these echo corrected measurements can be created.

# 6 Methodology

## 6.1 Model Selection Methodology

As mentioned in Section 5.3.1, linear models (LMs) and generalised additive models (GAMs) were both considered for predicting healthy aortic valve diameters. To identify the best model structure and eventually justify it, the Results section presents the outcomes for five different model structures. These five models are outlined below, followed by the method used to compare them.

To investigate whether a nonlinear approach is beneficial, linear models are compared to GAMs. Furthermore, since there is a huge amount of data, highly complex models are considered in attempt to achieve the highest accuracy possible. Specifically, these are the full LM and GAM.

**Full Linear Model:**

$$\begin{aligned} ln(\text{AV}) = {} & \beta_0 + \beta_1 \text{Sex} + \beta_2 ln(\text{Age}+1) + \beta_3 ln(\text{Weight}) + \beta_4 \sqrt{\text{Height}} \\ & + \beta_5 (ln(\text{Age}+1) \times \text{Sex}) + \beta_6 (ln(\text{Weight}) \times \text{Sex}) + \beta_7 (\sqrt{\text{Height}} \times \text{Sex}) \\ & + \beta_8 (ln(\text{Weight}) \times ln(\text{Age}+1)) + \beta_9 (\sqrt{\text{Height}} \times ln(\text{Age}+1)) + \beta_{10} (ln(\text{Weight}) \times \sqrt{\text{Height}}) \\ & + \beta_{11} (ln(\text{Weight}) \times ln(\text{Age}+1) \times \text{Sex}) + \beta_{12} (\sqrt{\text{Height}} \times ln(\text{Age}+1) \times \text{Sex}) \\ & + \beta_{13} (ln(\text{Weight}) \times \sqrt{\text{Height}} \times \text{Sex}) + \beta_{14} (ln(\text{Weight}) \times \sqrt{\text{Height}} \times ln(\text{Age}+1)) \\ & + \beta_{15} (ln(\text{Weight}) \times \sqrt{\text{Height}} \times ln(\text{Age}+1) \times \text{Sex}) + \varepsilon. \end{aligned} \tag{13}$$

**Full Generalised Additive Model:**

$$\begin{aligned} ln(\text{AV}) = {} & \beta_0 + \beta_1 \text{Sex} + s(ln(\text{Age}+1)) + s(ln(\text{Weight})) + s(\sqrt{\text{Height}}) \\ & + s(ln(\text{Age}+1),\ \text{Sex}) + s(ln(\text{Weight}),\ \text{Sex}) + s(\sqrt{\text{Height}},\ \text{Sex}) \\ & + \text{ti}(ln(\text{Weight}), ln(\text{Age}+1)) + \text{ti}(\sqrt{\text{Height}}, ln(\text{Age}+1)) + \text{ti}(ln(\text{Weight}), \sqrt{\text{Height}}) \\ & + \text{ti}(ln(\text{Weight}), ln(\text{Age}+1),\ \text{Sex}) + \text{ti}(\sqrt{\text{Height}}, ln(\text{Age}+1),\ \text{Sex}) \\ & + \text{ti}(ln(\text{Weight}), \sqrt{\text{Height}},\ \text{Sex}) + \text{ti}(ln(\text{Weight}), \sqrt{\text{Height}}, ln(\text{Age}+1)) \\ & + \text{ti}(ln(\text{Weight}), \sqrt{\text{Height}}, ln(\text{Age}+1),\ \text{Sex}) + \varepsilon, \end{aligned} \tag{14}$$

where $s(x)$ denotes a smooth function applied to the variable $x$, $s(x, \text{Sex})$ are separate smooth functions of the variable $x$ separated by sex, and $\text{ti}(\cdot)$ is a tensor product smooth interaction. A large value of 15 was used for the number of knots, this is because the large amount of data and smoothing penalties limit the risks of overfitting.

To examine whether overfitting may still be occurring despite the large amount of data, the full LM and GAM are compared to a much reduced LM and GAM. A backwards elimination approach was used to find reasonable reduced models. Using R-squared, BIC, AIC, and the residual plots, many interactions were removed and similar results for the performance metrics were maintained. The reduced models are shown below.

**Reduced Linear Model:**

$$\begin{aligned} \log(\text{AV}) = {} & \beta_0 + \beta_1 \text{Sex} + \beta_2 \log(\text{Age}+1) + \beta_3 \log(\text{Weight}) + \beta_4 \sqrt{\text{Height}} \\ & + \beta_5 (\log(\text{Age}+1) \times \text{Sex}) + \beta_6 (\log(\text{Weight}) \times \text{Sex}) \\ & + \beta_7 (\sqrt{\text{Height}} \times \text{Sex}) + \beta_8 (\log(\text{Weight}) \times \sqrt{\text{Height}}) \\ & + \beta_9 (\sqrt{\text{Height}} \times \log(\text{Age}+1)) + \beta_{10} (\log(\text{Weight}) \times \log(\text{Age}+1)) \\ & + \beta_{11} (\log(\text{Weight}) \times \sqrt{\text{Height}} \times \log(\text{Age}+1)) + \varepsilon. \end{aligned} \tag{15}$$

**Reduced Generalised Additive Model:**

$$\begin{aligned}\log(\text{AV}) = \beta_0 + \beta_1 \text{Sex} + s(\log(\text{Age}+1)) + s(\log(\text{Weight})) + s(\sqrt{\text{Height}}) \\ + s(\log(\text{Age}+1),\ \text{Sex}) + s(\log(\text{Weight}),\ \text{Sex}) + s(\sqrt{\text{Height}},\ \text{Sex}) + \varepsilon,\end{aligned} \tag{16}$$

Lastly, to investigate BSA as a predictor, the best performing model using BSA as a variable can be compared to the best model using height and weight separately. The model using BSA is the full GAM as shown below,

**Generalised Additive Model using BSA:**

$$\begin{aligned}\log(\text{AV}) = \beta_0 + \beta_1 \text{Sex} + s(\log(\text{Age}+1)) + s(\log(\text{BSA})) \\ + s(\log(\text{Age}+1),\ \text{Sex}) + s(\log(\text{BSA}),\ \text{Sex}) \\ + \text{ti}(\log(\text{BSA}),\ \log(\text{Age}+1)) \\ + \text{ti}(\log(\text{BSA}),\ \log(\text{Age}+1),\ \text{Sex}) + \varepsilon.\end{aligned} \tag{17}$$

Once the best model has been selected from these five options, the conditional variance can be modelled to compute the conditional standard deviation necessary for the Z-score computation.

**Generalised Additive Model for the Conditional Variance:**

$$\begin{aligned}r^2 = \beta_0 + \beta_1 \text{Sex} + s(\text{Age}) + s(\text{Weight}) + s(\text{Height}) \\ + s(\text{Age},\ \text{Sex}) + s(\text{Weight},\ \text{Sex}) + s(\text{Height},\ \text{Sex}) + \varepsilon,\end{aligned} \tag{18}$$

where $r$ denotes the residuals. The residuals are computed on the original scale i.e. the predictions are back-transformed. Taking the square root of predictions based on this model gives the conditional standard deviation.

### 6.1.1 Comparing the Accuracy and Uncertainty of Models

This section covers the methodology used to compare the five models under consideration. When applying an **in-sample evaluation**, the following method is utilised:

1. Propose a model specification
2. Goodness-of-fit measures i.e. R-squared, various residuals plots
3. Model selection criteria i.e. AIC and BIC
4. Evaluation metrics i.e. root mean squared error (RMSE), mean absolute error (MAE)

Two approaches are used for the **out-of-sample evaluation**:

1. Introduce test/train splits using K-fold cross validation and take an average.
2. Utilise external validation datasets, namely the Lopez *et al.* and the UMCG echocardiographic datasets.

The evaluation metrics, RMSE and MAE, can then be applied to the test and validation datasets.

Aside from the measures and metrics applied for in-sample and out-of-sample evaluation, it was of particular interest to see how the models performed on extreme subsets of the data. To investigate this, the following method was employed:

1. Create five extreme case subsets of the data:
   (a) **Obese individuals:** those with BMI greater than 40
   (b) **Very short adults:** individuals above 18 and below the 3rd percentile in height
   (c) **Older females:** females older than 55
   (d) **Older males:** males older than 55
2. For each subset:
   (a) For each model, the RMSE and MAE can be computed via the method described in Figure 24.
   (b) Repeat this 100 times, store the obtained RMSEs and MAEs
   (c) Return the average RMSE and MAE for both models

Via this method, the models are trained and tested on the same data for each run. This ensures a fair comparison.

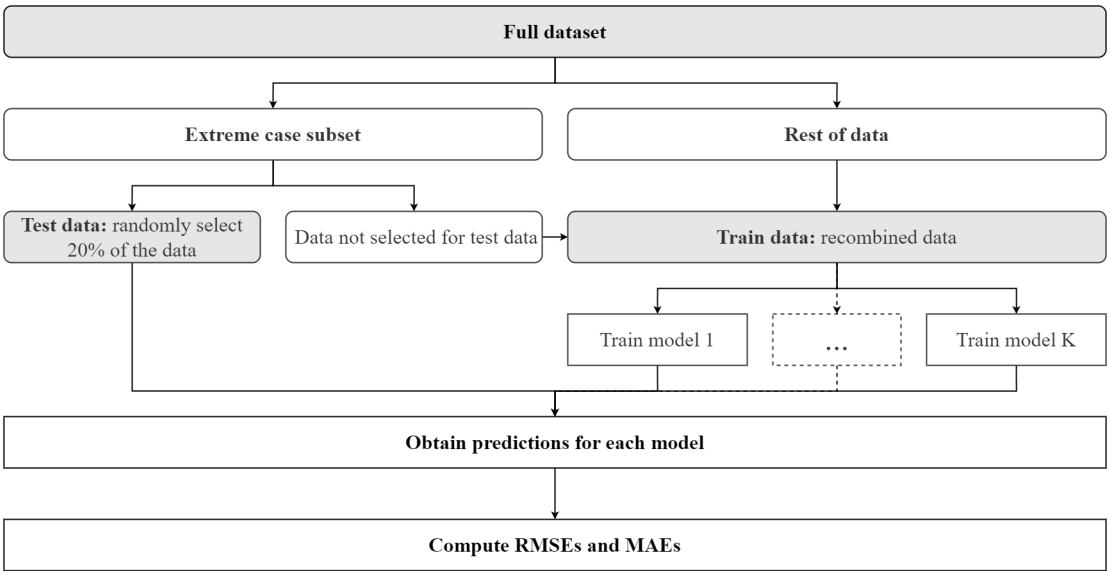



Figure 24: Flowchart demonstrating the method used to compare the performance of models on extreme subgroups.

Moreover, to better compare the uncertainty and not just accuracy, the conditional standard deviation for each model can be plotted against the variables. All the models can be used to make conditional variance predictions based on the donor data, and then the conditional standard deviations can be easily plotted.

## 6.2 Implementation of a Segment Neighbourhood Algorithm for Trend Correction

The segment neighbourhood algorithm identifies the optimal segmentation of time-series data by minimising a penalised cost function across potential changepoints and minimum window/segment size constraint. It uses dynamic programming to efficiently compute the cost of segments. The code can be found in the Appendix C.

**Input:**

- Data sequence $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ where $y_i \in \mathbb{R}$
- Maximum number of changepoints $Q$
- Penalty constant $\beta$
- Cost function $\mathcal{C}(\mathbf{y}, \beta, q)$: computing the cost (measure of fit) and includes a penalty term dependent on $\beta$ and the number of changepoints $q$
- Minimum window (segment) size $\omega$

**Output:** The location of the optimal changepoints as the vector $s$ and total cost $\mathcal{C}_T$.

**Initialisation:**

- Initialise cost matrix $M \in \mathbb{R}^{N+1,Q+1}$ with $\infty$ as elements (arbitrarily large elements)
- Initialise matrix $L \in \mathbb{R}^{N+1,Q+1}$ with $-1$s to store positions of last changepoints
- Start with $\mathcal{C}_0 = \mathcal{C}(\mathbf{y}_{1:0}, \beta, 0) = 0$ i.e. the cost of zero segments and no data is zero

**Dynamic Programming:**

1. For each potential number of changepoints $q$ from 1 to $Q+1$
   (a) For each potential ending index $i$ from 2 to $N+1$
   - If $q = 1$ and $i - 1 \geq \omega$,
   $$M_{i,q} = \mathcal{C}(\mathbf{y}_{1:(i-1)}, \beta, q-1).$$
   - If $q > 1$,
   $$M_{i,q} = \min_{\substack{0<j<i \\ i-j\geq\omega}} (\mathcal{C}(\mathbf{y}_{j:(i-1)}, \beta, q-1) + M_{j,q-1})$$
   $$L_{i,q} = \operatorname*{argmin}_{\substack{0<j<i \\ i-j\geq\omega}} (\mathcal{C}(\mathbf{y}_{j:(i-1)}, \beta, q-1) + M_{j,q-1}).$$
2. Return the matrices $M$ and $L$

**Reconstruction of Optimal Segmentation:**

1. Compute the total cost $\mathcal{C}_T$ from the last row of the cost matrix $M$. The vector $m := M_{N+1,(1:Q+1)}$ contains the total costs for segmenting the full data sequence for zero up to $Q$ changepoints, $m_i \in \mathbb{R}$ is an element of $m$. The lowest total cost $\mathcal{C}_T$ is
$$\mathcal{C}_T = \min_i m_i$$
2. Identify the optimal number of changepoints $q^*$ that result in the minimum total cost:
$$q^* = \operatorname*{argmin}_i (m_i) - 1$$

3. Extract the segmentation path from the matrix $L$ using the optimal number of changepoints. Starting at the last position $N+1$ and tracing back through the entries in $L$ which point to the indices of the minimising changepoint for each segment:
   (a) Initialise a zeros vector to collect the final changepoints $s = \vec{0}$
   (b) Initialise a current position index $p = N+1$ and current changepoint index $k = q^*$
   (c) While $k > 0$:
      i. Record the segment ending at $s_k = L_{p,k+1} - 1$ and update $p = L_{p,k+1}$
      ii. Decrement $k$ by one to move to backtrace to the previous segment

**Output:** Return the vector of changepoints $s$ and the total cost $\mathcal{C}_T$

**Neutralising the Trend(s)**

Upon identifying the changepoints in the residuals, a linear model is fitted to the residuals within each segment to capture the trend. The correction is then applied to the aortic valve (AV) diameter measurements themselves by subtracting the difference between the linear model(s) predictions and the AV diameter measurements. This method effectively neutralises the unwarranted trend(s).

**Example and Intuitive Explanation:**

Consider a simple example where $N = 5$ and $Q = 2$, and this can be simplified further by ignoring the penalty constant $\beta$ and the minimum window size $\omega$. The matrix $M$ keeps track of the minimal cost of segmenting the data with different numbers of changepoints:

- Rows in $M$ represent the number of data points considered from the beginning of the dataset (from 0 up to N)
- Columns in M represent the number of changepoints used (from 0 up to Q)
- The entry $M_{i,q}$ represents the minimal cost of segmenting the first $i$ data points with $q$ changepoints

The cost matrix $M$ of this example would be computed as follows:

1. **First Column ($q = 1$):**
   - $M_{1,1}$ is 0 because there are no data points and no segments meaning no cost
   - $M_{i,1}$ for $i > 0$ calculates the cost of fitting all data points up to $i$ with no changepoints. These entries can be used for further computation in the next column
2. **Second Column ($q = 2$):**
   - For each data point $i$, the cost of the best single changepoint segmentation up to $i$ is computed in $M_{i,1}$. This involves finding the index $0 < j < i$ that minimises the cost of segmenting the data into two parts, i.e. the minimal cost is the sum of two terms:
     (a) The cost of fitting data from 1 to $j$ as one segment, which were already computed for all $j$ in the previous column i.e. they are simply equal to the elements $M_{j,1}$
     (b) The cost from $j+1$ to $i$ considering it as a separate segment (thus introducing one changepoint)
3. **Further Columns ($q > 2$):**
   - Similar to $q = 1$, the minimal cost involves finding the index $0 < j < i$ such that is the sum of the following two terms is minimised:
     (a) The cost of fitting $q-1$ segments from 1 to $j$ were already computed for all $j$ in the previous column i.e. they are simply equal to the elements $M_{j,q-1}$
     (b) The cost of the segment from $j+1$ to $i$ (thus introducing another changepoint)

The resulting matrix $M$ for the example can be seen below:

$$
M = \begin{bmatrix}
0 & \infty & \infty \\
C(y_1,0) & \infty & \infty \\
C(y_{1:2},0) & C(y_2,1)+C(y_1,0) & \infty \\
C(y_{1:3},0) & \min\limits_{0<j<3}\{C(y_{j:3},1)+C(y_{1:j},0)\} & C(y_3,2)+C(y_2,1)+C(y_1,0) \\
C(y_{1:4},0) & \min\limits_{0<j<4}\{C(y_{j:4},1)+C(y_{1:j},0)\} & \min\limits_{0<j<4}\{C(y_{j:4},2)+\min\limits_{i<l<j}\{C(y_{l:j},1)+C(y_{i:l},0)\}\} \\
C(y_{1:5},0) & \min\limits_{0<j<5}\{C(y_{j:5},1)+C(y_{1:j},0)\} & \min\limits_{0<j<5}\{C(y_{j:4},2)+\min\limits_{i<l<j}\{C(y_{l:j},1)+C(y_{i:l},0)\}\}
\end{bmatrix}
$$

$$
= \begin{bmatrix}
0 & \infty & \infty \\
C(y_1,0) & \infty & \infty \\
C(y_{1:2},0) & C(y_2,1)+M_{2,1} & \infty \\
C(y_{1:3},0) & \min\limits_{0<j<3}\{C(y_{j:3},1)+M_{j,1}\} & C(y_3,2)+C(y_2,1)+M_{2,1} \\
C(y_{1:4},0) & \min\limits_{0<j<4}\{C(y_{j:4},1)+M_{j,1}\} & \min\limits_{0<j<4}\{C(y_{j:4},2)+\min\limits_{i<l<j}\{C(y_{l:j},1)+M_{l,1}\}\} \\
C(y_{1:5},0) & \min\limits_{0<j<5}\{C(y_{j:5},1)+M_{j,1}\} & \min\limits_{0<j<5}\{C(y_{j:4},2)+\min\limits_{i<l<j}\{C(y_{l:j},1)+M_{l,1}\}\}
\end{bmatrix}
$$

$$
= \begin{bmatrix}
0 & \infty & \infty \\
C(y_1,0) & \infty & \infty \\
C(y_{1:2},0) & C(y_2,1)+M_{2,1} & \infty \\
C(y_{1:3},0) & \min\limits_{0<j<3}\{C(y_{j:3},1)+M_{j,1}\} & C(y_3,1)+M_{3,1} \\
C(y_{1:4},0) & \min\limits_{0<j<4}\{C(y_{j:4},1)+M_{j,1}\} & \min\limits_{0<j<4}\{C(y_{j:4},2)+M_{j,2}\} \\
C(y_{1:5},0) & \min\limits_{0<j<5}\{C(y_{j:5},1)+M_{j,1}\} & \min\limits_{0<j<5}\{C(y_{j:5},2)+M_{j,2}\}
\end{bmatrix}
$$

The matrix $L$ stores the indices $j$ of the changepoints that yielded the minimal costs in matrix $M$. This facilitates the "reconstruction of optimal segmentation". The process begins by determining the minimal total cost of segmenting the entire data sequence, which is found in the last row of $M$. Specifically, it identifies the column that contains the smallest element in the final row, indicating the optimal number of changepoints $q^*$. Using this, we can traceback via the matrix $L$ to find the locations of the changepoints i.e.

1. Store the element of final row and column $q^*$ i.e. $L_{N+1,q*}$
2. Store the element $L_{i,q^*-1}$ with $i = L_{N+1,q^*}$
3. Store the element $L_{i,q^*-2}$ with $i = L_{i,q^*-1}$
4. Repeat until reaching the first changepoint
5. Output as the vector $s$

The first column of $L$ is made up entirely of negative ones and are not included in the locations of the final, optimal changepoints. This makes sense, since the first column of the cost matrix represents zero changepoints.

Introducing the penalty parameter $\beta$ and the minimum window size $\omega$ does not change the understanding of the algorithm. The penalty parameter simply allows one to tweak how heavily adding changepoints is penalised, and the minimum window size speeds up the algorithm. To further speed up the algorithm – since there are over 70,000 data points – the data is aggregated by taking a weekly average. This reduces the data down to 1307 data points. Considering there is still a large amount of data, choosing a window size of 50 speeds up the algorithm considerably and would be a safe choice.

## 6.3 Methodologies for the Echocardiographic Bias Correction

There are two considerations: whether there is a statistically significant difference and if there is, how can that difference be quantified.

### 6.3.1 Investigating the Statistical Significance of a Possible Bias

To investigate the statistical significance of a potential echo bias, the same method can be employed on both the UMCG and Lopez *et al.* datasets. The method used is as follows:

1. Combine the datasets using a subset of the donor data that matches the demographics of the echocardiographic dataset used
2. Introduce a new binary variable, "source", which records whether the data is from the echo or the donor dataset
3. Create a full GAM with AV as the outcome variable, taking into account all explanatory variables and their interactions as is done in equation 14, and include the source variable as a main effect only
4. Check the statistical significance of the source variable

### 6.3.2 Quantifying the Systematic Difference

Since there are two distinct datasets, the UMCG dataset which only includes adults and the Lopez *et al.* dataset only includes individuals 18 and under, slightly different approaches are used to obtain pseudo-echo measurements for the donor data. Separate bridging models will be made for the adults and individuals 18 and under. Furthermore, the Lopez *et al.* dataset was only acquired later into the project. Thus, it is interesting to compare the outcomes of the method used before this data was available.

**For Adults:** First, train the best model structure found on a subset of the donor data that matches the demographics of the UMCG data i.e. above 18 years old and and a maximum weight of 105kg. Next, obtain pseudo-physical AV diameter measurements for the UMCG echo dataset. Finally, build a bridging model between the physical AV diameter predictions and the UMCG echo AV diameter measurements. The model trained on the UMCG data is an LM:

$$\mathrm{AV}_{\mathrm{echo}} = \beta_0 + \beta_1 \mathrm{AV}_{\mathrm{Physical}} + \beta_2 \mathrm{Age} + \varepsilon, \tag{19}$$

A simple model is used due to the small amount of data. This model can then be used to obtain pseudo-echo measurements for the donor data for adults.

**For Individuals 18 and Under:** A subset of the donor data that more closely matches the demographics of the Lopez *et al.* dataset is used i.e. 18 years old and under and a maximum weight of 100kg. Two methods are employed:

1. Predict pseudo-echo AV diameter predictions for the donor data subset using the Lopez *et al.* model. Build a bridging model between the pseudo-echo and the physical measurements
2. Using the best model structure trained on the donor data subset, predict pseudo-physical AV diameter predictions for the Lopez *et al.* data. Build a bridging model between the pseudo-physical and the echo measurements. This relationship is modelled using a GAM:

$$\mathrm{AV}_{\mathrm{echo}} = \beta_0 + s(\mathrm{AV}_{\mathrm{Physical}}) + s(\mathrm{Age}) + s(\mathrm{Height}) + s(\mathrm{Weight}) + \varepsilon, \tag{20}$$

   The number of knots $k$ has been limited to $k = 3$, this avoids overfitting. This model can then be used to obtain pseudo-echo measurements for the donor data for individuals 18 and under.

Prior to obtaining the Lopez *et al.* data, the regression equations from the other established models were considered for method (1). However, for these other models, comparing the use of regression equations to using their respective datasets was not feasible due to their datasets not being publicly available. Therefore, it made more sense to focus on the bridging model created using the Lopez model compared to using the Lopez data itself. Thus, method (1) will be compared to method (2) by plotting the $\text{AV}_{\text{echo}}$ predictions against each other, if the points closely fit the line $y = x$ then method (1) would have been a good approximation of using the data itself. Finally, to obtain pseudo-echo AV diameter measurements for individuals under 18, only method (2) will be used. This approach is independent of the modelling choices of Lopez *et al.* and instead assumes our model is representative of physical AV diameter measurements. This is in contrast to assuming that the less elaborate Lopez model is representative of echocardiographic AV diameter measurements. Since our model is based on a larger dataset, it is more logical to quantify the differences in this way.

# 7 Results and Discussion

## 7.1 Model Performance Analysis

This section presents a detailed comparison of the goodness-of-fit measures and evaluation metrics for full and reduced models, using both linear models (LM) and generalised additive models (GAM). The comparison is based on several criteria, including R-squared, AIC, BIC, RMSE, and MAE, across different samples and subgroups. Moreover, a comparison of the uncertainty via estimated standard deviation plots is also included.

Table 5: Performance metrics for full and reduced models.

| Model | R-squared | AIC | BIC |
|---|---|---|---|
| Full LM | 0.902 | -148**636.0** | -148**480.0** |
| Reduced LM | 0.902 | -148**638.7** | -148**519.4** |
| Full GAM | 0.904 | -149**932.9** | -149**235.9** |
| Reduced GAM | 0.904 | -149**810.7** | -149**343.0** |

In Table 5, it can be seen that all models achieve very high adjusted R-squared values. The GAMs have a slightly higher value compared to the LMs, by only 0.002. Comparable AIC and BIC values are observed between the full and reduced models. For linear models, the reduced LM has slightly better AIC and BIC values. Despite having slightly worse AIC and BIC values, the reduced GAM is a much more parsimonious model compared to the full GAM – with 10 terms versus 22 terms, respectively – while maintaining very similar AIC and BIC values.

Table 6: In-sample and out-of-sample RMSE and MAE results for full and reduced models.

| Eval. Metric | Sample | Full LM | Reduced LM | Full GAM | Reduced GAM |
|---|---|---|---|---|---|
| **RMSE (mm)** | In-sample | 1.763985 | 1.764133 | **1.752673** | 1.754413 |
| | K-fold (Avg.) | 1.764222 | 1.764097 | **1.754455** | 1.755377 |
| | UMCG Echo | 2.543261 | 2.544735 | **2.315956** | 2.318628 |
| | Lopez Echo | 1.391259 | 1.391380 | **1.358851** | 1.359828 |
| **MAE (mm)** | In-sample | 1.384991 | 1.385223 | **1.374275** | 1.375764 |
| | K-fold (Avg.) | 1.385285 | 1.385154 | **1.375803** | 1.376667 |
| | UMCG Echo | 2.037670 | 2.039143 | **1.851335** | 1.860262 |
| | Lopez Echo | 1.104750 | 1.104991 | **1.080442** | 1.080516 |

In Table 6, the full GAM outperforms all other models across every evaluation metric for every test sample. While it is expected for the full GAM to perform best for in-sample evaluation metrics, it is noteworthy that it also outperforms in out-of-sample metrics. This indicates that the full GAM is not overfitting, as it performs better on the validation datasets compared to the more parsimonious reduced GAM. The same observation applies to the full linear model, which does not overfit when compared to the reduced LM.

Table 7: RMSE and MAE results for the different extreme subgroups for full and reduced models.

| Eval. Metric | Group | Full LM | Redu. LM | Full GAM | Redu. GAM |
|---|---|---|---|---|---|
| **Avg. RMSE (mm)** | Obese Individuals | 1.875770 | 1.875622 | **1.864358** | 1.866417 |
| | Very short adults | 1.832929 | 1.832246 | 1.819279 | **1.818732** |
| | Older females | 1.494494 | 1.496397 | **1.489364** | 1.502334 |
| | Older males | 1.596249 | 1.595268 | 1.570000 | **1.566699** |
| **Avg. MAE (mm)** | Obese Individuals | 1.468307 | 1.468306 | **1.460621** | 1.463904 |
| | Very short adults | 1.461398 | 1.461121 | **1.450046** | 1.454131 |
| | Older females | 1.191640 | 1.193283 | **1.172177** | 1.183168 |
| | Older males | 1.275015 | 1.274464 | **1.262003** | 1.262182 |

The full GAM outperforms in each extreme subgroup, except for older males, where the reduced GAM performs better for the average MAE by 0.003301. Notably, the GAM models only perform slightly better than the linear models. This aligns with the data exploration, in Figure 6 which suggested potential improvements with a non-linear approach, but also showed that the linear fits were already very good. Thus, it is not surprising that the linear models still perform very well comparatively. Overall, each model achieves very similar metrics.

With the concerns of overfitting stamped out, it is reasonable to proceed with the uncertainty comparison with just the full linear model and the full generalised additive models.

### 7.1.1 Results for using BSA Instead of Height and Weight Separately

The GAM model using Body Surface Area (BSA) as a variable achieved the same adjusted R-squared as the models that use height and weight separately. Since the BSA model is not nested within any of the height and weight models, their AIC and BIC values are not directly comparable.

Table 8: In-sample and out-of-sample RMSE and MAE results comparing the BSA model and the full GAM using height and weight.

| Eval. Metric | Sample | BSA Model | Height & Weight |
|---|---|---|---|
| **RMSE (mm)** | In-sample | 1.769331 | **1.752673** |
| | K-fold (Avg.) | 1.770114 | **1.754455** |
| | UMCG Echo | **2.139797** | 2.315956 |
| | Lopez Echo | 1.369318 | **1.358851** |
| **MAE (mm)** | In-sample | 1.386055 | **1.374275** |
| | K-fold (Avg.) | 1.386709 | **1.375803** |
| | UMCG Echo | **1.700905** | 1.851335 |
| | Lopez Echo | 1.088719 | **1.080442** |

On average, the differences between the evaluation metrics were expected to be subtle. The full GAM with height and weight performs better in every case, except on the UMCG echo data. While this is surprising, it could be due to the small size of this dataset, which may be affecting the validity of the finding.

Table 9: RMSE and MAE results for extreme subgroups for the BSA model and the full GAM using height and weight.

| Group | Avg. RMSE (mm) | | Avg. MAE (mm) | |
|---|---|---|---|---|
| | BSA | Height & Weight | BSA | Height & Weight |
| Obese Individuals | 1.879414 | **1.864358** | 1.471787 | **1.460621** |
| Very Short Adults | 1.894769 | **1.819279** | 1.493917 | **1.450046** |
| Older Females | 1.518651 | **1.489364** | 1.174030 | **1.172177** |
| Older Males | 1.574918 | **1.570000** | 1.262424 | **1.262003** |

The differences between the models are not as significant as expected. The full GAM with height and weight performs better consistently. The largest improvement – that is still a very small difference in actuality – is observed for the very short adults group, with a difference of approximately 0.08mm for the average RMSE and approximately 0.04mm for the average MAE.

Using a model of the residuals for each model, as specified in equation 18, the conditional standard deviations were plotted against the continuous variables and split by sex.

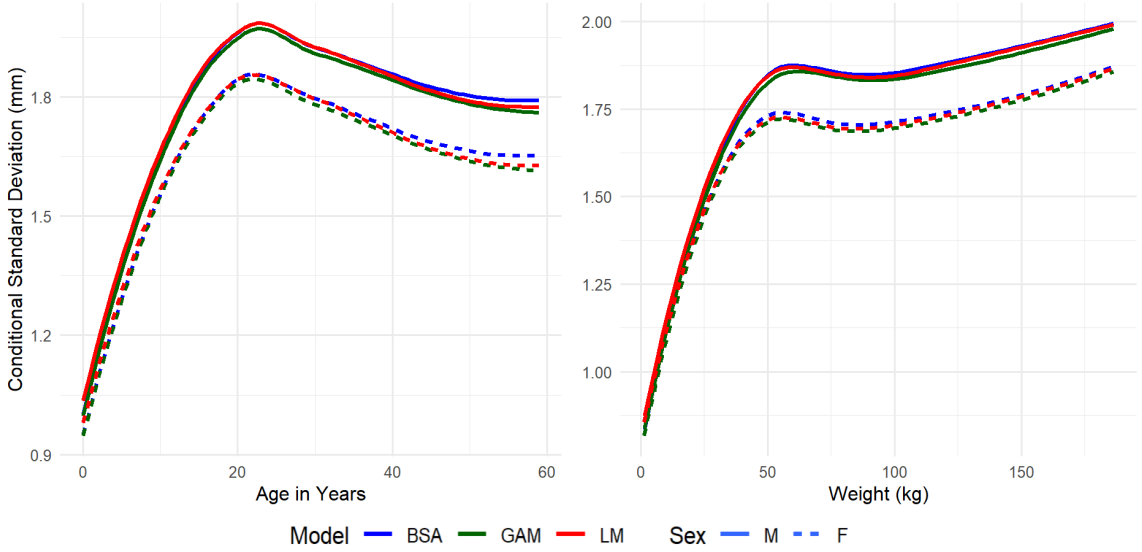



Figure 25: The conditional standard deviation plotted against age and weight and split by sex.

When looking at the standard deviation with respect to age and weight in Figure 25, the standard devation of the full GAM is always consistently below the GAM using BSA and the full LM.

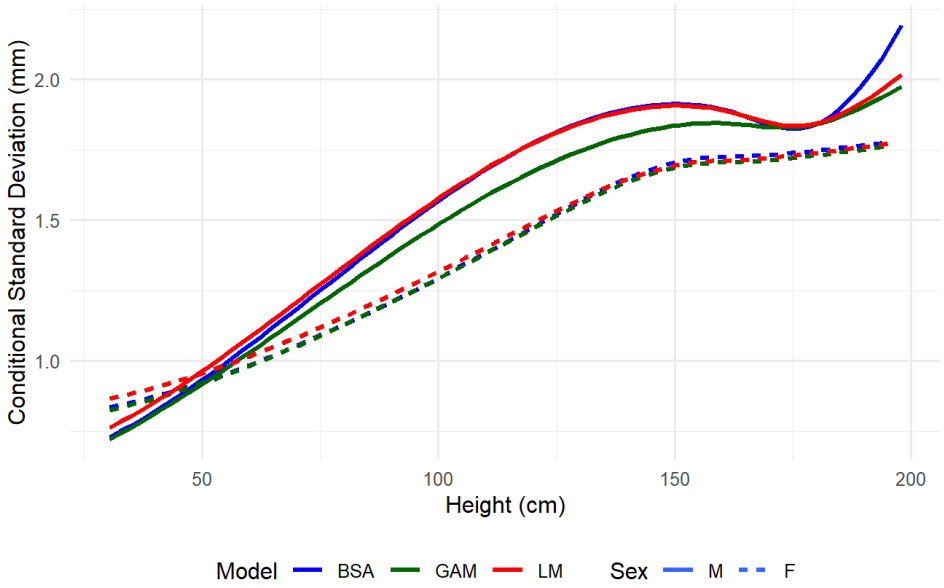
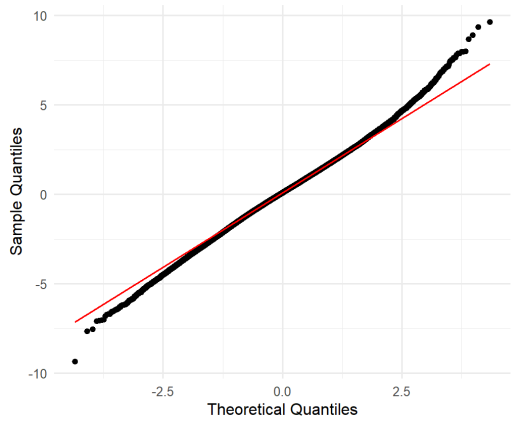



Figure 26: The conditional standard deviation plotted against the height and split by sex.

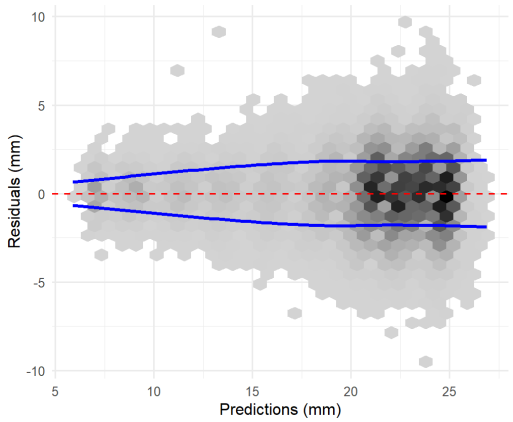
The standard deviation with respect to height in Figure 26 reveals something interesting: the GAM using BSA shows a dramatic increase in standard deviation after around 160cm for males. In this plot the full GAM demonstrates the most visible improvement, seen for males.

From the goodness-of-fit measures, evaluation metrics, and the estimated uncertainty plots, the best overall model is the full GAM. Thus, this is the model that will be used going forward.

Below are the residuals plotted against the predictions as well as a QQ-plot, seen in Figure 27.




(a) **QQ-plot for the residuals.** The residuals are computed on the original scale. The QQ-line is in red.




(b) **Residuals versus predictions plot.** The estimated standard deviation with respect to the predictions is given in blue.

Figure 27: Residual plots for the full GAM on the donor data.

The QQ-plot looks very good, showing only very slight skewness away from the QQ-line towards both ends. This implies that the residuals do not exhibit significant deviations or outliers, and there is little evidence of skewness or kurtosis. This indicates that the residuals of the model are approximately normally distributed, which validates the model's fit and robustness. The residuals versus the predictions also look quite good, though this plot does seem to indicate towards some heteroscedasticity. This can be further investigated by plotting the residuals against the predictors, as is done in Figure 28.

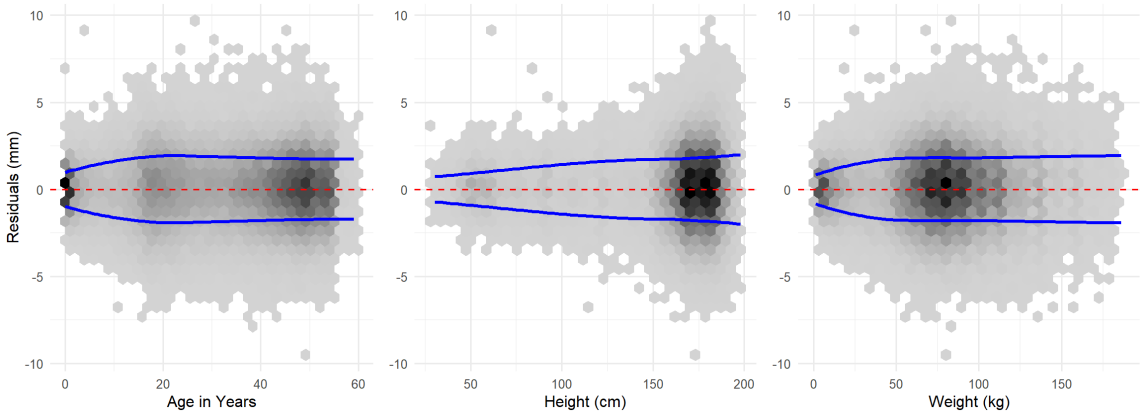



Figure 28: Plots of the residuals against the three continuous predictors for the full GAM applied to the donor data.

When looking at the residuals plotted against the predictors, it initially seems there is quite heavy heteroscedasticity in the plot for height. However, notice that the conditional standard deviations – modelled with respect to the plotted variable only – show that the vast majority of the points are centred around zero. Overall, the residuals of the full GAM appear reasonable. There is some heteroscedasticity present, particularly notable when looking at the standard deviation in the plot of the residuals with respect to age. Given the wide demographic variability of the dataset, this is unsurprising. It would be expected that for very low ages there is minimal variability, which increases significantly after puberty, and then reduces somewhat in adulthood, as discussed in the Theory section. And this is indeed what is seen in the residual plot and standard deviation plot

(Figure 25). The conditional standard deviation plots show slightly different shapes since those are partial effect plots of the standard deviation modelled on all variables, isolating the effect of each predictor. Lastly, the residuals versus the dissection date are shown below in Figure 29.

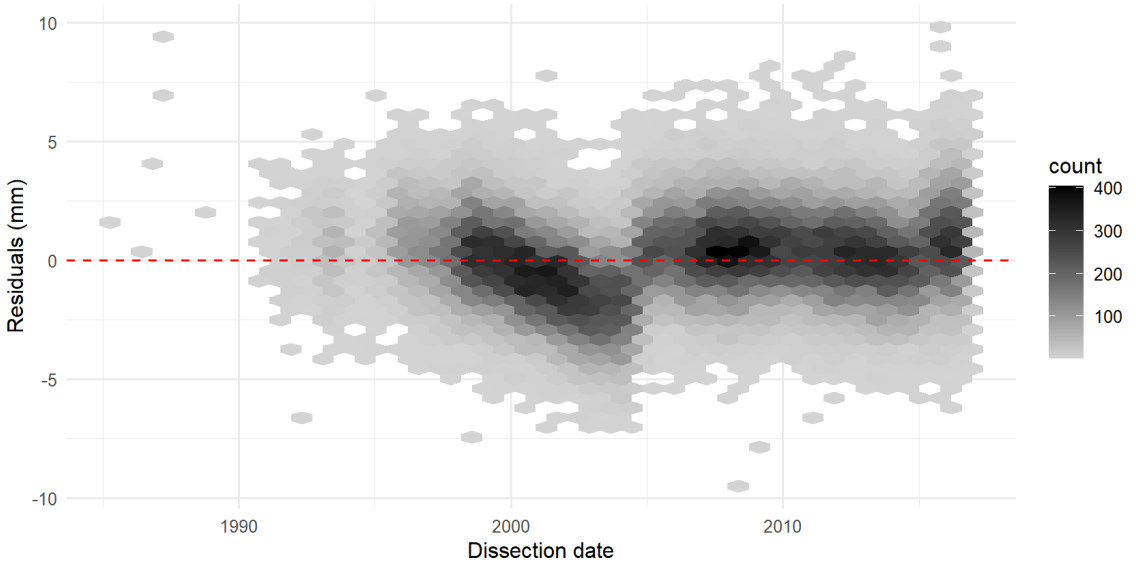


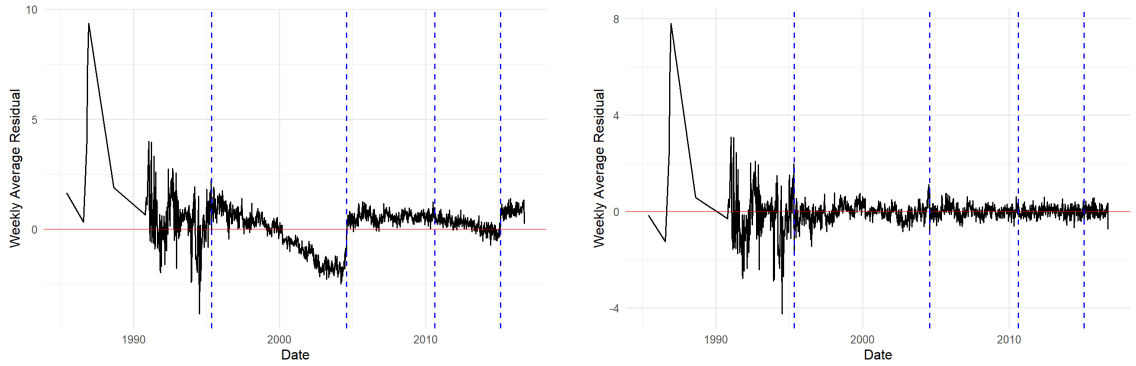

Figure 29: Plot of the residuals against the dissection date for the full GAM applied to the donor data.

The residuals over time clearly exhibit a similar negative trend to what was seen in the AV over time from the exploration. In fact, the trend seen here looks more significant and the possible changepoints look more obvious than the trend seen in the AV diameter measurements (Figure 7). There is also an indication of other plausible changepoints, which will be explored further in the next section.

## 7.2 Outcome of Trend Correction

The use of a tailored SN algorithm and the correction technique, as specified in section 6.2, improved the trends seen in the residuals. This is demonstrated below with figures of the weekly average of the residuals over time before and after corrections have been applied to each segment identified.

In Figure 30, it can be seen that four changepoints were found. These are located at the dates 1995/05/07, 2004/08/01, 2010/08/15, and 2015/02/08.




(a) **Before correction.** (b) **After correction.**

Figure 30: The plots depict the residuals aggregated by taking a weekly average, the four changepoints in blue dashed lines, and a horizontal line is plotted at zero in red.

A visible difference can be observed for each segment. The initial trend observed, the negative trend between 1995/05/07 and 2004/08/01, looks much improved. There other changepoints were not as clearly visible in the residuals (Figure 29), but after being objectively identified and corrected it has resulted in a noticeable improvement. There is a lot of variation seen before the first changepoint at 1995/05/07, this is due to the much smaller amount of data before this point. The algorithm does not identify any changepoints in this first segment, this is a positive sign as it implies the penalty is working as it should i.e. applying lots of small corrections in a region with sparse data would not be appropriate.

The residuals against the dissection date after the trend correction process are shown in Figure 31. This can be compared to Figure 29, depicting the residuals before the correction. Visually, it looks as though all of the identified trends have been eliminated entirely. The residuals now look very well-distributed around zero. It is also wise to take another look at the other residual plots for the trend-corrected model.

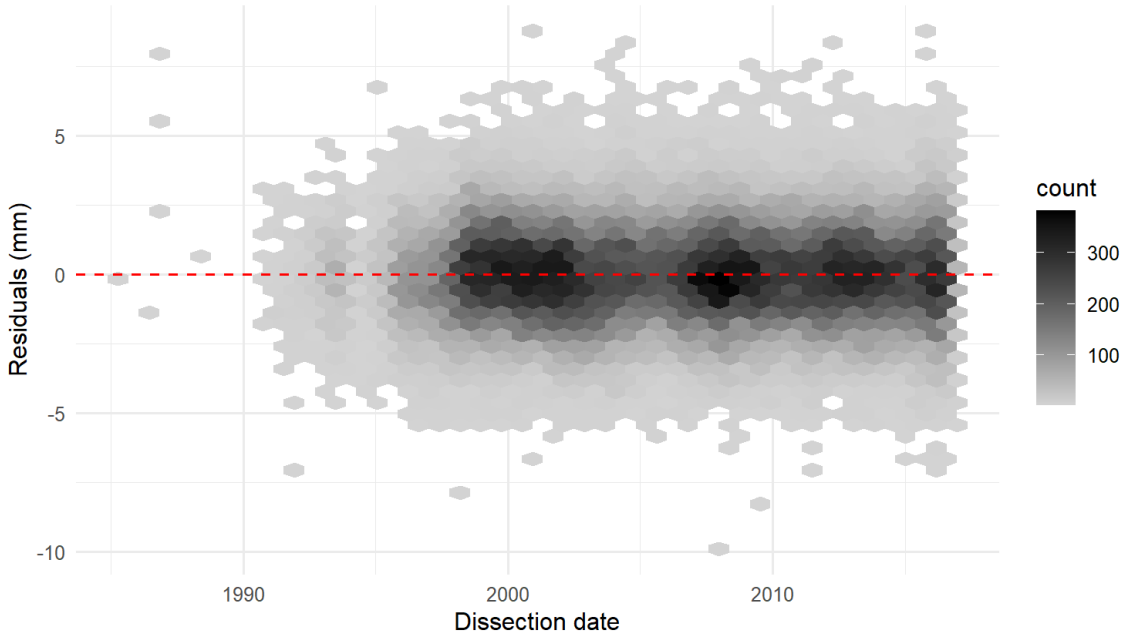
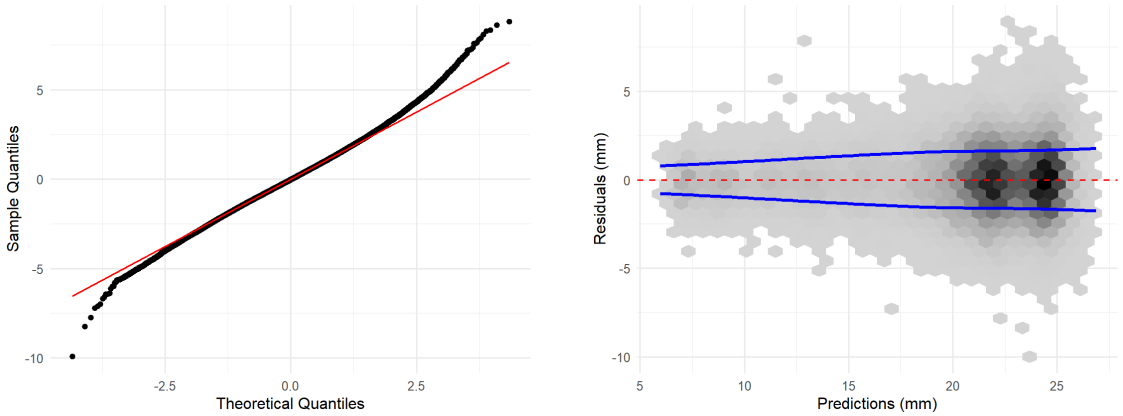



Figure 31: Plot of the residuals against the dissection date for the full GAM applied to the donor data after the trend correction process.

From Figure 32a, the QQ-plot looks slightly better when compared to the plot of the full, uncorrected GAM (Figure 27a). There is less deviation from the QQ-line seen at the upper end. The residuals versus predictions plot looks indistinguishable from Figure 27b. When plotting the residuals with respect to the predictors, the same can be seen i.e. they are indistinguishable from Figure 28.




(a) **QQ-plot for the residuals.** The residuals are computed on the original scale. The QQ-line is in red.

(b) **Residuals versus predictions plot.** The estimated standard deviation with respect to the predictions is given in blue.

Figure 32: Residual plots for the trend-corrected model on the trend-corrected donor data.

The R squared value increased to 0.922 from 0.904. What remains to be seen is how the model now performs with respect to the evaluation metrics, these are shown in Table 10.

Table 10: **Trend correction.** In-sample and out-of-sample RMSE and MAE results comparing the full GAM before and after the correction.

| Eval. Metric | Sample | Before Correction | After Correction |
|---|---|---|---|
| **RMSE (mm)** | In-sample | 1.752673 | **1.579097** |
| | K-fold (Avg.) | 1.754455 | **1.580592** |
| | UMCG Echo | **2.315956** | 2.393776 |
| | Lopez Echo | 1.358851 | **1.306000** |
| **MAE (mm)** | In-sample | 1.374275 | **1.235021** |
| | K-fold (Avg.) | 1.375803 | **1.236103** |
| | UMCG Echo | **1.851335** | 1.917620 |
| | Lopez Echo | 1.080442 | **1.022533** |

The in-sample and K-fold evaluation metrics show quite a significant improvement of about 0.18mm in the RMSE and 0.14mm in the MAE. There is an improvement seen in the Lopez echo data. For the UMCG echo data the model before correction is performing slightly better. One of the goals was that the model after the trend correction would generalise better. Overall, this seems to have been achieved since there is an improvement for the performance on the Lopez data and there is no notable deterioration on the UMCG data.

Considering the significant improvement in the residuals over time and the evaluation metrics, the final models will include a trend correction.

## 7.3 Effects of Echo Bias Correction

The results of the Echo correction has been separated into four sections: the statistical significance, the adults, those 18 and under, and combining the two corrections after having found pseudo-echo AV diameters for each subset.

### 7.3.1 Statistical Significance of an Echo Bias

To investigate whether there is a statistically significant difference between the physical and echocardiographic measurements, the method used is described in Section 6.3.

It was found that the binary predictor for the source of the data was significant with a p-value smaller than 0.001 for both the UMCG and Lopez *et al.* datasets. The summary of the results can be seen in Tables 15 and 16.

### 7.3.2 The Results for Adults

Quantifying the difference in adults was straightforward. Using the full GAM trained on a subset of the donor data, predictions for the UMCG data were made. This difference was then modeled via a full linear model. A GAM was not used due to the limited amount of data, a nonlinear approach would overfit, as was indicated to in the exploration in Figure 17.

The difference of the donor model pseudo-physical AV predictions and the UMCG echo AV diameter measurements with respect to the predictors are plotted below in Figure 33.

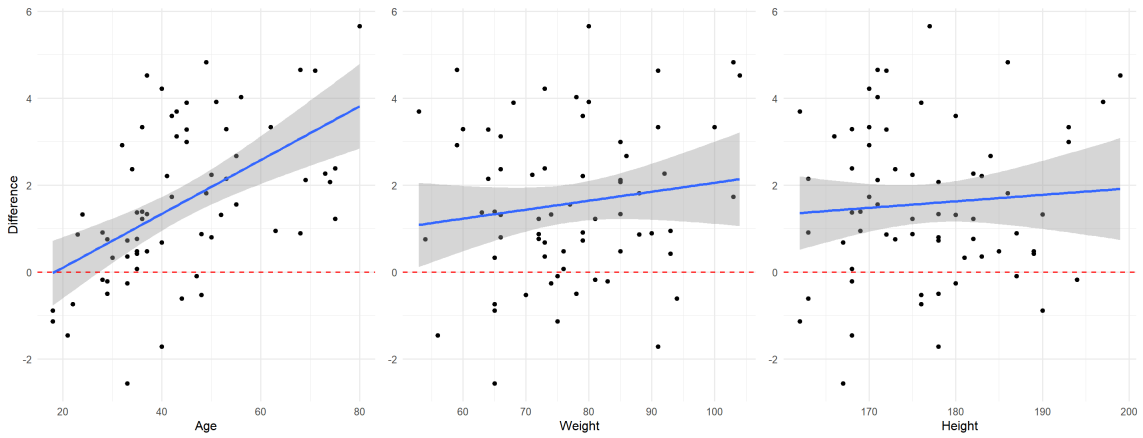



Figure 33: Plots of the difference between the donor model predictions and the UMCG AV diameter measurements against the variables. A linear fit is applied to the data in blue, and the shaded area represents the confidence interval.

Despite the limited data, there does seem to be a plausible indication of a consistent echo bias. In both height and weight, the donor measurements appear to be consistently larger than the echo estimates by just under 2mm on average. There is a visible positive gradient seen in the difference against age. This evidence towards age being included in the bridging model. It should be noted that the donor model is trained on data that has a maximum age of 59, thus making AV diameter predictions up to age 80 is quite a large extrapolation.

An echo correction is applied to the adult subset of the donor data. However, the reliability of this echo bias estimate is questionable since this difference is based on a very small dataset.

### 7.3.3 The Results for Individuals 18 and Under

As described in the methodology section 6.3, two different methods were discussed for those 18 and under. These two methods will continue to be referred to as methods (1) and (2).

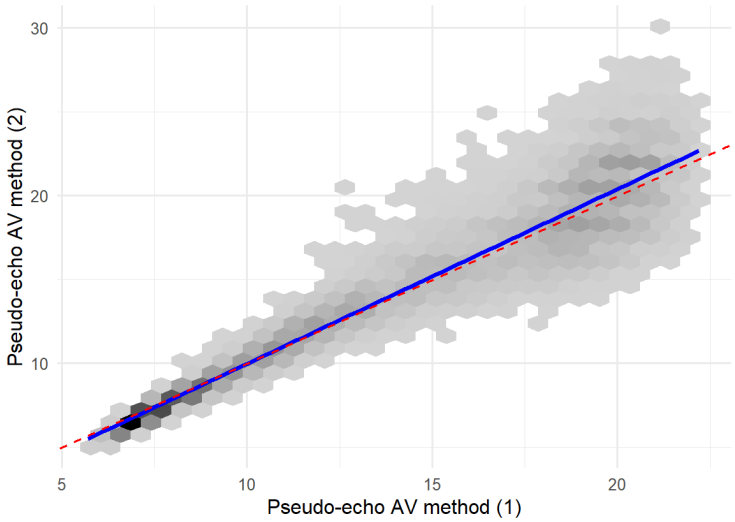



Figure 34: Plots of the pseudo-echo AV diameters for a subset of the donor data. In red is the line $y = x$, and in blue is a linear fit to the relationship.

From Figure 34 it can be seen that, if the Lopez data was not available and only the model was, method (1) would have been a very close approximation of creating a bridging model using the data itself. The linear fit to the data, in blue, is extremely close to the line $y = x$. This shows that, for the demographic studied by Lopez *et al.*, their model is similarly representative of echo AV diameters measurements as our model is for the physical AV diameters of the donor data. However, notice that there is cone-shaped variance around the linear fit i.e. on average the two methods are very close but there is increasing variability as the AV diameters get larger.

Using method (2), the relationship between the bias and the predictors has been visualised below:

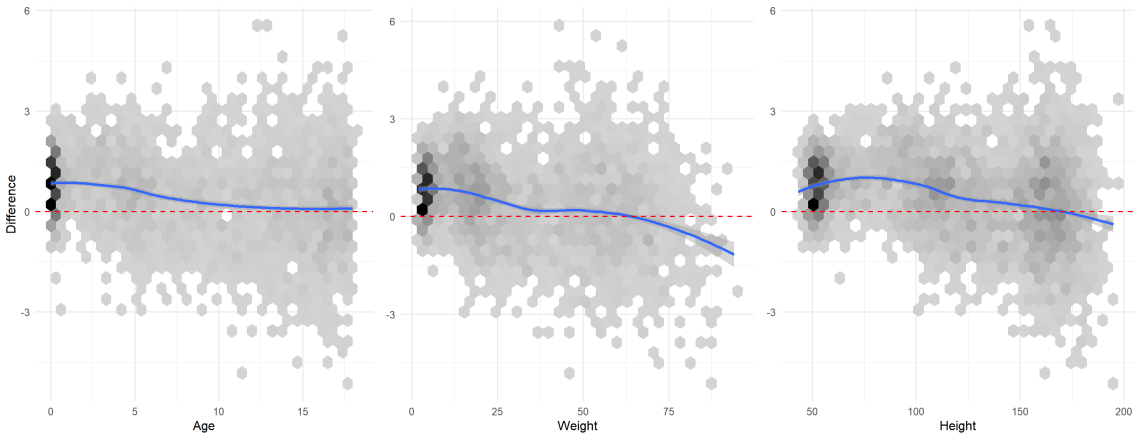



Figure 35: Plots of the difference between the donor model predictions and the Lopez AV diameter measurements against the variables. A LOESS fit is applied to the data in blue.

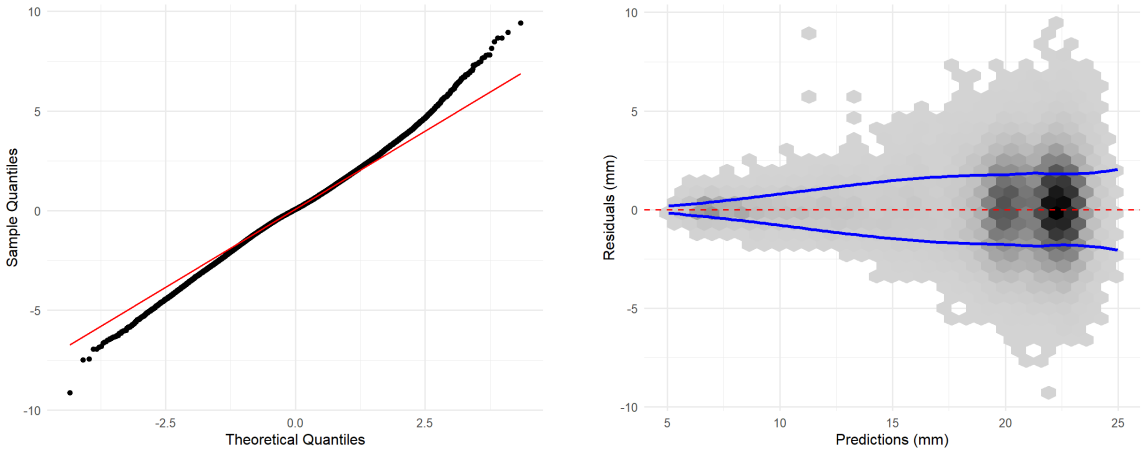
Figure 35 matches what would be expected from the exploration of the difference. Overall, the bias does not look very large. Notably, donor measurements are approximately 1mm larger than echo estimates for the very low end of age and weight, and approximately 1mm smaller for the higher end of weight. This is in contrast to the relationship seen for adults, where the bias increases with respect to all the predictors. This could result in disjointed AV diameter measurements when both bias correction models are applied if too much flexibility is used for the bridging models.

This correction will be applied, however it would not be unreasonable to conclude that this bias is not clinically significant i.e. it is not greater than 2mm on average (based on expert opinion).

### 7.3.4 The Combined Echo Corrections and Resulting Model

Once pseudo-echo AV diameters were obtained for both adults and individuals 18 and under, the donor subsets could be recombined and the best model structure was retrained on this data. This results in a model for pseudo-echo AV diameters for a very large demographic. It should be noted that this now includes an extrapolation for weights larger than 100kg.

The R squared of the echo-corrected model increased to a value of 9.22. Next, the residuals were plotted as usual in Figure 36.




(a) **QQ-plot for the residuals.** The residuals are computed on the original scale. The QQ-line is in red.

(b) **Residuals versus predictions plot.** The estimated standard deviation with respect to the predictions is given in blue.

Figure 36: Residual plots for the full GAM on the donor data after the echo bias corrections.

From the QQ-plot (Fig. 36a) the residuals look closer to normally distributed as there seems to be less deviation from the QQ-line (when compared to the full GAM without correction). However, in Figure 36b, there appear to be a few new outliers introduced when looking at the predictions between 10mm and 15mm. Moreover, the overall shape has noticeably changed. There is less variation for lower predicted values leading to a more accentuated cone shape.

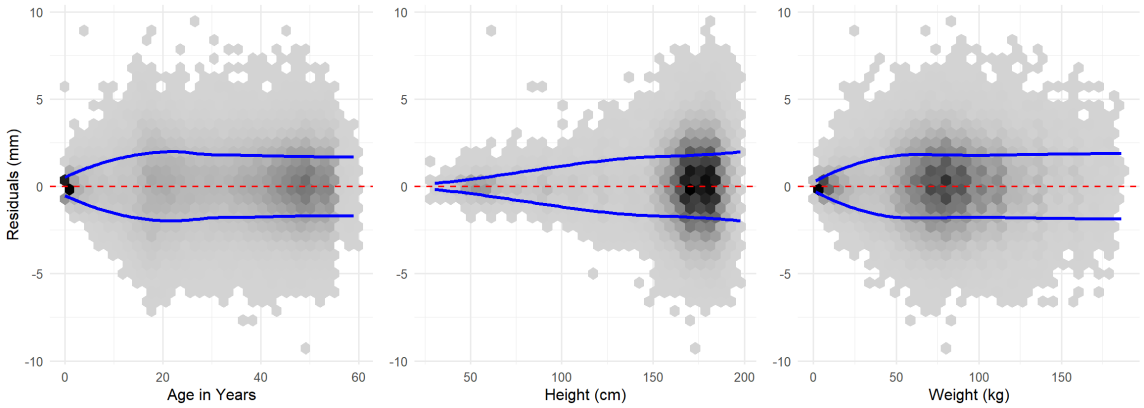



Figure 37: Plots of the residuals against the three continuous predictors for the full GAM applied to the donor data after the echo bias corrections.

When potting the residuals against the predictors in Figure 37, it can be seen that the shape of the variation of the residuals is more pronounced for all three continuous predictors. Overall, it can be seen that after the echo correction there is more deviation from the homoscedasticity assumption, further motivating the necessity of using a conditional standard deviation for the Z-score computations. Lastly, notice that there is slightly more bulging in the residuals at around age 18 after the combined echo correction. This is likely due to the separate echo corrections applied to adults and individuals under 18.

Table 11: **Echocardiographic bias correction.** In-sample and out-of-sample RMSE and MAE results comparing the full GAM before and after the correction.

| Eval. Metric | Sample | Before Correction | After Correction |
|---|---|---|---|
| **RMSE (mm)** | In-sample | 1.752673 | **1.721622** |
| | K-fold (Avg.) | 1.754455 | **1.722995** |
| | UMCG Echo | 2.315956 | **1.509165** |
| | Lopez Echo | 1.358851 | **1.206946** |
| **MAE (mm)** | In-sample | 1.374275 | **1.328663** |
| | K-fold (Avg.) | 1.375803 | **1.329901** |
| | UMCG Echo | 1.851335 | **1.254647** |
| | Lopez Echo | 1.080442 | **0.919731** |

The evaluation metrics in Table 11 match what would be expected. Since the corrections were based on the UMCG and Lopez datasets, the evaluation metrics improved significantly on those samples. This does indicate that there is an improvement on echo data. However, it is important to note that these are no longer validation datasets as they are not independent from the data the model is trained on. It is also noteworthy that the in-sample and K-fold evaluations improved.

## 7.4 The Final Models and a Comparison to the Established Models

As discussed in Theory section 5.3.4, two separate models are required to predict the healthy AV diameter: one for diagnosis and another for aortic valve replacement (AVR) surgery. In both cases, the full generalised additive model is used. The final models cannot reasonably be written out explicitly with the estimated coefficients since there are over 400 terms. For the clinical implementation, a web tool will be developed. An example of how this web tool may look in practice is given in Figure 49. Additionally, the summary tables can be found in the Appendix and are referenced in the text.

### A Model to Predict Healthy Aortic Valve Diameters for AVR Surgery

The model used to predict healthy AV diameters for AVR surgery is the full GAM with the trend correction applied. The summary is given in Table 17. The evaluation metrics compared to the two most commonly used models are given below.

Table 12: **Model for AVR surgery.** RMSE and MAE results comparing the full GAM trend-corrected to the Lopez *et al.* and Cantinotti *et al.* models.

| Eval. Metric | Sample | AVR Model | Lopez | Cantinotti |
|---|---|---|---|---|
| **RMSE (mm)** | Donor | **1.579097** | 2.564206 | 2.114441 |
| | UMCG Echo | 2.393776 | **1.770278** | 1.941664 |
| | Lopez Echo | 1.306000 | **1.275022** | 1.481374 |
| **MAE (mm)** | Donor | **1.235021** | 2.043886 | 1.652006 |
| | UMCG Echo | 1.917620 | **1.491492** | 1.500794 |
| | Lopez Echo | 1.022533 | **0.977995** | 1.127006 |

Table 12 demonstrates that, for the donor data, the AVR model surpasses the Lopez and Cantinotti models by 0.98mm and 0.53mm, respectively, in terms of RMSE. The Lopez model, however, performs better on both the UMCG and Lopez echo datasets. While the Lopez model only slightly outperforms the AVR model on the Lopez dataset, it significantly outperforms it on the UMCG dataset. Given that both datasets utilise echo measurements, this result is not too unexpected. However, the Lopez model's superior performance on the UMCG data is noteworthy, as it is a significant extrapolation beyond the age range it was trained on, making this outcome somewhat surprising.

In this case we are most concerned with how it performs on the donor data, since this model will be used to predict the physical AV diameter needed for the aortic valve replacement. Lets compare the estimated uncertainty of each model with respect to the predictors, in Figure 38.

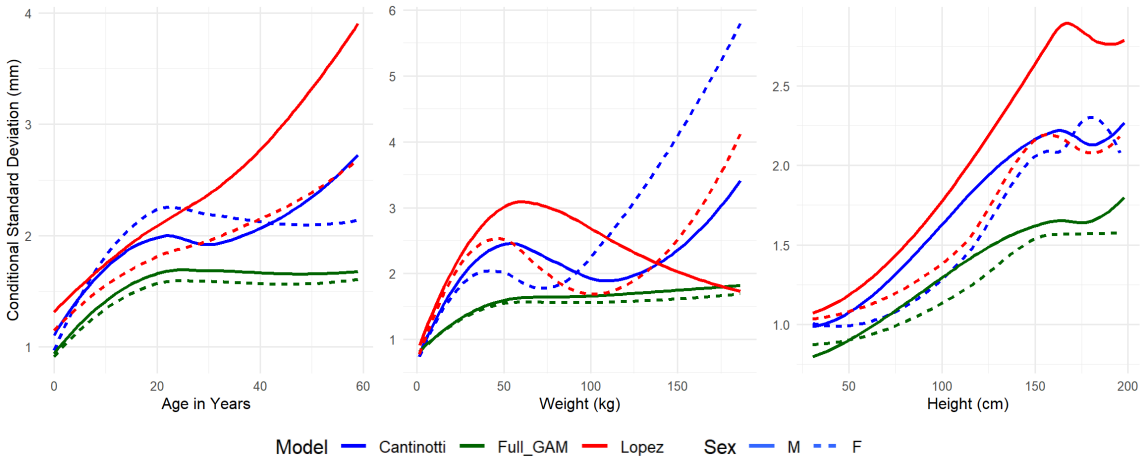



Figure 38: **Conditional standard deviations applied to the trend-corrected donor data.** Plots of the standard deviations against the three continuous predictors.

The final donor model trend-corrected massively outperforms the Lopez and Cantinotti models in terms of certainty with respect to each predictor. The Lopez model has exponentially increasing uncertainty with respect to age, it is especially drastic for males. The Cantinotti model shows a similar exponential behaviour in the standard deviation after a weight of about 100kg for females and under 75kg for males. These behaviours make sense since – for both the age and weight – these

are massive extrapolations for the Lopez and Cantinotti models. Nonetheless, even within the demographics that the Cantinotti and Lopez models are trained on, the donor model consistently demonstrates better certainty. Interestingly, when looking at the standard deviation against weight, the Lopez model depicts very different behaviour for the different sexes. Notice that the standard deviation remains very similar for both sexes for the final donor model, but not for the other models. This is because the donor model takes sex into account.

Overall, the trend-corrected, full GAM is the better performing and more suitable model to predict healthy aortic valve diameters for aortic valve replacement surgery.

**Separate Models for Diagnosis That Incorporate the Echo Correction**

The model to predict the healthy average AV diameter and the variance are full GAMs and incorporate both the trend correction *and* echocardiographic bias correction. First the trend correction is applied, and then the echo correction. The model summaries for the AV diameter prediction and the variance are given in Tables 18 and 19.

As was done for the previous model, the evaluation metrics of the diagnostic model are compared to the Lopez *et al.* and Cantinotti *et al.* models in Table 13.

Table 13: **Model for Diagnosis.** RMSE and MAE results comparing the full GAM including the trend and echo bias correction to the Lopez *et al.* and Cantinotti *et al.* models.

| Eval. Metric | Sample | Diagnostic Model | Lopez | Cantinotti |
|---|---|---|---|---|
| **RMSE (mm)** | Donor (pseudo-echo) | **1.550151** | 1.940521 | 2.198464 |
| | UMCG Echo | **1.493045** | 1.770278 | 1.941664 |
| | Lopez Echo | **1.205917** | 1.275022 | 1.481374 |
| **MAE (mm)** | Donor (pseudo-echo) | **1.195610** | 1.502287 | 1.698825 |
| | UMCG Echo | **1.236335** | 1.491492 | 1.500794 |
| | Lopez Echo | **0.918554** | 0.977995 | 1.127006 |

The diagnostic donor model outperforms when tested on every dataset. As mentioned before, since the corrections were applied using the UMCG and Lopez datasets, these are no longer fully independent validation datasets for the donor model. However, it is still notable that the diagnostic model now outperforms the Lopez model on the Lopez dataset. Although, the model without an echo correction already performed comparably to the Lopez model on the Lopez data.

As was seen with the model for AVR, the diagnostic model achieves better estimated uncertainty with respect to all the variables when compared to the Lopez and Cantinotti. The plots can be seen in Figure 39. While the diagnostic model is outperforming, it is also noteworthy that the Lopez and Cantinotti models are performing better on the pseudo-echo donor data when compared to Figure 38, before the echo correction is applied. This is also reflected in the evaluation metrics.

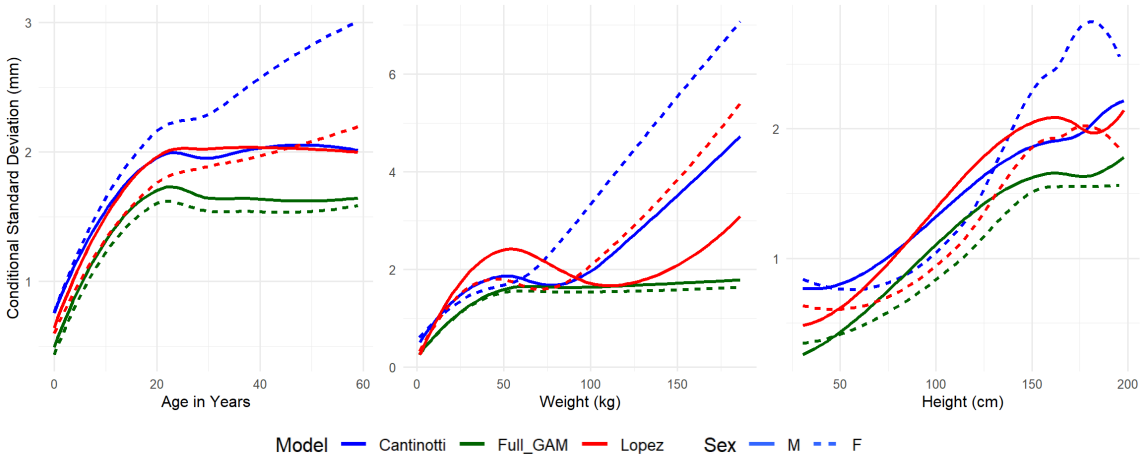



Figure 39: **Conditional standard deviations applied to the trend and echo-corrected donor data.** Plots of the standard deviations against the three continuous predictors.

While it is definitely worthwhile to see how much better the donor model performs for a wider demographic, comparing the models on the entire donor dataset goes well beyond the scope they are trained on. In this case it is also useful to check the conditional standard deviations on the Lopez *et al.* dataset. This is a way to analyse and compare performance on true echo measurements, and within the same demographics that the Lopez and Cantinotti models were trained on.

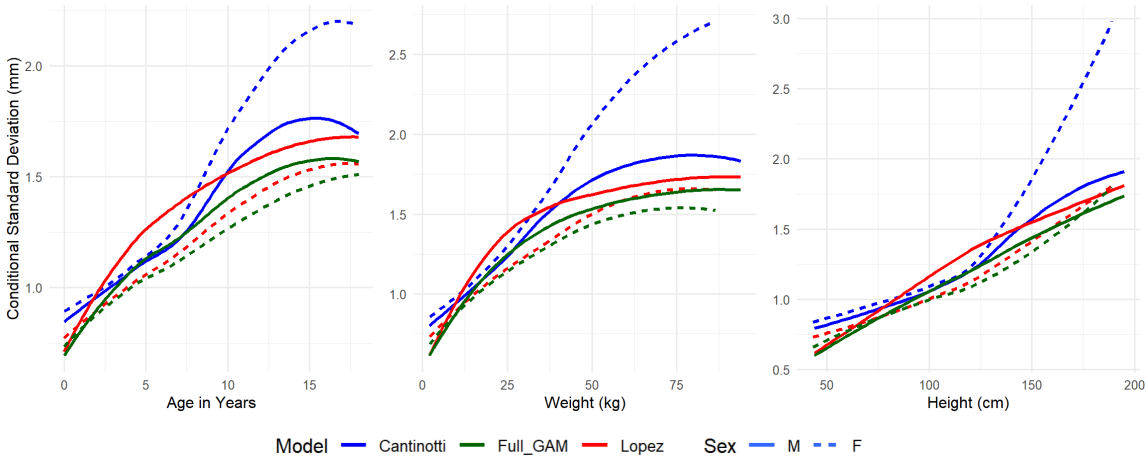



Figure 40: **Conditional standard deviations applied to the Lopez data.** Plots of the standard deviations against the three continuous predictors.

Figure 40 shows that the diagnostic donor model and the Lopez *et al.* model perform similarly, however the standard deviation of the donor model remains consistently below in each plot. Again, the Lopez and Cantinotti models demonstrate different behaviour for males and females, especially the Cantinotti model which has increasingly larger standard deviations for females.

Ultimately, it could be argued either way whether the echocardiographic bias correction is most appropriate and necessary. Since the UMCG data is so limited, any performance metrics on this dataset should be taken with a pinch of salt. Thus, the lack-luster performance of the uncorrected donor model could be reasonably disregarded. The performance of the donor model on the Lopez data is very good regardless of whether an echo bias correction is included or not. Having said that, the performance of the echo-corrected donor model on the Lopez data outperforms the Lopez model itself. Thus, a case could certainly be made to include the correction.

# 8 Conclusion

This study aimed to develop a refined model for aortic valve (AV) diameter measurements that accounts for demographic variability and measurement technique discrepancies. The goal was to enhance predictive accuracy and uncertainty quantification in AV diameter measurements, particularly for extreme demographic cases. In the case of diagnosis, this would allow for more accurate Z-score computations. In preparation for surgery, this would result in more accurate healthy AV diameter predictions when choosing a prosthetic for aortic valve replacement (AVR).

The results demonstrate that the full generalised additive model (GAM) with height and weight consistently performs better than other models, including the linear model (LM) and the GAM using body surface area (BSA). Explicitly, the full GAM – including tensor product smooth interaction terms – shows the best performance in terms of goodness-of-fit measures and evaluation metrics, as well as in the estimated uncertainty plots. Although, the improvements are modest in clinical terms.

Using a tailored segment neighbourhood (SN) algorithm, trend corrections were applied to the AV diameter measurements through an iterative changepoint analysis via the residuals. Significant improvements were observed in the in-sample evaluation metrics and the residuals, particularly for the negative trend observed in the residuals between 1995/05/07 and 2004/08/01. Furthermore, using an objective algorithm led to the identification of changepoints and trends that would have otherwise gone undetected. Although the echocardiographic datasets – namely the UMCG and Lopez et al. datasets – used as validation did not show improvements, there was no notable deterioration in performance, supporting the inclusion of a trend correction in the final models. Additionally, implementing the SN algorithm myself developed a better understanding of how it works and how it can be best implemented. Despite finding a package towards the end of my project that would likely have been sufficient, this experience was very valuable.

For diagnostic purposes, the donor model that incorporated an echocardiographic bias correction consistently demonstrated better performance across all three datasets when compared to the Lopez and Cantinotti models. It was noted that the performance of the donor model on the Lopez data, regardless of whether an echo bias correction is included, was very good. However, the echo-corrected donor model outperformed on the Lopez data when compared to the Lopez model itself, making a case for including this correction.

With regards to future research, a larger dataset of echocardiograpic AV diameter measurements spanning both adults and individuals under 18 would be ideal to better investigate whether there is a bias between physical and echo measurements. This would allow for a more comprehensive analysis as to whether the donor dataset can be reliably applied for diagnosis. Moreover, the Cryolife donor dataset also includes pulmonary valve measurements, these were not explored and a predictive model could be made for these as well.

This research was conducted in collaboration with my supervisors, Prof. M.A. Grzegorczyk and Prof. G.A. Lunter, as well as specialists in the field of Cardiology Prof. T. Ebels and Dr J. P. van Melle. Their expertise were invaluable in this study and its outcomes. We are currently in the process of writing a paper and developing an accompanying web tool that will present our findings and final models in an accessible and practical format. We aim (and hope) to publish this work soon.

Overall, the models to predict AV diameters developed in this study offer significant improvements in predictive accuracy and uncertainty. For the first time, the uncertainty – that includes the natural variation seen in wide demographics – has been modelled. These things combined are crucial for clinical applications, allowing for more reliable diagnosis and prediction of healthy AV diameter for AVR surgery.

# A Appendix: Tables

| Parametric Coefficients | | | | |
|---|---|---|---|---|
| **Term** | **Estimate** | **Std. Error** | **t value** | **Pr($>\|t\|$)** |
| (Intercept) | 3.067245 | 0.007983 | 384.22 | $< 2e-16^{***}$ |
| Sex1 | -0.061782 | 0.001794 | -34.43 | $< 2e-16^{***}$ |
| **Approximate Significance of Smooth Terms** | | | | |
| **Smooth Term** | **edf** | **Ref.df** | **F** | **p-value** |
| s(logAge) | 13.170 | 14 | 59.998 | $< 2e-16^{***}$ |
| s(logAge):Sex0 | 2.645 | 14 | 4.256 | $< 2e-16^{***}$ |
| s(logAge):Sex1 | 0.055 | 14 | 0.004 | $0.00243^{**}$ |
| s(logWeight) | 7.182 | 14 | 8.045 | $< 2e-16^{***}$ |
| s(logWeight):Sex0 | 1.622 | 14 | 0.915 | $2.74e-05^{***}$ |
| s(logWeight):Sex1 | 0.002 | 14 | 0.000 | $0.01063^{*}$ |
| s(sqrtHeight) | 9.446 | 14 | 71.284 | $< 2e-16^{***}$ |
| s(sqrtHeight):Sex0 | 0.000 | 14 | 0.000 | 0.42787 |
| s(sqrtHeight):Sex1 | 0.001 | 14 | 0.000 | 0.38667 |
| ti(logWeight,sqrtHeight) | 2.257 | 16 | 0.342 | $0.00564^{**}$ |
| ti(logWeight,sqrtHeight):Sex0 | 0.919 | 16 | 0.075 | $0.02285^{*}$ |
| ti(logWeight,sqrtHeight):Sex1 | 0.000 | 16 | 0.000 | 0.16863 |
| ti(sqrtHeight,logAge) | 0.000 | 16 | 0.000 | 0.48991 |
| ti(sqrtHeight,logAge):Sex0 | 0.000 | 16 | 0.000 | 0.37306 |
| ti(sqrtHeight,logAge):Sex1 | 0.000 | 16 | 0.000 | 0.41644 |
| ti(logWeight,logAge) | 8.298 | 16 | 4.593 | $< 2e-16^{***}$ |
| ti(logWeight,logAge):Sex0 | 0.617 | 16 | 0.049 | 0.06737 |
| ti(logWeight,logAge):Sex1 | 0.000 | 16 | 0.000 | 0.43634 |
| ti(logWeight,sqrtHeight,logAge) | 10.410 | 48 | 0.759 | $< 2e-16^{***}$ |
| ti(logWeight,sqrtHeight,logAge):Sex0 | 8.515 | 48 | 0.467 | $1.20e-05^{***}$ |
| ti(logWeight,sqrtHeight,logAge):Sex1 | 0.000 | 48 | 0.000 | 0.86403 |

Signif. codes: $0^{***}, 0.001^{**}, 0.01^{*}, 0.05$ .

R-sq.(adj) = 0.904, Deviance explained = 90.4%

fREML = -74868, Scale est. = 0.0071192, n = 71197

Table 14: Summary of the model fit for the **full GAM**.

| Parametric Coefficients | | | | |
|---|---|---|---|---|
| **Term** | **Estimate** | **Std. Error** | **t value** | **Pr($>\|t\|$)** |
| (Intercept) | 3.081974 | 0.009940 | 310.065 | $< 2e-16^{***}$ |
| Source1 | 0.071900 | 0.009929 | 7.241 | $4.5e-13^{***}$ |
| Sex1 | -0.067029 | 0.001903 | -35.230 | $< 2e-16^{***}$ |
| **Approximate Significance of Smooth Terms** | | | | |
| **Smooth Term** | **edf** | **Ref.df** | **F** | **p-value** |
| s(logAge) | 6.340 | 14 | 109.460 | $< 2e-16^{***}$ |
| s(logAge):Sex0 | 3.340 | 14 | 7.123 | $< 2e-16^{***}$ |
| s(logAge):Sex1 | 0.001 | 14 | 0.000 | $0.00267^{**}$ |
| s(logWeight) | 6.654 | 14 | 78.660 | $< 2e-16^{***}$ |
| s(logWeight):Sex0 | 1.781 | 14 | 1.490 | $2.28e-06^{***}$ |
| s(logWeight):Sex1 | 0.001 | 14 | 0.000 | $0.04736^{*}$ |
| s(sqrtHeight) | 0.002 | 14 | 0.000 | $0.00497^{**}$ |
| s(sqrtHeight):Sex0 | 4.229 | 14 | 52.278 | $< 2e-16^{***}$ |
| s(sqrtHeight):Sex1 | 5.470 | 14 | 39.509 | $< 2e-16^{***}$ |
| ti(logWeight,sqrtHeight) | 0.000 | 16 | 0.000 | 0.93434 |
| ti(logWeight,sqrtHeight):Sex0 | 3.293 | 16 | 0.691 | $0.00276^{**}$ |
| ti(logWeight,sqrtHeight):Sex1 | 0.001 | 16 | 0.000 | 0.39393 |
| ti(sqrtHeight,logAge) | 0.001 | 16 | 0.000 | 0.46799 |
| ti(sqrtHeight,logAge):Sex0 | 1.568 | 16 | 0.417 | $0.00499^{**}$ |
| ti(sqrtHeight,logAge):Sex1 | 0.001 | 16 | 0.000 | 0.53884 |
| ti(logWeight,logAge) | 3.641 | 16 | 2.137 | $< 2e-16^{***}$ |
| ti(logWeight,logAge):Sex0 | 3.313 | 16 | 0.512 | $0.00521^{**}$ |
| ti(logWeight,logAge):Sex1 | 0.001 | 16 | 0.000 | 0.23193 |
| ti(logWeight,sqrtHeight,logAge) | 0.001 | 64 | 0.000 | 0.37873 |
| ti(logWeight,sqrtHeight,logAge):Sex0 | 2.846 | 64 | 0.069 | 0.10763 |
| ti(logWeight,sqrtHeight,logAge):Sex1 | 2.990 | 64 | 0.069 | 0.10300 |

Signif. codes: $0^{***}, 0.001^{**}, 0.01^{*}, 0.05$ .

R-sq.(adj) = 0.447, Deviance explained = 44.8%

fREML = -64525, Scale est. = 0.0063243, n = 58058

Table 15: **Donor and UMCG source significance.** Summary of the full GAM fit with a binary variable for the source of the data.

| Parametric Coefficients | | | | |
|---|---|---|---|---|
| **Term** | **Estimate** | **Std. Error** | **t value** | **Pr($>\|t\|$)** |
| (Intercept) | 2.655905 | 0.017928 | 148.142 | $< 2e-16^{***}$ |
| Source2 | -0.049223 | 0.002082 | -23.638 | $< 2e-16^{***}$ |
| Sex1 | -0.039785 | 0.004022 | -9.891 | $< 2e-16^{***}$ |
| **Approximate Significance of Smooth Terms** | | | | |
| **Smooth Term** | **edf** | **Ref.df** | **F** | **p-value** |
| s(logAge) | 4.854 | 14 | 2.854 | $< 2e-16^{***}$ |
| s(logAge):Sex0 | 0.000 | 14 | 0.000 | 0.789741 |
| s(logAge):Sex1 | 0.000 | 14 | 0.000 | 0.758164 |
| s(logWeight) | 8.009 | 14 | 7.088 | $< 2e-16^{***}$ |
| s(logWeight):Sex0 | 0.000 | 14 | 0.000 | 0.581012 |
| s(logWeight):Sex1 | 0.000 | 14 | 0.000 | 0.566644 |
| s(sqrtHeight) | 3.497 | 14 | 2.256 | $< 2e-16^{***}$ |
| s(sqrtHeight):Sex0 | 0.000 | 14 | 0.000 | 0.896791 |
| s(sqrtHeight):Sex1 | 0.000 | 14 | 0.000 | 0.841144 |
| ti(logWeight,sqrtHeight) | 1.161 | 16 | 0.099 | $0.011658^{*}$ |
| ti(logWeight,sqrtHeight):Sex0 | 1.444 | 16 | 0.178 | $0.027875^{*}$ |
| ti(logWeight,sqrtHeight):Sex1 | 0.000 | 16 | 0.000 | 0.588503 |
| ti(sqrtHeight,logAge) | 4.035 | 16 | 1.072 | $< 2e-16^{***}$ |
| ti(sqrtHeight,logAge):Sex0 | 1.206 | 16 | 0.120 | $0.008301^{**}$ |
| ti(sqrtHeight,logAge):Sex1 | 0.001 | 16 | 0.000 | 0.197733 |
| ti(logWeight,logAge) | 0.000 | 16 | 0.000 | 0.746000 |
| ti(logWeight,logAge):Sex0 | 3.137 | 16 | 0.723 | $0.000153^{***}$ |
| ti(logWeight,logAge):Sex1 | 0.000 | 16 | 0.000 | 0.067495 |
| ti(logWeight,sqrtHeight,logAge) | 0.120 | 64 | 0.531 | $< 2e-16^{***}$ |
| ti(logWeight,sqrtHeight,logAge):Sex0 | 0.000 | 64 | 0.000 | 0.396452 |
| ti(logWeight,sqrtHeight,logAge):Sex1 | 0.000 | 60 | 0.000 | 0.748679 |

Signif. codes: $0^{***}, 0.001^{**}, 0.01^{*}, 0.05$ .

R-sq.(adj) = 0.926, Deviance explained = 92.6%

fREML = -14166, Scale est. = 0.01035, n = 16436

Table 16: **Donor and Lopez source significance.** Summary of the full GAM fit with a binary variable for the source of the data.

| Parametric Coefficients | | | | |
|---|---|---|---|---|
| **Term** | **Estimate** | **Std. Error** | **t value** | **Pr($>\|t\|$)** |
| (Intercept) | 3.074066 | 0.008256 | 372.32 | $< 2e-16^{***}$ |
| Sex1 | -0.061052 | 0.001802 | -33.88 | $< 2e-16^{***}$ |
| **Approximate Significance of Smooth Terms** | | | | |
| **Smooth Term** | **edf** | **Ref.df** | **F** | **p-value** |
| s(logAge) | 13.210 | 14 | 48.042 | $< 2e-16^{***}$ |
| s(logAge):Sex0 | 2.906 | 14 | 6.972 | $< 2e-16^{***}$ |
| s(logAge):Sex1 | 0.008 | 14 | 0.001 | $0.002343^{**}$ |
| s(logWeight) | 6.576 | 14 | 10.920 | $< 2e-16^{***}$ |
| s(logWeight):Sex0 | 1.667 | 14 | 0.781 | $0.000123^{***}$ |
| s(logWeight):Sex1 | 0.000 | 14 | 0.000 | 0.114077 |
| s(sqrtHeight) | 7.933 | 14 | 26.789 | $< 2e-16^{***}$ |
| s(sqrtHeight):Sex0 | 0.001 | 14 | 0.000 | 0.374278 |
| s(sqrtHeight):Sex1 | 0.001 | 14 | 0.000 | 0.319876 |
| ti(logWeight,sqrtHeight) | 6.537 | 16 | 3.056 | $< 2e-16^{***}$ |
| ti(logWeight,sqrtHeight):Sex0 | 0.001 | 16 | 0.000 | 0.213513 |
| ti(logWeight,sqrtHeight):Sex1 | 0.000 | 16 | 0.000 | 0.568359 |
| ti(sqrtHeight,logAge) | 5.637 | 16 | 1.188 | $7.22e-06^{***}$ |
| ti(sqrtHeight,logAge):Sex0 | 0.001 | 16 | 0.000 | 0.056236 . |
| ti(sqrtHeight,logAge):Sex1 | 2.091 | 16 | 0.223 | $0.009851^{**}$ |
| ti(logWeight,logAge) | 5.554 | 16 | 3.158 | $< 2e-16^{***}$ |
| ti(logWeight,logAge):Sex0 | 2.193 | 16 | 0.363 | $0.000941^{***}$ |
| ti(logWeight,logAge):Sex1 | 0.000 | 16 | 0.000 | 0.403036 |
| ti(logWeight,sqrtHeight,logAge) | 10.110 | 48 | 0.499 | $1.13e-05^{***}$ |
| ti(logWeight,sqrtHeight,logAge):Sex0 | 2.393 | 48 | 0.096 | $0.005800^{**}$ |
| ti(logWeight,sqrtHeight,logAge):Sex1 | 0.000 | 47 | 0.000 | 0.408821 |

Signif. codes: $0^{***}, 0.001^{**}, 0.01^{*}, 0.05$ .

R-sq.(adj) = 0.922, Deviance explained = 92.3%

fREML = -81960, Scale est. = 0.0058339, n = 71199

Table 17: Summary of the **full GAM with the trend correction** applied.

| Parametric Coefficients | | | | |
|---|---|---|---|---|
| **Term** | **Estimate** | **Std. Error** | **t value** | **Pr($>\|t\|$)** |
| (Intercept) | 3.042591 | 0.008141 | 373.74 | $< 2e-16^{***}$ |
| Sex1 | -0.063385 | 0.001514 | -41.85 | $< 2e-16^{***}$ |
| **Approximate Significance of Smooth Terms** | | | | |
| **Smooth Term** | **edf** | **Ref.df** | **F** | **p-value** |
| s(logAge) | 11.65 | 14 | 34.965 | $< 2e-16^{***}$ |
| s(logAge):Sex0 | 0.009 | 14 | 0.001 | $0.000831^{***}$ |
| s(logAge):Sex1 | 3.631 | 14 | 11.600 | $< 2e-16^{***}$ |
| s(logWeight) | 7.502 | 14 | 13.389 | $< 2e-16^{***}$ |
| s(logWeight):Sex0 | 1.681 | 14 | 1.294 | $5.92e-06^{***}$ |
| s(logWeight):Sex1 | 0.001 | 14 | 0.000 | $0.014477^{*}$ |
| s(sqrtHeight) | 7.468 | 14 | 30.839 | $< 2e-16^{***}$ |
| s(sqrtHeight):Sex0 | 0.001 | 14 | 0.000 | 0.332864 |
| s(sqrtHeight):Sex1 | 0.133 | 14 | 0.010 | 0.254426 |
| ti(logWeight,sqrtHeight) | 6.290 | 16 | 3.875 | $< 2e-16^{***}$ |
| ti(logWeight,sqrtHeight):Sex0 | 0.000 | 16 | 0.000 | 0.931679 |
| ti(logWeight,sqrtHeight):Sex1 | 0.000 | 16 | 0.000 | 0.844650 |
| ti(sqrtHeight,logAge) | 3.089 | 16 | 0.838 | $0.000116^{***}$ |
| ti(sqrtHeight,logAge):Sex0 | 0.000 | 16 | 0.000 | 0.772222 |
| ti(sqrtHeight,logAge):Sex1 | 0.001 | 16 | 0.000 | 0.667928 |
| ti(logWeight,logAge) | 6.547 | 16 | 10.611 | $< 2e-16^{***}$ |
| ti(logWeight,logAge):Sex0 | 0.001 | 16 | 0.000 | $0.052153\ .$ |
| ti(logWeight,logAge):Sex1 | 2.348 | 16 | 0.457 | $0.000372^{***}$ |
| ti(logWeight,sqrtHeight,logAge) | 11.39 | 48 | 0.729 | $1.38e-06^{***}$ |
| ti(logWeight,sqrtHeight,logAge):Sex0 | 0.001 | 48 | 0.000 | 0.871501 |
| ti(logWeight,sqrtHeight,logAge):Sex1 | 0.001 | 47 | 0.000 | 0.928091 |

Signif. codes: $0^{***}, 0.001^{**}, 0.01^{*}, 0.05\ .$

R-sq.(adj) = 0.927, Deviance explained = 92.7%

fREML = -83323, Scale est. = 0.0056168, n = 71199

Table 18: Summary of **the diagnostic model**, a full GAM with the trend and echo bias correction.

| Parametric Coefficients | | | | |
|---|---|---|---|---|
| **Term** | **Estimate** | **Std. Error** | **t value** | **Pr($>\|t\|$)** |
| (Intercept) | 2.50540 | 0.02036 | 123.085 | $< 2e-16^{***}$ |
| Sex1 | -0.30073 | 0.03412 | -8.815 | $< 2e-16^{***}$ |
| **Approximate Significance of Smooth Terms** | | | | |
| **Smooth Term** | **edf** | **Ref.df** | **F** | **p-value** |
| s(Age) | 8.618 | 14 | 21.414 | $< 2e-16^{***}$ |
| s(Age):Sex0 | 1.755 | 14 | 0.498 | $0.008599^{**}$ |
| s(Age):Sex1 | 0.009 | 13 | 0.001 | $0.027544^{*}$ |
| s(Weight) | 2.234 | 14 | 0.389 | $0.028441^{*}$ |
| s(Weight):Sex0 | 1.360 | 14 | 0.333 | $0.016009^{*}$ |
| s(Weight):Sex1 | 0.005 | 14 | 0.000 | 0.360066 |
| s(Height) | 0.001 | 14 | 0.000 | 0.821821 |
| s(Height):Sex0 | 5.630 | 14 | 1.299 | $0.000669^{***}$ |
| s(Height):Sex1 | 0.001 | 14 | 0.000 | 0.886755 |

Signif. codes: $0^{***}, 0.001^{**}, 0.01^{*}, 0.05$ .

R-sq.(adj) = 0.0324, Deviance explained = 3.26%

fREML = 1.9763e+05, Scale est. = 15.066, n = 71199

Table 19: Summary of the **conditional variance GAM** for the diagnostic model.

# B Appendix: Figures

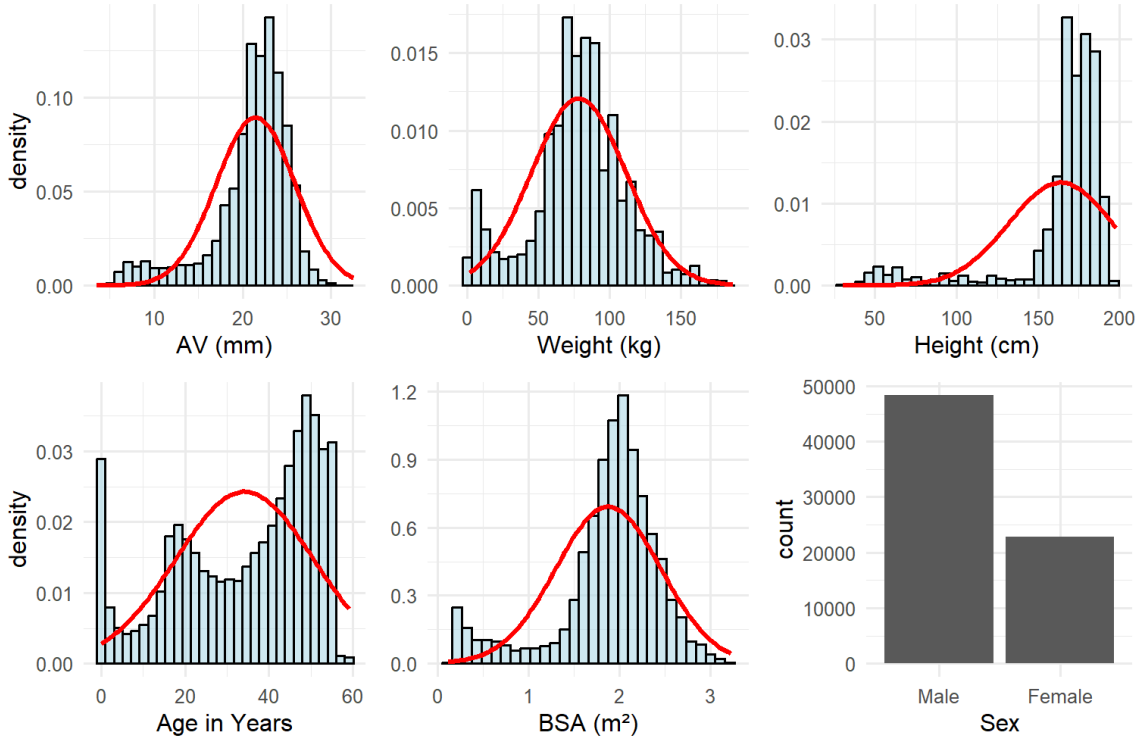



Figure 41: Histograms and bar chart for the raw donor data.

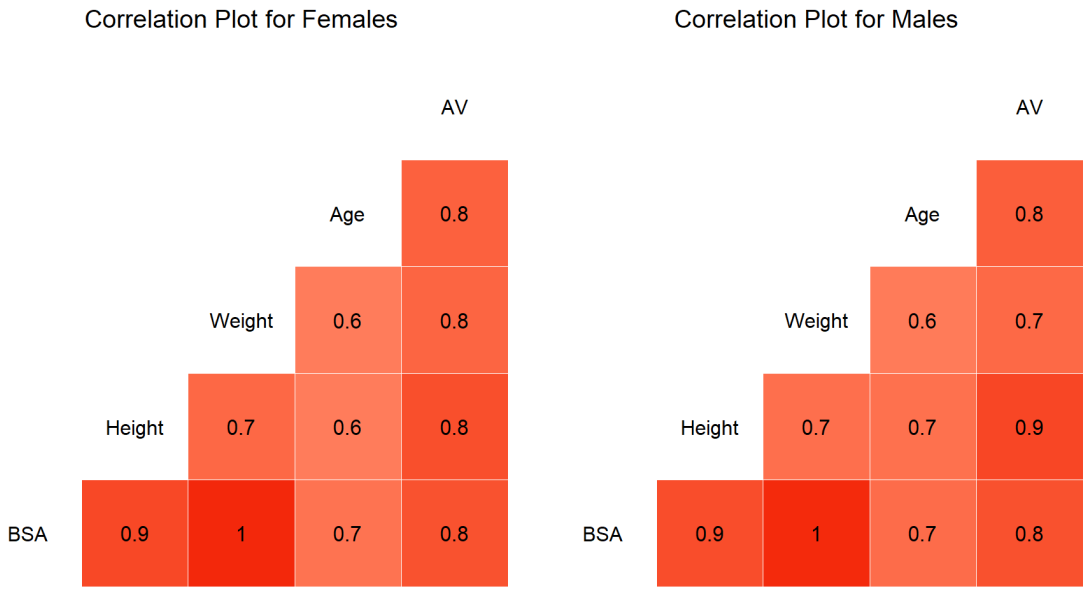



Figure 42: Correlation matrix for the continuous variables for the donor data.

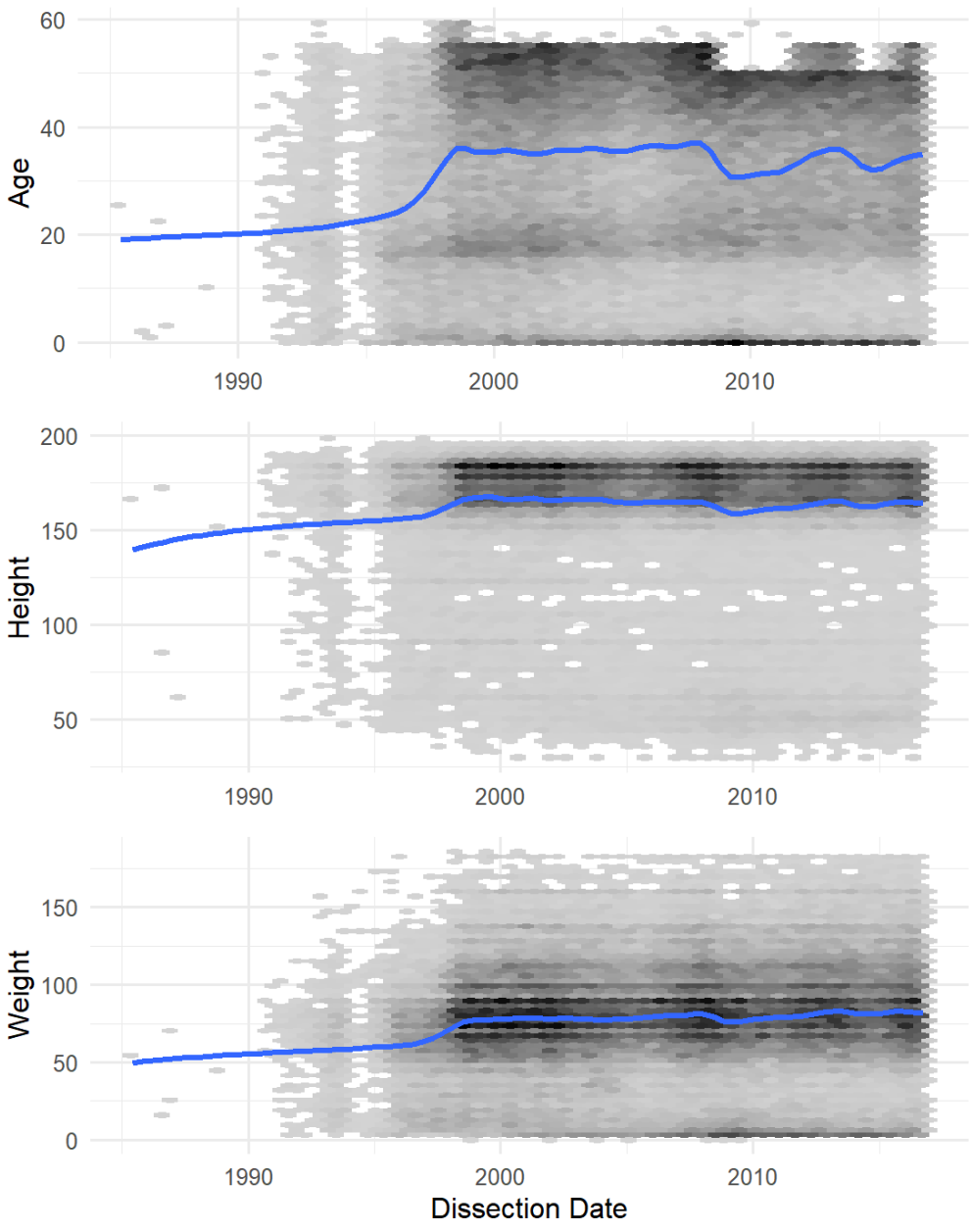



Figure 43: The predictors over time with LOESS fits for the donor data.

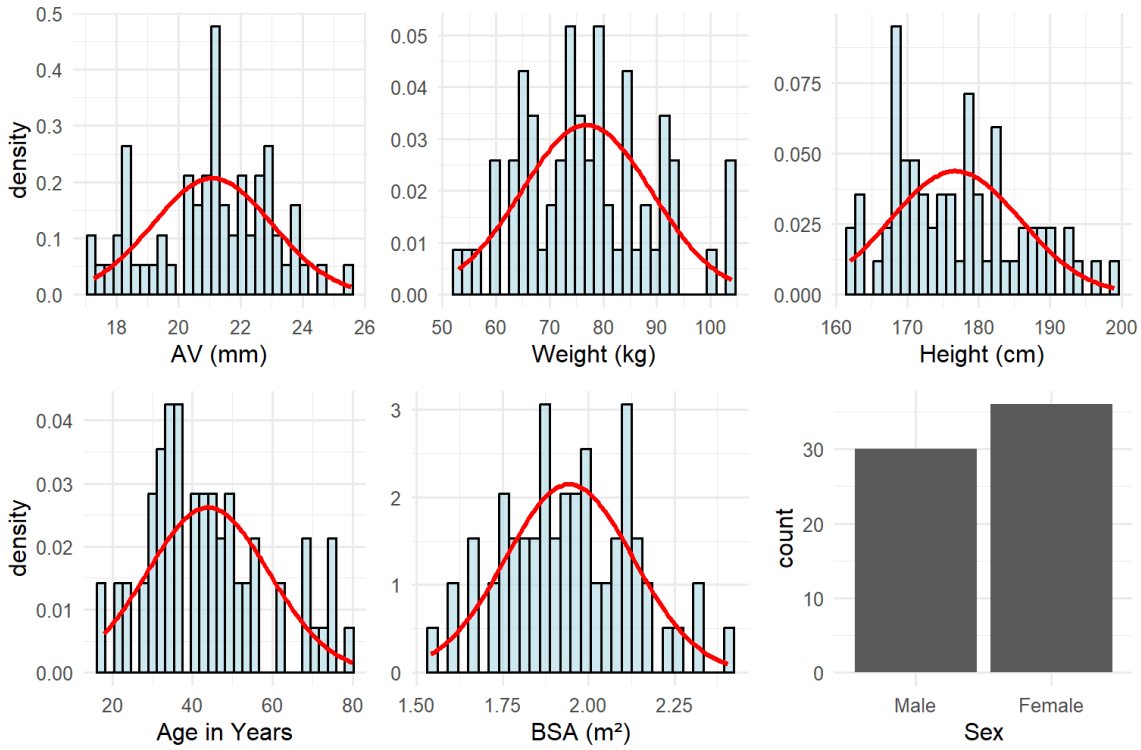



Figure 44: Histograms and bar chart for the raw echo data.

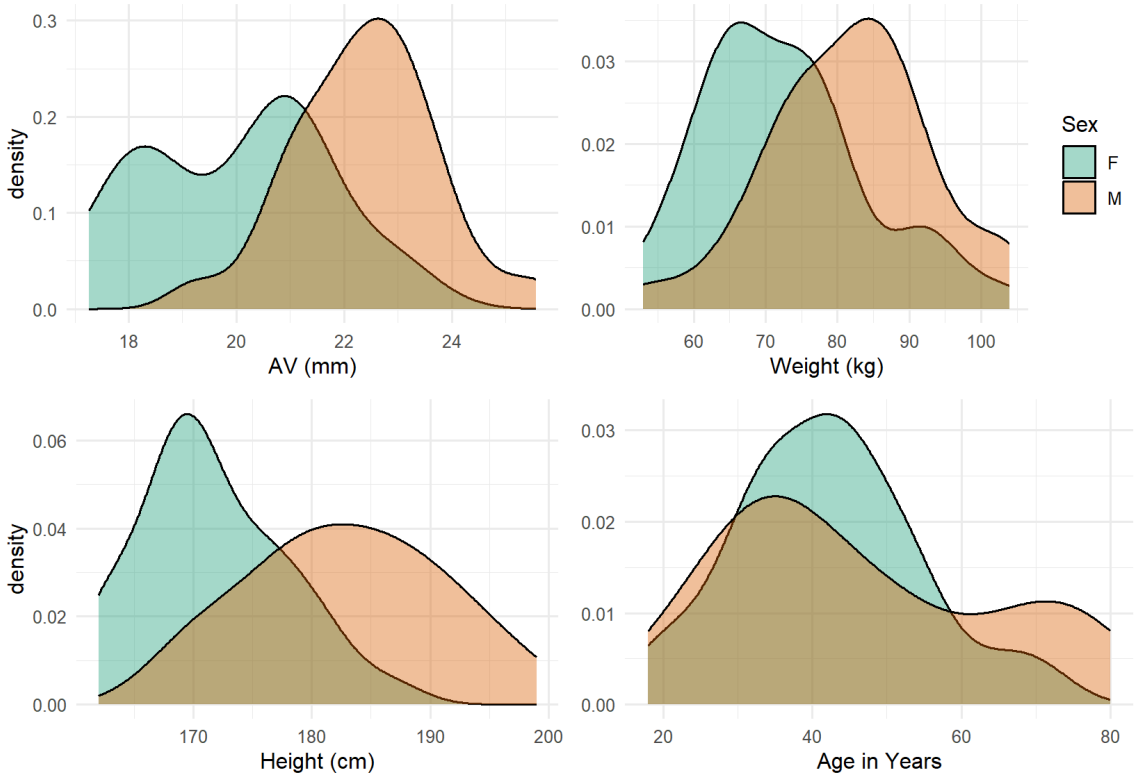



Figure 45: Density plots split by sex for the adult echo data. Female is **orange**, male is **teal**.

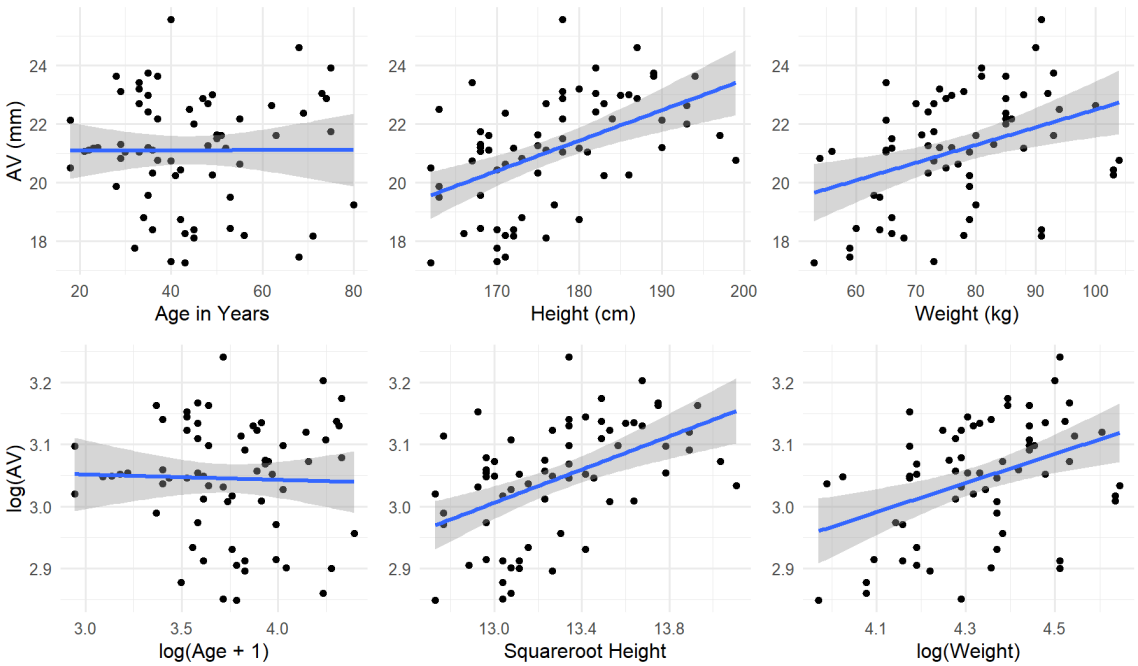



Figure 46: Linearity transformations for the adult UMCG echo data. The plots on the top are before the chosen transformations are applied and the plots on the bottom are after.

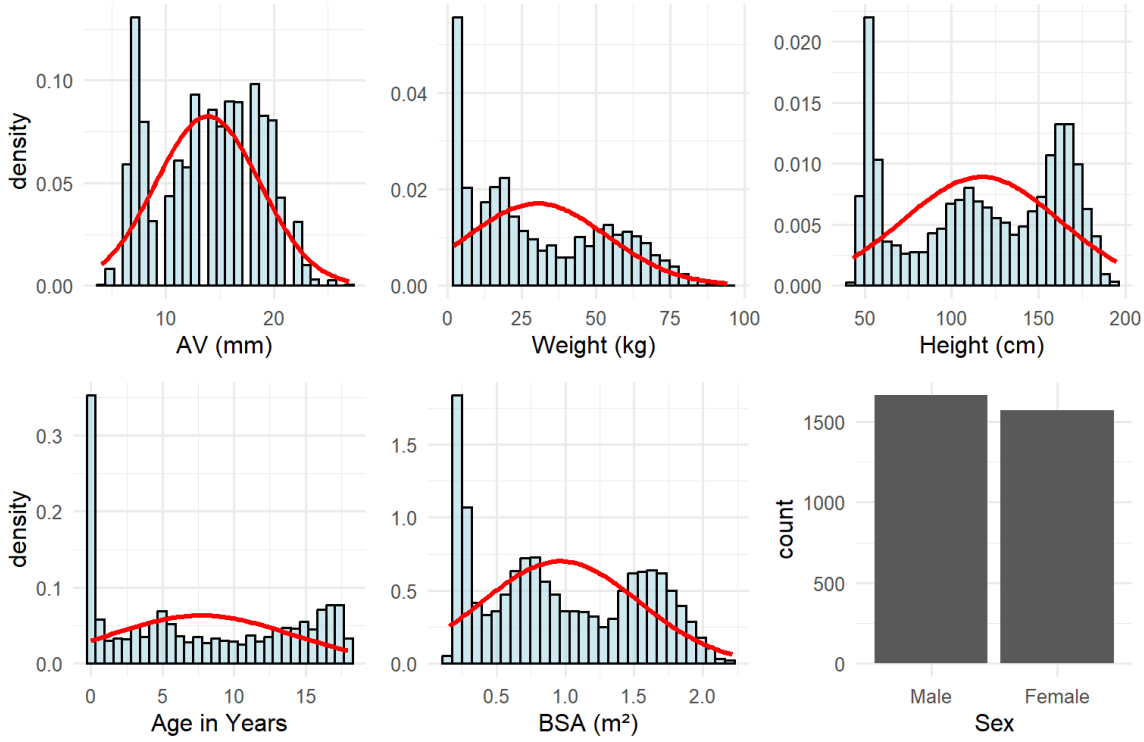



Figure 47: Histograms and bar chart for the Lopez *et al.* echo data.


Figure 48: Correlation matrix for the continuous variables for the Lopez *et al.* data.

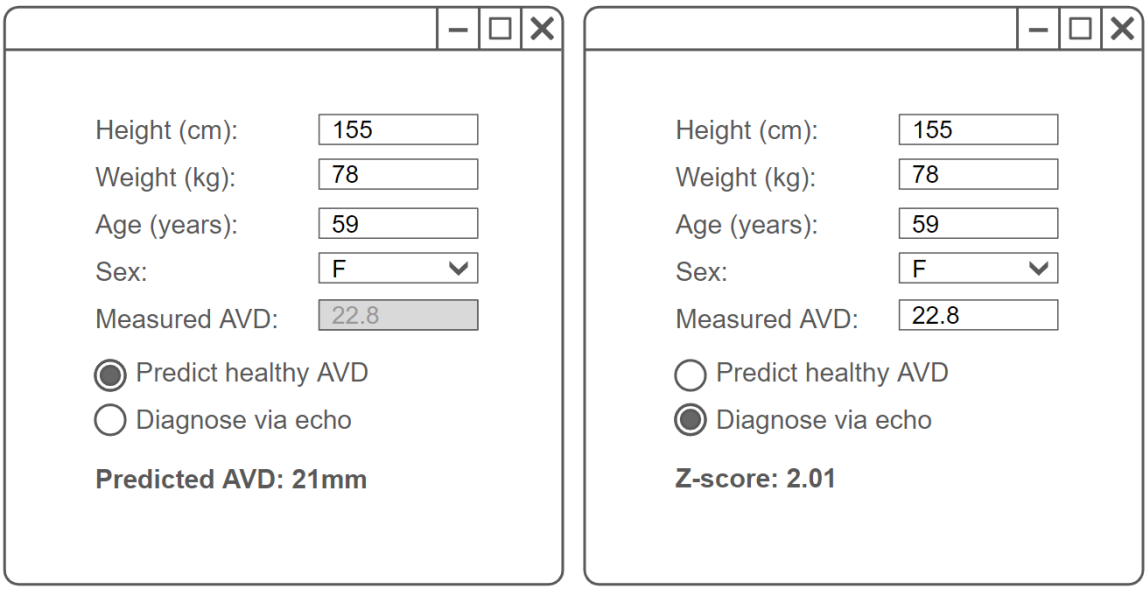



Figure 49: **An example of a the web tool would look.** When 'predict healthy AVD' is toggled, the model for AVR surgery would be used. When 'diagnosis via echo' is toggled, the echo-corrected AVD model and the model to estimate the standard deviation is used to compute a Z-score.

# C Appendix: Code

The code for the implementation of the segment neighbourhood algorithm and trend correction are included below. **For the complete code, please visit:** `https://github.com/LeahDijkshoorn/AVD-MSc-thesis`

### Segment Neighbourhood Function

```r
segmentNeighbourhood <- function(data, maxChangepoints, costFunction, penalty, minWindowSize = 1) {
  n <- length(data)

  # Initialise matrices to store costs and positions of last changepoints
  costMatrix <- matrix(Inf, n + 1, maxChangepoints + 1)
  lastChangepointMatrix <- matrix(-1, n + 1, maxChangepoints + 1)
  totalCost <- rep(Inf, maxChangepoints + 1)

  # Initialise the cost of zero segments, no data
  costMatrix[1, 1] <- 0

  # Dynamic programming to fill the matrices
  for (q in 1:(maxChangepoints + 1)) {
    for (i in 2:(n + 1)) {
      if (q == 1) {
        # Only calculate if segment is large enough to be valid
        if (i - 1 >= minWindowSize) {
          costMatrix[i, q] <- costFunction(data[1:(i - 1)], penalty, q - 1)
        }
      } else {
        # Ensure j starts at least minWindowSize before i
        for (j in 1:(i - 1)) {
          if (i - j >= minWindowSize) {
            currentCost <- costFunction(data[j:(i - 1)], penalty, q - 2) + costMatrix[j, q - 1]
            if (!is.na(currentCost) && currentCost < costMatrix[i, q]) {
              costMatrix[i, q] <- currentCost
              lastChangepointMatrix[i, q] <- j
            }
          }
        }
      }
    }
    totalCost[q] <- costMatrix[n+1, q] + penalty * (q - 1)
  }

  # Determine optimal number of changepoints based on total cost
  optimalChangepoints <- which.min(totalCost) - 1
  segments <- numeric(optimalChangepoints)

  # Reconstruct optimal segmentation
  currentSegment <- optimalChangepoints
  currentPosition <- n + 1
  while (currentSegment > 0) {
    segments[currentSegment] <- lastChangepointMatrix[currentPosition, currentSegment + 1] - 1 # First run starts at end of data
    currentPosition <- lastChangepointMatrix[currentPosition, currentSegment + 1] # Update current position to end of previous segment
    currentSegment <- currentSegment - 1
  }

  return(list(changepoints = segments, cost = totalCost[optimalChangepoints + 1]))
}
```

### Trend Correction Function

```r
correct_trend_and_update <- function(data, start_date, end_date) {
  start_date <- as.Date(start_date)
  end_date <- as.Date(end_date)

  # Subset the data based on the provided changepoints
  date_subset_data <- data[data$dissection_date >= start_date & data$dissection_date
     <= end_date, ]

  # Fit linear model and correct for trend
  lm_trend_correction <- lm(Residuals ~ as.Date(dissection_date), data = date_subset
    _data)
  trend_predictions <- predict(lm_trend_correction, newdata = date_subset_data)
  #trend_difference_FromMean <- trend_predictions - mean(data$AV)

  date_subset_data$AV_trend_correction <- date_subset_data$AV - trend_predictions

  # Merge corrected AV values back into the main dataset
  data <- merge(data, date_subset_data[c("DonorID", "AV_trend_correction")], by = "
    DonorID", all.x = TRUE)

  # Update AV_corrected in the original dataset
  data$AV_corrected <- ifelse(!is.na(data$AV_trend_correction), data$AV_trend_
    correction, data$AV)

  # Replace original column
  data$AV <- data$AV_corrected

  # Clean up by removing temporary columns
  data$AV_trend_correction <- NULL
  data$AV_corrected <- NULL

  return(data)
}
```

# References


[1] W. S. Aronow, C. Ahn, and I. Kronzon, "Comparison of echocardiographic abnormalities in african-american, hispanic, and white men and women aged¿ 60 years," *American Journal of Cardiology*, vol. 87, no. 9, pp. 1131–1133, 2001.

[2] D. S. Bach, J. I. Radeva, H. G. Birnbaum, A.-A. Fournier, and E. G. Tuttle, "Prevalence, referral patterns, testing, and surgery in aortic valve disease: leaving women and elderly patients behind?," *The Journal of heart valve disease*, vol. 16, no. 4, pp. 362–369, 2007.

[3] A. M. Iivanainen, M. Lindroos, R. Tilvis, J. Heikkilä, and M. Kupari, "Natural history of aortic valve stenosis of varying severity in the elderly," *The American journal of cardiology*, vol. 78, no. 1, pp. 97–101, 1996.

[4] C. M. Otto, "Timing of aortic valve surgery," *Heart*, vol. 84, no. 2, pp. 211–218, 2000.

[5] S. Coffey, R. Roberts-Thomson, A. Brown, J. Carapetis, M. Chen, M. Enriquez-Sarano, L. Zühlke, and B. D. Prendergast, "Global epidemiology of valvular heart disease," *Nature Reviews Cardiology*, vol. 18, no. 12, pp. 853–864, 2021.

[6] L. J. Davidson and C. J. Davidson, "Transcatheter treatment of valvular heart disease: a review," *Jama*, vol. 325, no. 24, pp. 2480–2494, 2021.

[7] A. P. Durko, P. Pibarot, P. Atluri, V. Bapat, D. E. Cameron, F. P. Casselman, E. P. Chen, G. Dahle, J. A. Elefteriades, P. Lancellotti, *et al.*, "Essential information on surgical heart valve characteristics for optimal valve prosthesis selection: expert consensus document from the european association for cardio-thoracic surgery (eacts)–the society of thoracic surgeons (sts)–american association for thoracic surgery (aats) valve labelling task force," *European Journal of Cardio-Thoracic Surgery*, vol. 59, no. 1, pp. 54–64, 2021.

[8] A. M. Kasel, S. Cassese, S. Bleiziffer, M. Amaki, R. T. Hahn, A. Kastrati, and P. P. Sengupta, "Standardized imaging for aortic annular sizing: implications for transcatheter valve selection," *JACC: Cardiovascular Imaging*, vol. 6, no. 2, pp. 249–262, 2013.

[9] M. J. Mack, M. B. Leon, C. R. Smith, D. C. Miller, J. W. Moses, E. M. Tuzcu, J. G. Webb, P. S. Douglas, W. N. Anderson, E. H. Blackstone, *et al.*, "5-year outcomes of transcatheter aortic valve replacement or surgical aortic valve replacement for high surgical risk patients with aortic stenosis (partner 1): a randomised controlled trial," *The Lancet*, vol. 385, no. 9986, pp. 2477–2484, 2015.

[10] M. Cantinotti, N. Assanta, M. Crocetti, M. Marotta, B. Murzi, and G. Iervasi, "Limitations of current nomograms in pediatric echocardiography: Just the tip of the iceberg—a call for standardization," *Journal of the American Society of Echocardiography*, vol. 27, no. 3, p. 339, 2014.

[11] F. Dallaire, J.-L. Bigras, M. Prsa, and N. Dahdah, "Bias related to body mass index in pediatric echocardiographic z scores," *Pediatric cardiology*, vol. 36, pp. 667–676, 2015.

[12] J. Mahgerefteh, W. Lai, S. Colan, F. Trachtenberg, R. Gongwer, M. Stylianou, A. H. Bhat, D. Goldberg, B. McCrindle, P. Frommelt, *et al.*, "Height versus body surface area to normalize cardiovascular measurements in children using the pediatric heart network echocardiographic z-score database," *Pediatric cardiology*, vol. 42, pp. 1284–1292, 2021.

[13] L. Lopez, S. Colan, M. Stylianou, S. Granger, F. Trachtenberg, P. Frommelt, G. Pearson, J. Camarda, J. Cnota, M. Cohen, *et al.*, "Relationship of echocardiographic z scores adjusted for body surface area to age, sex, race, and ethnicity: the pediatric heart network normal echocardiogram database," *Circulation: Cardiovascular Imaging*, vol. 10, no. 11, p. e006979, 2017.

[14] M. Cantinotti, R. Giordano, M. Scalese, B. Murzi, N. Assanta, I. Spadoni, C. Maura, M. Marco, S. Molinaro, S. Kutty, *et al.*, "Nomograms for two-dimensional echocardiography derived valvular and arterial dimensions in caucasian children," *Journal of Cardiology*, vol. 69, no. 1, pp. 208–215, 2017.

[15] M. D. Pettersen, W. Du, M. E. Skeens, and R. A. Humes, "Regression equations for calculation of z scores of cardiac structures in a large cohort of healthy infants, children, and adolescents: an echocardiographic study," *Journal of the American Society of Echocardiography*, vol. 21, no. 8, pp. 922–934, 2008.

[16] Centers for Disease Control and Prevention, "National health and nutrition examination survey: Demographics data." `https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics&CycleBeginYear=2015`, Sep 2017. Accessed: 2024-07-16.

[17] W. H. Organization *et al.*, *WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development*. World Health Organization, 2006.

[18] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

[19] S. N. Wood, *Generalized additive models: an introduction with R*. chapman and hall/CRC, 2017.

[20] G. Haycock, G. Schwartz, and D. Wisotsky, "Geometric method for measuring body surface area: A height weight formula validated in infants, children and adults," *Journal of Pediatrics*, vol. 93, no. 1, pp. 62–66, 1978.

[21] E. Boyd, "The growth of the surface area of the human body," *University of Minnesota. Institute of Child Welfare. Monograph Series Number 10*, vol. 19, pp. 1–15, 1935.

[22] E. F. DuBois and D. DuBois, "The measurement of the surface area of man," *Archives of Internal Medicine*, vol. 17, no. 6, p. 863, 1916.

[23] M. Savva, *Pharmaceutical Calculations*. Springer, 2019.

[24] R. D. Mosteller, "Simplified calculation of body-surface area," *New England Journal of Medicine*, vol. 317, no. 17, p. 1098, 1987.

[25] T. Bement and M. Waterman, "Locating maximum variance segments in sequential data," *Journal of the International Association for Mathematical Geology*, vol. 9, pp. 55–61, 1977.

[26] I. E. Auger and C. E. Lawrence, "Algorithms for the optimal identification of segment neighborhoods," *Bulletin of mathematical biology*, vol. 51, no. 1, pp. 39–54, 1989.