The development and implementation of automated methods for the reproducible assessment of aortic stenosis and mitral regurgitation

Matthew Shun-Shin
BM BCh MA (Oxon) MRCP(UK)

A thesis submitted to Imperial College London
for the degree of
Doctor of Philosophy

International Centre for Circulatory Health
National Heart & Lung Institute
Imperial College London
London W12 0HS

Submitted 2nd August 2018
Final 1st March 2019

Abstract

Valve disease is the commonest reason for a patient to undergo repeated echocardiography, which is itself by far the commonest type of cardiac imaging. Of the various valve diseases, aortic stenosis and mitral regurgitation are the two that most commonly require intervention, typically with surgery.

Echocardiographic evaluation, alongside symptoms, is the most powerful determinant of the timing of surgery for aortic stenosis and mitral regurgitation. The purpose of sequential follow-up echocardiography is to make a major contribution by accurately defining the severity of the valve lesion at any point in time.

However, under blinded conditions conventional echocardiographic evaluation of these conditions has surprisingly poor test-retest reproducibility. In clinical practice this leads to indecision and wasteful clinical activity. In clinical trials this leads to the need for larger (i.e. more expensive) trials to detect the same effect size.

Improvements in the precision of quantification would allow clinical services to improve the certainty of diagnostic staging and reduce the frequency of follow-up. It would allow clinical trials to successfully test treatments with fewer patients.

In this thesis I demonstrate in Chapter 2 how the reproducibility of existing techniques has been misleadingly reported in the literature, compared with what is achieved even in the most stringent real-world scenarios.

I then develop methods for segregating and then displaying to clinicians the sources of variability in echocardiographic measurements (Chapter 3). I apply this to the valve condition to where this is most important, namely aortic stenosis.

I then develop in Chapter 4 a theoretical algorithm which should improve the test-retest reproducibility of echocardiographic assessment of aortic stenosis severity. I create an open-source software suite to perform the analysis automatically. To make it maximally applicable I designed this to operate on image streams from any model and any manufacturer of echocardiograph.

Having developed the software suite, I enrolled 93 patients in clinical study of validity and test-retest reproducibility with an approximately 12-month interval between scans. To minimise the possibility of me biasing the results favourably, I engaged a colleague in another university to check the validity in a cohort of 10 patients in her hospital.

Having completed my initial ambition to improve reproducibility of aortic stenosis evaluation, I explored the application of what I had learnt to mitral regurgitation (Chapter 5). This is a very different problem of physics to solve, and it seems to have a different solution. I have proposed a method of using colour M-mode traces to improve the reproducibility of measurements of effective regurgitant orifice area. I then perform a pilot validation study of the improved method at my own institution.

In my thesis overall, I detect a systematic problem in the way reproducibility is described, and then focus on one valve condition (aortic stenosis). I bring together insights from cardiology, statistics, and computer science to design and build a method of improving precision evaluation of stenosis. The subsequent testing suggests that this method does indeed reduce variance about 4-fold.

The method I present for decomposing and displaying the variably may have general applicability throughout clinical measurement science. The approach I have taken to these valve diseases could be taken with other conditions in the future.

Declaration of originality

I hereby declare that the work presented in this thesis is my own and all else is referenced appropriately.

Matthew James Shun-Shin

Copyright declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Dedication

To my wife, my parents, and my sister, who have always supported and encouraged me; and to my daughter Ada, whose recent arrival has filled me with joy.

Acknowledgements

I have worked with some fantastic colleagues and friends throughout this PhD and my time at Imperial.

I thank Prof. Darrel Francis who has supported me throughout my cardiology career and provided the initial idea for this thesis. With Prof. Jamil Mayet he has created a department that supports and inspires fellows, and I am thankful for being able to be a part of it.

Thank you to Dr. Nilesh Sutaria and Dr. Graham Cole for their help with this thesis and for teaching me cardiac imaging.

Thank you to Nina Bual, for both teaching me echocardiography and for her patience as I disrupt her echocardiography department for yet another study.

Thank you to my friends and colleagues. Thank you to Rasha Al-Lamee, Judy Finegold, James Howard, Ahran Arnold, Dan Keene, Chris Cook, and Yousif Ahmad. We have collaborated on some fantastic studies and trials together and it has been great to be part of this team of fellows. A second thank you to Rasha and Judy for their support and ensuring I made it to the end.

Thank you to the British Heart Foundation who supported this research.

A final thanks to all the patients who took part for their time and effort whilst I was taking "just one more set of images".

Acronyms

AQURO: Automated QUantification of Regurgitant Orifice area

AS: Aortic Stenosis

CV: Coefficient of Variation

CW: Continuous Wave

DI: Dimensionless Index. In this thesis it is calculated as the ratio of the aortic velocity to the LVOT velocity so that values are >1.

EF: Ejection Fraction

EROA: Effective Regurgitant Orifice Area

iDI: instantaneous Dimensionless Index

ISO: International Standards Organisation

LVOT: Left Ventricular Outflow Tract

MR: Mitral Regurgitation

PW: Pulsed Wave

RCT: Randomised Clinical Trial

RR: Relative Risk

SE: Standard Error

SD: Standard Deviation

SDD: Standard Deviation of the Difference

Table of contents

1 INTRODUCTION 17

1.1 INTRODUCTION 18
1.2 PREVALENCE OF VALVULAR DISEASE 24
1.3 THE ROLE OF IMAGING TO ASSESS AORTIC STENOSIS 24
1.4 ECHOCARDIOGRAPHY AND AORTIC VALVE DISEASE 25
1.5 ECHOCARDIOGRAPHY AND MITRAL VALVE DISEASE 28
1.6 QUANTITATIVE ECHO IN PRACTICE 31
1.7 THE REPRODUCIBILITY CRISIS 32
1.8 THE IMPORTANCE OF MEASUREMENT PRECISION 32
1.9 IMPORTANCE OF OPEN BIOMARKERS 34
1.10 GOING FORWARD 35

2 CAN WE RELY ON REPRODUCIBILITY STUDIES? A SYSTEMATIC REVIEW AND META-ANALYSIS OF 6727 PATIENTS FROM 97 STUDIES. 37

2.1 ABSTRACT 38
2.1.1 Background 38
2.1.2 Method 38
2.1.3 Results 39
2.1.4 Conclusions 39
2.2 INTRODUCTION 41
2.2.1 Why reproducibility is important 41
2.2.2 How is reproducibility measured 41
2.2.3 Potential sources of reproducibility data 42
2.2.4 Aims of this study 43
2.3 METHODS 45
2.3.1 Eligibility and search strategy 45
2.3.2 Data abstraction 47
2.3.3 Calculation of standard deviation of difference from reproducibility studies 47
2.3.4 Calculation of the standard deviation of difference from randomised controlled trials 49
2.3.5 Reproducibility test conditions 51
2.3.6 Data analysis 52
2.4 RESULTS 53
2.4.1 Results of the searches 53
2.4.2 Intra-reader reproducibility of echocardiographic measurements 54
2.4.3 Inter-reader reproducibility of echocardiographic measurements 57
2.4.4 Test-retest variability of echocardiographic measurements 60
2.4.5 Test-retest variability from randomised controlled trials 63
2.4.6 Comparison between intra-reader, inter-reader, and test-retest variability derived from reproducibility studies and randomised controlled trials 66
2.4.7 Comparison of RCT data to reproducibility study data for specific parameters 70
2.5 DISCUSSION 75
2.5.1 Biomarker features associated with large coefficient of variation 75
2.5.2 Limitations 77
2.5.3 Clinical implications 80
2.5.4 Implications for research 81
2.6 CONCLUSION 81

3 THE CAUSES OF IRREPRODUCIBILITY IN DOPPLER MEASUREMENTS OF AORTIC STENOSIS 83

3.1 ABSTRACT 84
3.1.1 Introduction 84
3.1.2 Methods 85
3.1.3 Results 86
3.1.4 Conclusions 87
3.2 INTRODUCTION 88
3.2.1 What are the causes of irreproducibility in the Doppler assessment of aortic stenosis? 88
3.2.2 Aims of this study 89

3.3 METHODS 92
3.3.1 Distinguishing the sources of variability even though they can only be measured nested 92
3.3.2 Quantifying the sources of variability 93
3.3.3 Patients 94
3.3.4 Data collection 96
3.3.5 Measurement of the echocardiographic data 98
3.3.6 Data analysis 100
3.4 RESULTS 105
3.4.1 Baseline characteristics 105
3.4.2 Display of variability components in clinically appreciable manner 105
3.4.3 Application of display method to the peak aortic velocity data 113
3.4.4 Application to LVOT peak velocity 121
3.4.5 Components of variability 125
3.5 DISCUSSION 128
3.5.1 Surprising dominance of probe placement component 128
3.5.2 Opportunities for reducing variability originating at probe placement level 129
3.5.3 These data suggest that current advice given to clinicians and researchers is wrong and harmful 132
3.5.4 Targeting variability 134
3.5.5 Effect of variability on management 135
3.5.6 Variability limits the maximum agreement between two measurement techniques 138
3.5.7 Limitations 138
3.5.8 Comparison with other literature 139
3.6 CONCLUSION 141

4 THE DEVELOPMENT AND VALIDATION OF THE "INSTANTANEOUS DIMENSIONLESS INDEX": A NOVEL TOOL TO MORE RELIABLY FOLLOW-UP AORTIC STENOSIS 142

4.1 ABSTRACT 143
4.1.1 Introduction 143
4.1.2 Methods 143
4.1.3 Results 144
4.1.4 Conclusions 145
4.2 INTRODUCTION 146
4.2.1 Non-simultaneity of a large element of error. 148
4.2.2 Overwhelming advantages of simultaneous measurement. 149
4.2.3 Steps to the solution 151
4.2.4 Validation of the instantaneous Dimensionless Index. 151
4.3 METHODS 153
4.3.1 Development of the instantaneous Dimensionless Index 153
4.3.2 Study population 181
4.3.3 Doppler echocardiography 181
4.3.4 Data acquisition 181
4.3.5 Automated image analysis 182
4.3.6 Calculation of iDI 184
4.3.7 Data analysis 184
4.3.8 Statistical analysis 189
4.4 RESULTS 190
4.4.1 Patient Group A (Imperial College NHS Healthcare Trust, London) 190
4.4.2 Validation of the instantaneous Dimensionless Index in Patient Group B 200
4.5 DISCUSSION 205
4.5.1 Main findings 205
4.5.2 Paradoxical increase in precision 206
4.5.3 Peak velocity or velocity time integral 206
4.5.4 Limitations 208
4.5.5 Implications for clinical practice 212
4.5.6 Implications for research 212
4.5.7 Future directions 213
4.6 CONCLUSION 217

5 THE DEVELOPMENT AND PILOT VALIDATION OF "AQURO", AN AUTOMATED METHOD FOR THE QUANTIFICATION OF MITRAL REGURGITATION SEVERITY 218

5.1 ABSTRACT 219

5.1.1 Introduction *219*
5.1.2 Methods *219*
5.1.3 Results *220*
5.1.4 Conclusions *220*
5.2 INTRODUCTION 222
5.2.1 Difficulties encountered by clinicians in calculating the EROA using the PISA method *224*
5.2.2 What is the true PISA shape? *225*
5.2.3 The AQURO method – resolving these difficulties *232*
5.2.4 The AQURO principle *232*
5.2.5 AQURO: Overcoming the difficulties of PISA *244*
5.3 METHODS 246
5.3.1 The AQURO algorithm *246*
5.3.2 Study Population *263*
5.3.3 Data acquisition ..... ***Error! Bookmark not defined.***
5.3.4 Image acquisition *264*
5.3.5 Image analysis *264*
5.3.6 Data analysis *266*
5.3.7 Feasibility and data exclusion ..... ***Error! Bookmark not defined.***
5.3.8 Statistical analysis ..... ***Error! Bookmark not defined.***
5.4 RESULTS 268
5.4.1 Patient population *268*
5.4.2 Relationship between the AQURO and PISA derived EROA and visual assessment *268*
5.4.3 Reproducibility of AQURO and PISA derived EROA *272*
5.4.4 Relationship and agreement between AQURO derived EROA and PISA derived EROA *272*
5.5 DISCUSSION 275
5.5.1 Main findings *275*
5.5.2 Implications of imprecision in EROA assessment *275*
5.5.3 Calibration *276*
5.5.4 Limitations *277*
5.5.5 Implications for clinical practice *279*
5.5.6 Implications for research *279*
5.6 CONCLUSION 280

6 SYNTHESIS 281

6.1 RECOGNISING THAT THE LITERATURE ON REPRODUCIBILITY IS ALMOST WORTHLESS 281
6.2 TARGETING THE TRUE SOURCES OF VARIABILITY 282
6.3 DESIGNING AND VALIDATING A NEW BIOMARKER – THE INSTANTANEOUS DIMENSIONLESS INDEX – WITH THE AIM TO IMPROVE REPRODUCIBILITY 282
6.4 IMPROVING AND VALIDATING AQURO DERIVED EROA FOR MITRAL REGURGITATION 283
6.5 FUTURE PLANS 283

7 REFERENCES 284

List of tables

TABLE 1.1 MEASUREMENTS REQUIRED FOR ECHOCARDIOGRAPHIC BIOMARKERS 33
TABLE 2.1 TEST-RETEST COEFFICIENTS OF VARIATION FOR ECHOCARDIOGRAPHIC BIOMARKERS FOR WHICH DATA FROM BOTH THE PLACEBO-ARM OF RCTS AND DEDICATED REPRODUCIBILITY STUDIES WERE AVAILABLE. 67
TABLE 2.2 REPRODUCIBILITY OF AORTIC STENOSIS PARAMETERS 70
TABLE 2.3 REPRODUCIBILITY OF LVEF MEASUREMENTS 73
TABLE 3.1 FOR THE SUM OF TWO VARIABLES IT IS THE VARIANCES RATHER THAN STANDARD DEVIATIONS THAT SUMMATE 107
TABLE 3.2 CONTINUOUS WAVE: SOURCES OF VARIABILITY AND THEIR COMPONENT AND OBSERVED COEFFICIENT OF VARIATION (CV) 118
TABLE 3.3 PULSED WAVE: SOURCES OF VARIABILITY AND THEIR COMPONENT AND OBSERVED COEFFICIENT OF VARIATION (CV) 122
TABLE 6.1 ARCHITECTURE OF COLOUR M-MODE DOPPLER NEURAL NET 257

List of figures

FIGURE 1-1 NOISE ON THE X-AXIS BUT NOT NOISE ON THE Y-AXIS BIASES THE SLOPE OF THE REGRESSION 23
FIGURE 1-2 NUMBER OF NEW PUBLICATIONS EACH YEAR DESCRIBING A "NOVEL BIOMARKER" 35
FIGURE 2-1 INDIVIDUAL INTRA-READER COEFFICIENTS OF VARIATION FROM REPRODUCIBILITY STUDIES 55
FIGURE 2-2 SUMMARISED INTRA-READER COEFFICIENTS OF VARIATION FROM REPRODUCIBILITY STUDIES 56
FIGURE 2-3 INDIVIDUAL INTER-READER COEFFICIENTS OF VARIATION FROM REPRODUCIBILITY STUDIES 58
FIGURE 2-4 WEIGHTED MEAN INTER-READER COEFFICIENTS OF VARIATION 59
FIGURE 2-5 INDIVIDUAL TEST-RETEST COEFFICIENTS OF VARIATION FROM REPRODUCIBILITY STUDIES 61
FIGURE 2-6 WEIGHTED MEAN TEST-RETEST COEFFICIENTS OF VARIATION FROM REPRODUCIBILITY STUDIES 62
FIGURE 2-7 INDIVIDUAL TEST-RETEST COEFFICIENTS OF VARIATION FROM THE PLACEBO ARMS OF RCTS 64
FIGURE 2-8 WEIGHTED MEAN TEST-RETEST COEFFICIENTS OF VARIATION FROM THE PLACEBO ARMS OF RCTS ... 65
FIGURE 2-9 INDIVIDUAL TEST-RETEST CV OF MEASUREMENTS FOR WHICH DATA WAS AVAILABLE FOR BOTH REPRODUCIBILITY STUDIES AND THE PLACEBO ARMS OF RCTS 68
FIGURE 2-10 SUMMARY TEST-RETEST CV OF MEASUREMENTS WHICH DATA WAS AVAILABLE FOR BOTH REPRODUCIBILITY STUDIES AND THE PLACEBO ARMS OF RCTS 69
FIGURE 2-11 INTRA-READER, INTER-READER, AND TEST-RETEST COEFFICIENTS OF VARIATION FROM REPRODUCIBILITY STUDIES AND RCTS 71
FIGURE 2-12 INDIVIDUAL CV OF AORTIC STENOSIS MEASUREMENTS FROM REPRODUCIBILITY STUDIES AND THE PLACEBO ARMS OF RCTS 72
FIGURE 2-13 SUMMARY CV OF AORTIC STENOSIS MEASUREMENTS FROM REPRODUCIBILITY STUDIES AND THE PLACEBO ARMS OF RCTS 72
FIGURE 2-14 INDIVIDUAL COEFFICIENTS OF VARIATION FOR CONTRACT (C-) AND NON-CONTRAST (NC-) LVEF FROM REPRODUCIBILITY STUDIES AND THE PLACEBO ARMS OF RCTS 74
FIGURE 2-15 SUMMARY COEFFICIENTS OF VARIATION OF CONTRAST (C-) AND NON-CONTRAST (NC-) LVEF FORM REPRODUCIBILITY STUDIES AND THE PLACEBO ARMS OF RCTS 74
FIGURE 3-1 TWO COHORTS OF PATIENTS WERE USED TO PROVIDE DATA FOR EVERY SOURCE OF VARIABILITY. COHORT A CONTRIBUTED DATA AT THE BEAT, PROBE PLACEMENT, AND VISIT LEVEL OF VARIABILITY. COHORT B CONTRIBUTED DATA AT THE TRACING, BEAT, PLACEMENT, OPERATOR, AND READER LEVEL. 95
FIGURE 3-2 RELATIVE SIZES OF TWO COMPONENTS CONTRIBUTING TO REPRODUCIBILITY 109
FIGURE 3-3 THE ADDITION OF TWO COMPONENTS OF VARIATION 110
FIGURE 3-4 SUBTRACTION OF ONE SOURCE OF VARIABILITY. 112
FIGURE 3-5 THE "TRACING" AND "BEAT" LEVEL COMPONENTS OF VARIATION 114
FIGURE 3-6 THE **COMPONENT VIEW** OF THE **PEEL-AWAY-**VARIANCE **MODEL** OF THE SOURCES OF VARIABILITY CONTRIBUTING TO THE ECHOCARDIOGRAPH MEASUREMENT OF THE CONTINUOUS WAVE LVOT PEAK VELOCITY 119
FIGURE 3-7 THE **PEEL-AWAY-VARIANCE REPRESENTATION** OF THE **PEEL-AWAY-VARIANCE** MODEL OF THE SOURCES OF VARIABILITY CONTRIBUTING TO THE ECHOCARDIOGRAPH MEASUREMENT OF THE CONTINUOUS WAVE PEAK TRANS-AORTIC VELOCITY 120
FIGURE 3-8 THE **COMPONENT VIEW** OF THE **PEEL-AWAY-VARIANCE MODEL** OF THE SOURCES OF VARIABILITY CONTRIBUTING TO THE ECHOCARDIOGRAPH MEASUREMENT OF THE PULSED WAVE LVOT PEAK VELOCITY 123
FIGURE 3-9 THE **PEEL-AWAY-VARIANCE REPRESENTATION** OF THE **PEEL-AWAY-VARIANCE MODEL** OF THE SOURCES OF VARIABILITY CONTRIBUTING TO THE ECHOCARDIOGRAPH MEASUREMENT OF THE PULSED WAVE LVOT PEAK VELOCITY 124
FIGURE 3-10 THE INDIVIDUAL COMPONENTS THAT CONTRIBUTE TO THE VARIABILITY OF BOTH CONTINUOUS WAVE AND PULSED WAVE DOPPLER. 127
FIGURE 3-11 THE REMOVAL OF READER AND OPERATOR VARIABILITY ONLY HAS A MODEST EFFECT ON OVERALL VARIABILITY (CONTINUOUS WAVE) 132
FIGURE 4-1 THE PRINCIPLE OF THE INSTANTANEOUS DIMENSIONLESS INDEX 149
FIGURE 4-2 THE SIMULTANEOUS MEASUREMENT OF THE LVOT AND PEAK VELOCITY CAN REDUCE MEASUREMENT VARIABILITY 150
FIGURE 4-3 A CONTINUOUS WAVE DOPPLER ACQUISITION FROM A PATIENT WITH AORTIC STENOSIS 156
FIGURE 4-4 AN EXAMPLE OF A CHALLENGING CONTINUOUS WAVE DOPPLER IMAGE FOR THRESHOLDING 158
FIGURE 4-5 HISTOGRAM OF ABOVE FIGURE 158
FIGURE 4-6 THRESHOLDING USING OTSU'S METHOD FAILS WHEN THE OUTER TRACE IS FAINT 159
FIGURE 4-7 RESULTS OF 16 DIFFERENT THRESHOLD ALGORITHMS ON THE EXAMPLE CONTINUOUS WAVE DOPPLER IMAGE 159
FIGURE 4-8 AN EXAMPLE PULSED-WAVE DOPPLER IMAGE 160

FIGURE 4-9 RESULTS OF 16 DIFFERENT THRESHOLD ALGORITHMS ON THE EXAMPLE PULSED-WAVE DOPPLER IMAGE 160
FIGURE 4-10 IDENTIFICATION OF THE OPTIMUM THRESHOLD TO CONVERT THE GRAYSCALE IMAGE TO A BINARY IMAGE FOR SUBSEQUENT TRACING 163
FIGURE 4-11 IMPROVED THRESHOLDING PERFORMANCE OF DESCRIBED METHOD AS COMPARED TO OTSU'S 164
FIGURE 4-12 STAGES OF TRACING AND PEAK DETECTION 166
FIGURE 4-13 DETECTION OF REGION OF PEAK IN EACH BEAT USING CONTINUOUS WAVELET TRANSFORM. 169
FIGURE 4-14 BEAT AVERAGING TO PRODUCE AN ENSEMBLED BEAT. 171
FIGURE 4-15 DETECTION OF THE INNER AND OUTER PROFILE FROM THE ENSEMBLED BEAT. 174
FIGURE 4-16 THE INSTANTANEOUS DIMENSIONLESS INDEX READINGS FROM 7 CONTINUOUS WAVE DOPPLER RECORDINGS IN ONE PATIENT 175
FIGURE 4-17 THE USE OF THE INSTANTANEOUS DIMENSIONLESS INDEX ON EXERCISE 216
FIGURE 4-18 VIEWING A FOLDER OF DICOM FILES 176
FIGURE 4-19 EXTRACTION OF THE STRIP, PEAK IDENTIFICATION, AND QUALITY CONTROL 178
FIGURE 4-20 VIEWING THE ENSEMBLED BEAT WITH THE INNER AND OUTER PROFILE 179
FIGURE 4-21 THE INTENSITY MAP 180
FIGURE 5-1 COMPARISON BETWEEN PW OF THE LVOT AND THE INNER PROFILE 191
FIGURE 5-2 COMPARISON OF THE VARIABILITY BETWEEN REPLICATE TYPICAL DI ASSESSMENTS (LEFT PANEL), REPLICATE IDI ASSESSMENTS (RIGHT PANEL), AND BETWEEN IDEALISED DI AND IDI (MIDDLE PANEL)..... 193
FIGURE 5-3 A SINGLE ACQUISITION FROM AN INDIVIDUAL PATIENT SHOWING HOW VARIANCE COMPOUNDS WHEN CALCULATING THE TRADITIONAL DIMENSIONLESS INDEX. 195
FIGURE 5-4 SIMULTANEOUS MEASUREMENTS CANCEL OUT VARIABILITY WITH THE INSTANTANEOUS DIMENSIONLESS INDEX 195
FIGURE 5-5 INDIVIDUAL VALUES FOR THE TYPICAL DIMENSIONLESS INDEX (DI), IDEALISED DI, AND THE INSTANTANEOUS DIMENSIONLESS INDEX (IDI) FOR THE FIRST 10 PATIENTS. 198
FIGURE 5-6 INTERACTION BETWEEN THE NUMBER OF REPEATED ACQUISITIONS (X-AXIS) AGAINST THE COEFFICIENT OF VARIATION (Y-AXIS). 199
FIGURE 5-7 AGREEMENT BETWEEN THE LVOT VELOCITY OBTAINED FROM PW AND FROM THE INNER PROFILE 201
FIGURE 5-8 AGREEMENT BETWEEN THE DIMENSIONLESS INDEX AND THE INSTANTANEOUS DIMENSIONLESS INDEX 202
FIGURE 5-9 INDIVIDUAL VALUES FOR THE TYPICAL CLINICAL DIMENSIONLESS INDEX, THE IDEALISED CLINICAL DIMENSIONLESS INDEX, AND THE INSTANTANEOUS DIMENSIONLESS INDEX 204
FIGURE 5-10 THE PORTION OF THE TRACED PROFILE USED FOR THE IDI CALCULATION 208
FIGURE 6-1 USING THE PISA METHOD TO CALCULATE THE EROA OF A REGURGITANT MITRAL VALVE 223
FIGURE 6-2 DEPENDING ON THE DIRECTION OF FLOW, THE SAME SPEED WILL HAVE A DIFFERENT VELOCITY COMPONENT IN THE DIRECTION "SEEN" BY THE ULTRASOUND PROBE 226
FIGURE 6-3 THE COLOUR DOPPLER PISA SHAPE IS MORE URCHINOID IN SHAPE THAN HEMISPHERICAL. COLOUR DOPPLER CAN ONLY MEASURE THE VELOCITY IN THE DIRECTION OF THE PROBE RATHER THAN THE TRUE SPEED. THEREFORE, THE THEORETICAL HEMISPHERICAL ISOSPEED CONTOURS BECOME URCHINOID ISOVELOCITY CONTOURS WHEN VISUALISED WITH DOPPLER 227
FIGURE 6-4 RELATIONSHIP BETWEEN DESCRIPTION PROVIDED OF PISA SHAPE AND MEASURED PISA DISTANCE. 230
FIGURE 6-5 THE MONA LISA PISA EFFECT 231
FIGURE 6-6 ACQUIRED COLOUR M-MODE OF SEVERE MITRAL REGURGITATION FOR AQURO 237
FIGURE 6-7 THE AQURO METHOD - INTRODUCTION 238
FIGURE 6-8 THE AQURO METHOD - INTERPRETING A SINGLE VERTICAL SCANLINE 239
FIGURE 6-9 THE AQURO METHOD - CALCULATING THE EROA FROM A SINGLE VERTICAL SCANLINE 241
FIGURE 6-10 THE AQURO METHOD - USING DATA FROM MULTIPLE VERTICAL SCANLINES TO CALCULATE THE EROA 242
FIGURE 6-11 COLOUR M-MODE THROUGH THE CONVERGENCE ZONE OF A REGURGITANT MITRAL VALVE. 248
FIGURE 6-12 COLOUR M-MODE THROUGH A CONVERGENCE ZONE SHOWING AN OFF-AXIS BEAT. 248
FIGURE 6-13 THE SEGMENTATION OF TWO EXCERPTS OF A COLOUR M-MODE DOPPLER STRIP INTO SYSTOLIC AND DIASTOLE 250
FIGURE 6-14 RESULTS AFTER TRAINING THE NEURAL NETWORK ON THE VALIDATION SET OF IMAGES. 259
FIGURE 6-15 SEGMENTATION OF THE COLOUR M-MODE STRIP INTO SYSTOLE AND DIASTOLE AND IDENTIFICATION OF THE ALIASING POINT. 260
FIGURE 6-16 SEGMENTS OF VERTICAL SCANLINES SHOWING MOST LINEAR CHANGE IN $VjetVr$ WITH DISTANCE ALONG THE VERTICAL SCANLINE 261
FIGURE 6-17 PLOT OF THE VELOCITY GRADIENTS OVER TIME 263

1 Introduction

1.1 Introduction

Cardiology, perhaps more than any other speciality, is a speciality founded on measurement. Whilst all specialties utilise quantitative measurements, the number and breadth of these measurements is probably greatest within cardiology, and particularly within echocardiography.

Measurements are never perfect – the same test, performed by the same observer, on the same patient will not produce the same value. The spread will be even wider when one considers different beats, probe positions, operators, and days.

Such knowledge is essential in day-to-day clinical practice. For example, if a clinician sees an apparent change in a measured value it is not automatically obvious whether this is a true change in an underlying value (which might have a clinical meaning), or simply random biological variability or measurement error (which does not).

Similarly, if a value is obtained that is near the threshold for treatment or diagnosis, without an understanding of the underlying variation the clinician may falsely confident of their categorisation (e.g. blood pressure).

For the *clinician* the variability of a test is important for the following reasons:

(1) **Minimum detectable difference**: The variability of a measurement limits the smallest change that can be reliably detected. The more reproducible a measurement the greater the confidence a clinician will have in an improvement or deterioration. For example, the reproducibility of ejection

fraction (EF) measurements limits the minimum detectable change for chemotherapy induced cardiomyopathy to approximately 10 EF units (Zamorano et al. 2016).

(2) **Effect on binary decisions:** Many decisions in medicine are based on whether a value is above or below a threshold. The reproducibility of a measurement affects the certainty with which this decision can be made. For example, central healthcare payers may put strict threshold on the EF at which a CRT will be reimbursed to the hospital. A chance high reading (such as 37%, rather than 33%) would lead to the patient being inappropriately refused the device (Brignole et al. 2013).

(3) **The need for repeated measurements.** If a measure has a significant amount of variability and is under serial surveillance (for example, the peak aortic velocity), occasionally a reading will by chance perhaps lower the last reading (which is not physiologically reasonable). The clinician may be suspicious of the results and may request a repeat scan to obtain a more expected result. These repeat scans cause patient inconvenience and anxiety, and increase cost.

For the *researcher*, wide variability of a measurement is harmful for the following reasons:

(1) **Increases the sample size of trials:** The sample size required for clinical trials increases the more variable a measurement is. The sample size linearly increases not with the standard deviation, but with the variance. That is to say, a doubling of the coefficient of variation of a measurement leads to a quadrupling of sample size required (assuming no treatment variability). Therefore, even small improvements in the precision of a measurement can result in large cost savings to trial funders (Altman 1990).

(2) **Limits the potential agreement between a new method and an old:** It is difficult for a new method to agree with an old method more than the old method agrees with its self. This can make it difficult for a new technique, even if it is more precise to supplant an existing technique. This concept is known as regression dilution (Spearman 1904). However, with careful experimental design and statistical analysis, it is possible to compensate for this (Chapter 4).

(3) **Limits the predictive ability of a variable:** Similarly, the noise on a measurement limits not only the statistical significant of a relationship (i.e. more patients are needed), but also if on the independent variable, reduces the observed magnitude (Spearman 1904) of the relationship (Figure 1-1).

(4) **Designing tests with an acceptable precision:** Without an known the true test-retest reproducibility of a measurement, it is difficult to know how to

improve it, or even that it needs improving (Joint Committee for Guides in Metrology 2008).

(5) **Understanding and recognising regression to the mean:** Often, physicians are referred patients because a biomarker (such as EF) has been measured and it has been found to have a clinically significant deviation from the norm. The physician, sceptical of measurements performed by others, repeats echocardiogram and obtains a more "normal" value. Annoyed, the physician curses the original sonographer and reassures the patient. The scenario is common, and in part, reflects regression to the mean. When a noisy variable is found to be significantly abnormal this often in parts represents a combination of true biological severity, but also a, by chance, lower than average reading. When repeating any measurement, it is statistically more likely to be less-extreme than the previous one. This phenomenon is called regression to the mean (Senn 2011). The greater the variability of the measurement, the greater the scope for regression to the mean.

Reliable sciences characteristically report precision with individual measurements.

"When reporting the result of a measurement of a physical quantity, it is obligatory that some quantitative indication of the quality of the result be given so that those who use it can assess its reliability. Without such an indication, measurement results cannot be compared, either among themselves or with reference values given in a

specification or standard." Evaluation of measurement data (Joint Committee for Guides in Metrology 2008).

However, in medicine in general and in cardiology specifically, this habit has not yet developed.

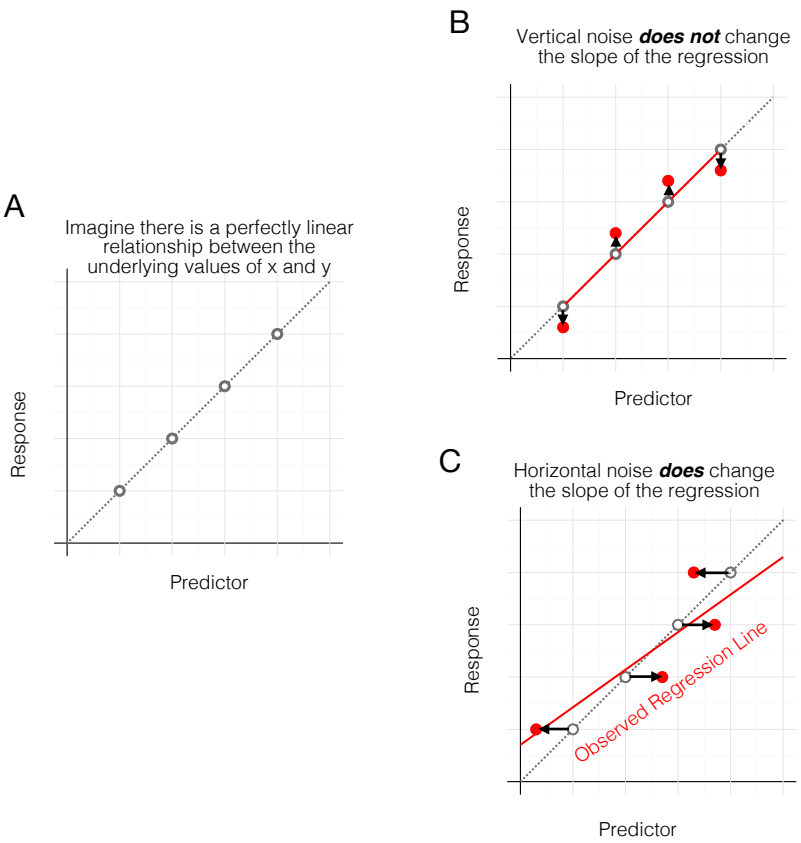


Figure 1-1 Noise on the x-axis but not noise on the y-axis biases the slope of the regression

Panel A: Imagine there is a perfectly linear relationship between the underlying values on the x- and y-axis. The regression line is shown as a dotted grey line.

Panel B: Noise on the y-axis ***does not*** affect the slope of regression. If there is noise in the measurement of the y-axis variable, this disturbs the data-points vertically in a random fashion. While this reduces the correlation coefficient, it does ***not*** reduce the regression slope (red) on average. An example of this is measuring E/E′ (y-axis) against Age (x-axis).

Panel C: Noise on the x-axis ***does affect*** both the slope of the regression and the correlation coefficient. If there is noise in the measurement of the x-axis variable, this disturbs the data-points horizontally in a random fashion. This reduces the correlation coefficient ***and*** reduces the regression slope (red) on average. An example of this is measuring pressure gradient (y-axis) against aortic valve area (x-axis).

1.2 Prevalence of valvular disease

The prevalence of moderate or more severe valvular heart disease is 2.5% across a general western population, with a meta-analysis of observational studies finding the prevalence increases with age: 0.7% in 18 to 44 year olds to 13.3% in those aged over 75 (Nkomo et al. 2006). In 2012 in the UK, significant valve disease led to 9,054 surgical aortic valve replacements, 7,524 mitral valve repairs or replacements (SCTS 2016), and 2,473 TAVIs (BCIS Audit 2016). Across the developed world aortic stenosis and mitral regurgitation are the dominant valvular diseases, accounting for 89% of significant valvular heart disease seen (Nkomo et al. 2006).

The prevalence of these two valvular diseases is therefore large. Diagnosed individuals typically require repeated screening. Therefore, quantification protocols need to be fast, cheap, and precise.

1.3 The role of imaging to assess aortic stenosis

Within current clinical practice, the quantification of aortic stenosis is directed at the identification of patients with enough of a combination of echocardiographic severity and symptoms to warrant intervention.

The echocardiographic grading uses flow-chart algorithms to combine multiple available biomarkers to aid a clinician in determining if the severity should be termed “severe” (Baumgartner et al. 2017).

With the prospect of disease modifying therapies for aortic stenosis (Dweck et al. 2013), the role of imaging biomarkers in aortic stenosis is changing. Biomarkers can be used to explore the biological pathways by which aortic stenosis develops and progresses (Pawade et al. 2016), and hence both evaluate the mechanistic effect of therapies, and predict the rate of progression. A researcher assessing the clinical efficacy of therapies requires that the biomarker is reproducible and responsive to changes, thereby maximising the signal-to-noise ratio and minimising the number of patients required within the trial.

1.4 Echocardiography and aortic valve disease

There is no true reference standard for aortic stenosis severity. Even a surgically explanted stenotic aortic valve will open to different extents depending on how much force is applied (Otto 2018). Echocardiography is the most widely used tool for the assessment of aortic stenosis. It has following desirable features:

(1) It reflects known physiology and pathophysiology.

(2) It is feasible in the majority of patients.

(3) It correlates with similar biomarkers.

(4) It predicts outcomes.

(5) It allows the selection of patients for therapy.

Careful work over the past four decades has shown that echocardiography meets the above requirements. Early work found that Doppler ultrasound could be used to examine the jets of stenotic and regurgitant lesions (Brubakk et al. 1977), laying the groundwork for Hatle (Hatle et al. 1980) to demonstrate not only that Doppler ultrasound was highly feasible (with traces obtained in 57 of 63 patients) in aortic stenosis, but also the peak jet velocity, converted into a gradient using the recently derived simplified Bernoulli equation (Holen et al. 1976), had relatively good agreement with invasive measurements. Subsequent work (Currie et al. 1985) confirmed this strong correlation between invasive and Doppler derived peak gradients. Otto (Otto et al. 1997), along with many others have demonstrated in prospective studies that echocardiographic biomarkers strongly predicted the need for intervention. Finally, the use of echocardiographic markers of severity as enrolment criteria in the PARTNER B randomized clinical trial (Leon et al. 2010) demonstrates its ability to allow selection of patients for beneficial therapy.

Whilst the peak gradient (expressed as a velocity) remains first line in the 2017 ESC aortic stenosis algorithm (Baumgartner et al. 2017), its dependence on LV function means that it cannot solely be relied on in to make clinical decisions. Furthermore, as the end-point for a clinical trial of therapy, this dependence on LV performance not only increases the variability of the measurement, but also clouds the assessment of possible mechanistic effects of therapies. Increases in cardiac output could be inappropriately interpreted as an increase in aortic stenosis severity.

Clinically, flow is always taken into consideration. In patients with high gradients, the algorithm ensures the clinician has excluded reversible high flow-states such as anaemia. In patients with a low gradient the physician is expected to make a judgement based on the and suspected significant aortic stenosis is suspected further calculations are required. Whilst guidelines (Baumgartner et al. 2009) have listed at least 10 other markers, the algorithm and clinical practice is dominated by the use of the aortic valve area (AVA) and to a lesser extent the Dimensionless Index (DI). Mathematically, by using ratio of the post-valve to pre-valve velocity profile, the AVA and DI corrects for flow. As with the peak and mean gradient, AVA and the DI have been found to predict outcome across multiple studies (Rusinaru et al. 2015; Bohbot et al. 2017).

Discordance between these measures can occur due to outright measurement error (such as misalignment of the jet), incorrect assumptions (such as the assuming the LVOT is circular), or from the inevitable random measurement and biological noise. More fundamentally, as clinical classification utilises arbitrary cut-points, there is potential for mis-calibration between the biomarkers. Across a population a mean gradient of 40mmHg results in an AVA closer to $0.8cm^2$ than $1.0cm^2$ (Minners et al. 2010), despite guidelines listing them as equivalent.

Nevertheless, disagreements between these multiple markers can truly reflect different aspects of pathophysiology, and many of these measures have additional predictive properties when combined (Rusinaru et al. 2015).

However, mechanistic discordance does exist, as the amount of opening varies with flow in a non-linear fashion. Fundamentally, there is no correction can be made for this at rest, and in these cases of low-flow low-gradient, flow must be temporarily increased (using dobutamine) to assess the valve at a typical flow rate. Alternatively, biomarkers from other modalities such as CT can be used (Pawade et al. 2018).

Researchers investigating the progression of, or modifying the progression of aortic stenosis, avoid this complexity of classification and focus on the within-patient changes of individual biomarkers. Better biomarkers have greater "signal-to-noise": the ratio between the standard deviation of serial changes to the magnitude of the average change. The higher the "signal-to-noise" the fewer the number of patients required, and the greater the ability to assess differences between sub-groups.

1.5 Echocardiography and mitral valve disease

As with aortic stenosis, transthoracic echocardiography has displaced invasive catheter assessment as the dominant method for the assessment of mitral regurgitation after symptoms (Tribouilloy et al. 1999).

The ESC (Baumgartner et al. 2017), ASE guidelines (Zoghbi et al. 2017), and expert reviews (Grayburn et al. 2014) recommend an integrated approach using multiple qualitative, and semi-quantitative parameters for the assessment of mitral regurgitation severity. However, only one quantitative technique, the EROA (effective regurgitant orifice area) calculated using the method PISA (and its

echocardiographic derivatives the regurgitant volume and fraction) is recommended (Chapter 5).

This recommendation is based on:

(1) **EROA reflecting regurgitant severity:** The PISA derived EROA cut-offs are reflect previous angiographic grading of mitral regurgitation (Dujardin et al. 1997).

(2) **EROA is prognostic:** EROA predicts prognosis in the asymptomatic patient. Even when adjusted for ejection fraction and presence of atrial fibrillation, an EROA of 20 to 39 mm^2 had a RR of 3.1, and >40 mm^2 of 5.2 as compared to those with an EROA < 20 mm^2 (Enriquez-Sarano et al. 2005).

As compared to aortic stenosis, the assessment of mitral regurgitation severity is more complex:

(1) **The aetiology has affected the cut-offs for severity**: The aetiology of the mitral regurgitation influences the quantification and management mitral regurgitation. Previously, the cut-off for severe *secondary* mitral regurgitation has been changed from >0.4cm^2 to >0.2cm^2 in 2014 (Marwick et al. 2014), before being increased again to >0.4cm^2 in 2017 (Grayburn et al. 2017). I.e. in the space of 3 years the cut-off was first reduced by 50%, then increased by 100%. This reflects not only the difficulties of assessment, but also what we want "severe" to mean when describing a lesion.

Furthermore, the original calculations of the cut-offs were obtained from a single study, enrolling 190 patients, and comparing it to LV angiography (Dujardin et al. 1997)

(2) **The aetiology affects the assessment of severity:** The regurgitant orifice of *secondary* mitral regurgitation is more crescentic in shape (Grayburn et al. 2014) and therefore the underlying assumptions are less valid.

(3) **The aetiology affects the management**: The importance of the above issues is somewhat diminished as isolated surgical intervention for mitral regurgitation is not recommended for *secondary* mitral regurgitation. However, the issues of calibration and violation of assumptions still apply, to a lesser extent with *primary* mitral regurgitation

(4) **Methodological difficulties in calculating the EROA by PISA:** Multiple issues, ranging from the measurement of a small, poorly defined distance, to the use of only a single snap-shot frame complicated the accurate assessment of the EROA by PISA (Section 5.2.1).

These issues result in great uncertainty in the assessment of mitral regurgitation for clinicians and researchers. This limits work to understand the underlying physiology, the development of new technologies, and decisions on the timing of surgery.

1.6 Quantitative echo in practice

Current European (Baumgartner et al. 2017) and North American (Grayburn et al. 2017) guidelines recommend regular follow-up, often at least yearly. Aside from symptoms, echocardiography has long been the mainstay of the assessment and quantification of these two valvular lesions.

Despite the echocardiography being the cornerstone of quantification, there are many features of the technique that would appear strange to metrological experts. Indeed, at multiple levels clinical staff have to manipulate the data to produce an acceptable experience for the patients:

(1) It is normal practice for the physician to look at the previous echocardiographic reports to get an approximate idea for how severe the valve disease is before performing the measurement. If we had provided the technician with reliable tests they would not need to do this routinely. For example, to measure a patient's weight, it is not required to enter their previous weight.

(2) Physicians often re-analyse the same acquired image when they are unhappy with the original result, with the aim of getting a different, supposedly more congruous result.

(3) Physicians reflect over many years of data to ensure that any action is based on a genuine trend, rather than a one-off fluke result. This might be a

manifestation of valve disease fluctuating in intensity, but one should only accept this if the test itself it reproducible. Otherwise this might just be measurement error.

(4) Recommendations make multiple measurements (e.g. 3 or 5) and average are routinely ignored. Furthermore, no rational is given for the number of measurements to be made, nor the overall precision that would be expected provided.

1.7 The reproducibility crisis

As clinicians we have to be adept in managing clinical variables that are poorly reproducible, ranging from patient symptoms to complex endocrine profiles. Such variables, whilst essential to the practice of clinical medicine, may be so irreproducible that they would not be accepted in a physical science. Nevertheless, physicians and their patients manage. Years of experience teaches them to observe and discard outliers, and to distinguish them from one-off flukes. For some variables this will always be unavoidable, but if we always assume that the situation cannot be improved upon, we will never identify the things can be improved upon.

1.8 The importance of measurement precision

The cardiology literature is littered with biomarkers (Figure 1-2), some have stood the test of time and are routinely used, and some have fallen by the wayside (Tzoulaki et al. 2011). Aortic stenosis is no exception, with recent guidelines providing guidance for the use of no less than 12 imaging biomarkers (Lancellotti et

al. 2013). Three of these biomarkers are directly derived from a single measurement; the other nine are a composite of up to six other measurements (Table 1.1).

Table 1.1 Measurements required for echocardiographic biomarkers

Biomarker	Measurements required
AS jet velocity	1
Mean gradient	1
Continuity equation valve area	3
Simplified continuity equation	3
Velocity ratio	2
Planimetry of valve area	1
LV % stroke work loss	2
Recovered pressure gradient	4
Energy loss index	6
Valvo-arterial impedance	4
Aortic valve resistance	3
Projected valve area at normal flow rate	3 (multiple repetitions required)

The use of imaging biomarkers (like all biomarkers) is an attempt to usefully model physiology so that it is interpretable by a clinician. Increasing model parameterisation to better reflect complex physiology is attractive but has a substantial drawback. The addition of each measured parameter to the biomarker increases the overall measurement as the errors become compounded and uncertainty propagates (Taylor & Kuyat 1994).

The increase in "uncertainty" of a biomarker dependent on more parameters may negate any improvements derived from the greater complexity. The more parsimonious biomarker may remain as the better predictor of morbidity or response.

In the development of risk model from a dataset of 107 patients with aortic stenosis, and including peak velocity, mean pressure gradient, aortic valve area, and indexed valve area, only peak velocity, despite its simplicity, was independently associated with death or valve intervention (Monin et al. 2009).

1.9 Importance of open biomarkers

The ever-increasing number of biomarkers in medicine (Figure 1-2) and echocardiography is unfortunately not mirrored by a rise in open-source protocols, algorithms, and software available for independent researchers or clinicians to test, validate, and use. This leads to two problems:

(1) As the algorithms are a black-box there are no guarantees that different commercial implementations will provide the same results. For example, one group, scanned patients and processed the videos with five different commercially available packages (Nagata et al. 2015). They found significant differences not only between the vendors, but also between different versions by the same manufacturer.

(2) Researchers are forced to use the software "as-designed". They are unable to, for example, easily collect and export multiple beats worth of data.

The algorithms and code I have developed within this PhD will therefore be released as open-source. Furthermore, I have packaged it up into easy to use software which I

have distributed to other research institutions for use in their clinical studies and trials, from which they have supplied independent validation data (Chapter 4).

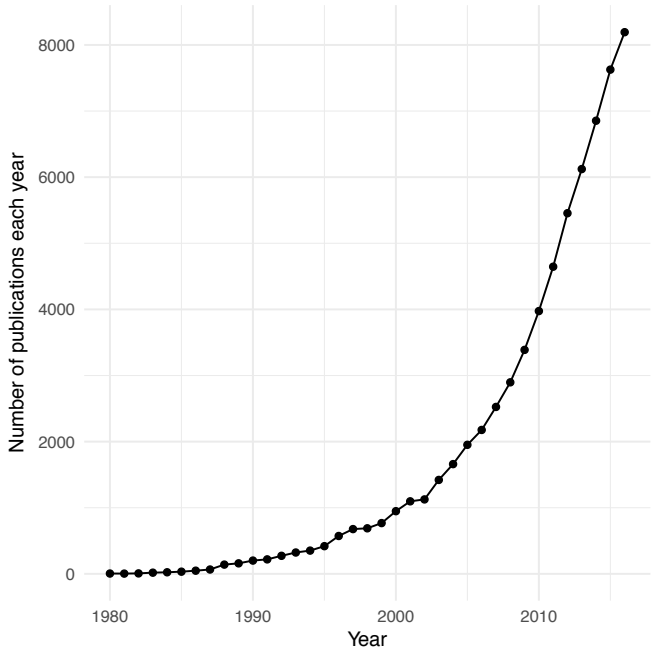


Figure 1-2 Number of new publications each year describing a "novel biomarker"

A Pubmed search was performed for "novel biomarker" with the publication date restricted to each year from 1980 to 2017. The number of articles published for each year is plotted against year. The number of articles published each year describing a "novel biomarker" has increased every year since 1980.

1.10 Going forward

Within this thesis I will address the following topics:

(1) Quantify the genuine reproducibility of cardiovascular variables (Chapter 2).

(2) Perform a reproducibility study to break down the *true* sources of noise in the measurement of aortic stenosis (Chapter 3), allowing them to be specifically targeted.

(3) Develop a more reproducible, clinically applicable method for quantifying aortic stenosis severity, the "instantaneous Dimensionless Index". This will be integrated into my easy to use, open-source, echocardiography quantification platform. I will then perform validation and testing of the instantaneous Dimensionless Index against conventional measures at both my own and an independent clinician's institution (Chapter 4).

(4) Further develop and enhance AQURO (Moraldo, Bergamini, et al. 2013), a more reproducible, clinically applicable method for quantifying mitral regurgitation severity, and integrate it into our echocardiography quantification platform "MagiQuant". I will then perform a pilot validation study at my own institution (Chapter 5).

In developing improved algorithms and protocols for the quantification of these valvular disease I will take several steps to identify the sources of variability, propose and implement solutions. I will then prospectively test their success.

2 Can we rely on reproducibility studies? A systematic review and meta-analysis of 6727 patients from 97 studies.

2.1 Abstract

2.1.1 Background

Sometimes, when a biomarker is published to have very good reproducibility, local experience can be poorer. The temptation is to assume that we are performing the test inexpertly. An alternative possibility is that the published reproducibility studies have, perhaps inadvertently, reported an artificially small value. This chapter explores this puzzle.

There are two sources of published information quantifying the reproducibility of imaging biomarkers. The first source is composed of studies specifically designed to report reproducibility. The second source is composed of the control arms of blinded randomised controlled trials which unintentionally report reproducibility because they restudy patients who have undergone no intervention at a later date. In this chapter I look for a contrast between these two types of reports of reproducibility for a range of imaging biomarkers.

Hypothesis: There is a systematic difference between the reproducibility of imaging biomarkers observed in randomised controlled clinical trials, and that reported in dedicated reproducibility studies.

2.1.2 Method

I systematically reviewed published studies referenced by ESC echocardiography guidelines which were designed to report reproducibility, and separately searched

for randomised controlled trials that unintentionally reported the reproducibility because of test and re-test scans in the control arm.

2.1.3 Results

Across the 329 ESC guideline referenced papers, 60 provided useable data on reproducibility with 295 assessments across 75 unique standard 2d-echocardiographic parameters. 38 trials with usable test-retest data in a placebo arm with 145 end-points across 39 unique echocardiographic were found.

There was a consistent pattern across echocardiographic biomarkers. In 16 of the 18 echocardiographic parameters in which test-retest data was available from both reproducibility studies and RCTs, the coefficient of variation was narrower in studies designed to test reproducibility than in the control arms of randomised controlled trials ($p<0.001$). Of note, for non-contrast LVEF, the test-retest coefficient of variation derived from RCTs was 2.8 times larger than that found in reproducibility studies (CV 19.2% vs 6.8%, $p=0.03$). On average, the test-retest CV derived from RCTs was 2.3 times larger than that from reproducibility studies.

2.1.4 Conclusions

Studies presented as quantifications of reproducibility of imaging biomarkers consistently report agreement on average 2.3-fold better than that seen in the control arms of clinical trials. One explanation may be that reproducibility studies fail to capture all sources of variability that develop over time elapsing between scans in RCTs. Researchers designing trials with imaging endpoints may unintendedly

designing their studies 5.1-fold too small if they rely on data from reproducibility studies.

2.2 Introduction

2.2.1 Why reproducibility is important

2.2.1.1 Clinical practice

Physicians planning efficient clinical services require reliable information on the test-retest reproducibility of echocardiographic measurements in circumstances that they can realistically be expected to achieve. For example, a patient with breast cancer having Herceptin treatment has attended for a repeat echocardiogram showing an ejection fraction of 47%, where three months previously it measured 53%. The clinician needs to know whether the observed decrease in EF is within the spontaneous variability of the test or is a true biological change. If the former, the treatment can continue. If the latter, the treatment should stop (Zamorano et al. 2016).

2.2.1.2 Clinical research

Similarly, a clinical trialist is planning a randomised control trial (RCT) to detect the ability of a new therapy to reduce the progression of aortic stenosis as assessed by peak velocity (V_{max}). She wonders how many patients to include in each arm to detect a particular size of change. She understands the need to know the variability of the measured parameter for sample size calculations. Will her calculation be accurate if she relies on this data?

2.2.2 How is reproducibility measured

Both the physician and researcher need to know the standard deviation of the difference (SDD). This is the standard deviation of the difference between two

measurements of the parameter under study. The SDD allows our first clinician to estimate how much random fluctuation to expect in a measurement. For example, if the SDD was 4 EF units, then you would expect 95% of the differences to lie within $\pm(4 \times 1.96)$ which approximates ±8 EF units; therefore she cannot be confident that a fall of 6 EF units truly represents a decline in left ventricular function.

The number of patients needed in each arm is proportional to the square of the SDD and is therefore a powerful determinant of sample size. If for example a novel technique of measuring aortic V_{max} produced half the SDD she would require a quarter of the sample size.

2.2.3 Potential sources of reproducibility data

Within the literature there are two possible sources of reproducibility data for biomarkers.

The first source are dedicated reproducibility studies which have examined the parameter in question. These studies are specifically designed to assess the reproducibility of biomarkers and reported as such.

A second, less obvious source are the placebo arms of RCTs that have measured the biomarker the beginning and the end of the study. These will naturally report the SDD of change of a variable over a clinically relevant period of time.

The availability of two separate sources of information allows for a comparison between the two. If systematic differences were found it may lead to a better understanding of how reproducibility is assessed and reported.

2.2.3.1 Reporting of reproducibility data

The SDD can be easily derived by calculating the standard deviation of the differences between two successively measured values. In the more general cause, where more than two measurements are made, the *within-patient* standard deviation (SD) of measurements can be calculated and presented. The is the standard deviation of a series of measurements within the same patient. The SD is smaller than the SDD by a factor, on average of $\sqrt{2} \approx 1.414$ and they can be freely converted. The coefficient of variation (also known as the within patient coefficient of variation) can approximated by dividing the SD by the mean value. Care must be taken as many reproducibility papers incorrectly calculate the CV by effectively dividing the SDD by the mean, producing a value that is approximately $\sqrt{2} \approx 1.414$ too large. Within this thesis I will use correct definition of the coefficient of variation (Bland 2006).

2.2.4 Aims of this study

Clinicians and researchers need data on the reproducibility of imaging biomarkers

Typically, reproducibility and validity data for biomarker is first derived from dedicated reproducibility studies. Those that are found to be both reproducible and predictive may be used in later RCTs as mechanistic or surrogate outcome measures.

Placebo-controlled RCTs, in which these biomarkers are used and measured at the beginning and end of a trial, also produce data on the reproducibility of the biomarker.

This allows interested clinicians to compare the estimates of the variability of the biomarker from both dedicated reproducibility studies and RCTs. If there was a systematic difference between the two sources of data then clinicians and researchers might reasonably be concerned that one source of reproducibility data had limited utility. Such information may allow novel biomarkers to be more critically assessed.

In this study I sought to:

- determine the coefficient of variation for many commonly used 2D echocardiographic biomarkers, and,
- determine if there was a systematic difference between the reproducibility studies and RCT estimates of the CV for each imaging biomarker.

Hypothesis: There is a systematic difference between the reproducibility of imaging biomarkers observed in randomised controlled clinical trials, and that reported in dedicated reproducibility studies.

2.3 Methods

2.3.1 Eligibility and search strategy

2.3.1.1 Search strategy for reproducibility studies

To replicate what a diligent clinician might do to form an opinion about the reproducibility of 2D-echocardiographic biomarkers I searched reference lists of the most recent ESC guidelines relating to echocardiography:

- Standardization of adult transthoracic echocardiography reporting in agreement with recent chamber quantification, diastolic function, and heart valve disease recommendations: an expert consensus document of the European Association of Cardiovascular Imaging (Galderisi et al. 2017).

- Echocardiographic assessment of valve stenosis: EAE/ASE recommendations for clinical practice (Baumgartner et al. 2009).

- Recommendations for the echocardiographic assessment of native valvular regurgitation: an executive summary from the European Association of Cardiovascular Imaging (Lancellotti et al. 2013).

- Recommendations of the European Association of Echocardiography: How to use echo-Doppler in clinical trials (Galderisi et al. 2011).

These guidelines are published by the European Society of Cardiology and its affiliated groups (the European Association of Cardiovascular Imaging). Similar guidelines are available from other large national bodies, such as the American

Society of Echocardiography (ASE). Practically, there is significant overlap between the two national agencies documents. One guideline was produced as a joint recommendation by the ESC and ASE (Baumgartner et al. 2009). One guideline (Galderisi et al. 2017) is a summary of recommendations of three joint ESC and ASE guidelines (Lang et al. 2015; Nagueh et al. 2016; Lancellotti et al. 2016). One guideline (Lancellotti et al. 2013) on valvular regurgitation is a update of two previous ESC guidelines on valvular regurgitation (Lancellotti, Tribouilloy, et al. 2010; Lancellotti, Moura, et al. 2010), which were updates of a joint ESC/ASE guideline on valvular regurgitation (Zoghbi 2003). Only one of the guidelines (Galderisi et al. 2011) was not produced in collaboration with ASE, but is more recent than the equivalent guideline (Douglas et al. 2009).

References to meta-analyses were followed to further extract their constituent studies. There were no language restrictions. I only considered studies in adult humans with either normal or acquired disease. Studies examining congenital heart disease or transplanted hearts were excluded.

2.3.1.2 Search strategy for randomised controlled trials

Similarly, to replicate what a clinician might do to find RCT trials that had an echocardiographic endpoint, I searched clinicaltrials.gov for registered randomized clinical trials with "results available" and the search-term "echocardiography". This was supplemented with the RCTs from the most recent Cochrane reviews of "Statins for Aortic Stenosis". I excluded trials in children, of patients with congenital disease,

and in transplanted hearts. I was only interested in groups that did not have an active intervention and excluded those comparing two (or more) active interventions. I also excluded studies in which both arms would be subjected to significant biological changes over time (e.g. recruited immediately after myocardial infarction).

2.3.2 Data abstraction

Two authors (myself and a colleague Chris Pavitt) independently extracted the data and any disagreements were resolved by consensus and if needed a senior author (DF). From the reproducibility studies we extracted the year of publication, time interval between measurements (days), the number of patients, the reproducibility test conditions, the mean population parameter value, and the coefficient of variation of each parameter. Some papers did not explicitly report the coefficient of variation but did report data that allowed its calculation.

2.3.3 Calculation of standard deviation of difference from reproducibility studies

2.3.3.1 Studies reporting Bland-Altman limits of agreement only

In studies reporting the Bland-Altman 95% limits of agreement the lower limit was subtracted from the upper limit and the result divided by 2×1.96, to obtain the SDD. This was divided by $\sqrt{2}$ to obtain the within-patient SD, and then by the mean value to obtain the coefficient of variation (Bland 2005).

2.3.3.2 Studies presenting correlation plots of raw data

If the graph was obtained as a vector image within a PDF (i.e. when zoomed in to the graph the edges of the lines and dots remain crisp), the PDF was opened in Inkscape (The Inkscape Team 2018), a vector drawing program and image features aside from the dots and the lines representing the axes were removed. The file was then saved as a Scalable Vector Graphic (SVG) file. I have developed custom scripts that are able to extract the raw co-ordinates of the data-points and axes and recreate the data-set.

If the graph was obtained as a raster image (i.e. when zoomed in to the graph it becomes blurry or pixelated), the graph was extracted and opened using plot digitising (Mitchell et al. 2018) software.

The CV was then calculated from the extracted data.

2.3.3.3 Studies presenting the mean absolute deviation

The standard deviation is the square-root of the mean of the squared deviations from the mean. Some studies reported the mean of the absolute deviations between the first and second measurement. In general, the mean absolute deviation is typically smaller than the standard deviation (Pham-Gia & Hung 2001). The SDD can be approximated (with the assumption that the underlying data is normally distributed) by *multiplying* the mean absolute deviation by $\sqrt{\frac{\pi}{2}} \approx 1.2533$. The *within-*

patient SD can therefore be approximated by *dividing* the mean absolute deviation by $\frac{2}{\sqrt{\pi}} \approx 1.128$.

2.3.3.4 Studies presenting the median absolute deviation

Some studies reported the median of the absolute deviations from the median. In general, the median absolute deviation is smaller than standard deviation. The SDD can be approximated (with the assumption that the underlying data is normal) by multiplying the median absolute deviation by $1/\Phi^{-1}(3/4) \approx 1.4826$ where $\Phi^{-1}$is the reciprocal of the Normal quantile function (Pham-Gia & Hung 2001). The *within-patient* SD can be calculated by *dividing* the median absolute deviation by ~ 1.048. Care must be taken as both the *Mean* and *Median* Absolute Deviation can be abbreviated to MAD within papers.

2.3.4 Calculation of the standard deviation of difference from randomised controlled trials

Typically, trials present the change from the beginning to the end of the trial in the placebo (and intervention) arm, alongside the primary comparison. If this data is presented, often it is also presented with the standard error (SE) of this mean change. To convert this to SDD, the SE is multiplied by the square-root of the number of patients in that arm, and then the CV calculated. Care must be taken as sometimes the standard deviation will be presented (despite referring to the SE) instead, and vice versa. If unlabelled, the magnitude of the value was used to help decide which value was presented. If this value was not explicitly presented the following techniques were used to calculate it.

2.3.4.1 Studies reporting the confidence interval of the mean change

If the 95% confidence interval was presented, the lower limit was subtracted from the upper limit and divided by 2×1.96 to generate the standard error. This was then multiplied by the square-root of the number of patients in the trial arm to generate the SDD (Altman & Bland 2005).

2.3.4.2 Studies with the plots of the mean change in each arm and confidence interval or standard error

Plot digitising software (Mitchell et al. 2018) was used to extract the mean and presented intervals. They were then processed as in Section 2.3.4.1.

2.3.4.3 Studies reporting the mean change and p-value for the change

If the study presented the p-value to two or more decimal places and the mean change, the p-value was transformed using the Student t-distribution into a t-score. The standard error was calculated by dividing the mean by the t-score, which was then processed as in Section 2.3.4.1.

2.3.4.4 Studies reporting the intraclass correlation coefficient

The intraclass correlation coefficient (ICC) is the ratio of the variance of the true value (i.e. if it were measured without error) to the variance of the measured value (i.e. measured, as it always is, with error). These components are typically calculated using ANOVA or a random-effects model. If the ICC is provided along with the population standard deviation, the within patient standard deviation can be calculated as follows (Bland 2004).

$$SD_{within-patient} = \sqrt[2]{(1 - ICC) \times (SD_{measured})^2}$$

The coefficient of variation can then be calculated by dividing by the mean.

2.3.5 Reproducibility test conditions

It is important to consider the conditions under which a "reproducibility" study is conducted as these will have a significant effect on the magnitude of the variability.

The International Standards Organisation defined two test conditions for this (International Standards Organisation 2017):

- **Reproducibility conditions**: In which as many relevant parameters are made as different as possible – e.g the same patient is assessed by a different lab, on a different day, by a different operator, and a different reader.

- **Repeatable conditions**: In which as many relevant parameters are made the same as possible – e.g. the same patient is assessed by the same operator, with the shortest possible time interval, and the images read by the same reader.

Unfortunately, the cardiology literature does not use standard descriptions. For example, inter-observer may be taken to mean a different person analysing the same images, or a different person acquiring a new set of images and then analysing them.

Within this chapter I will use the terms:

- **Intra-reader** to refer to the ***same*** reader analysing the ***same*** set of acquired images (at any time interval).

- **Inter-reader** to refer to a ***different*** reader analysing the ***same*** set of acquired images (at any time interval).

- **Test-retest** to refer to a ***new*** set of images being acquired (by either the same of a different sonographer and analysed by the same or different reader). All data from the placebo arm of RCTs will fall into this category.

The latter category encompasses a few possible test conditions. In the Chapter 3 I will detail these and conduct a reproducibility study of my own to accurately attribute the sources of variability.

If RCTs reported multiple time-points, or if the reproducibility studies reported multiple protocols these were separately extracted.

2.3.6 Data analysis

Using the SDD and mean parameter value we approximated the coefficient of variation (CV) for each reproducibility experiment and RCT end-point. I did this to allow comparison of the parameters on the same (unitless) scale. To assess if the CV was biased in one direction between reproducibility and RCT data the Mann-Whitney U-test was used. Analyses were performed using the statistical environment "R" (R Core Team 2016) using the ggplot2 package (Wickham 2009).

2.4 Results

2.4.1 Results of the searches

2.4.1.1 Reproducibility studies

The reproducibility study search strategy identified 329 papers. From these (including studies from referenced meta-analyses), 60 studies provided useable data on reproducibility. In total, the reducibility studies provided 295 assessments across 75 unique parameters measured on standard 2d-echocardiography.

Across the 295 reproducibility assessments, 108 were from the same reader re-measuring an already acquired set of images and 116 were from a different reader re-measuring an already acquire set of images. In only 71 reproducibility assessments covering 44 parameters was a new set of images acquired to assess the test-retest variability. This test-retest data was provided by only 15 of 60 (25%) of the included studies. The time interval between scans ranged from 1 hour to 12 months. The median interval was 1 day (IQR 1 hour to 10 days).

2.4.1.2 Randomised controlled trials

The RCT search strategy identified 221 studies. Including the four studies (SALTIRE, SEAS, ASTRONOMER, and PROCAS) from a recent Cochran meta-analysis for "Statins for aortic valve stenosis"(Thiago et al. 2016), 37 RCTs met the inclusion criteria, providing useable test-retest data in the placebo arm. In total, the RCTs provided 144 end-points across 39 unique parameters on standard 2d-echocardiography. The median interval between scans was 182 days (IQR 141 to 365,

mean 216), and the median number of patients for which data was available was 32 (IQR 13 to 85).

2.4.2 Intra-reader reproducibility of echocardiographic measurements

50 Studies provided 108 assessments of intra-reader reproducibility across 61 unique parameters. The intra-reader CV ranged from 1.5% for the dimensionless index of aortic stenosis to 16.8% for the left ventricular mass. The intra-reader CV for a non-contrast left ventricular ejection fraction was 9.4% from data across 7 studies including 172 patients. The raw and summarised data are shown in Figure 2-1 and Figure 2-2 respectively.

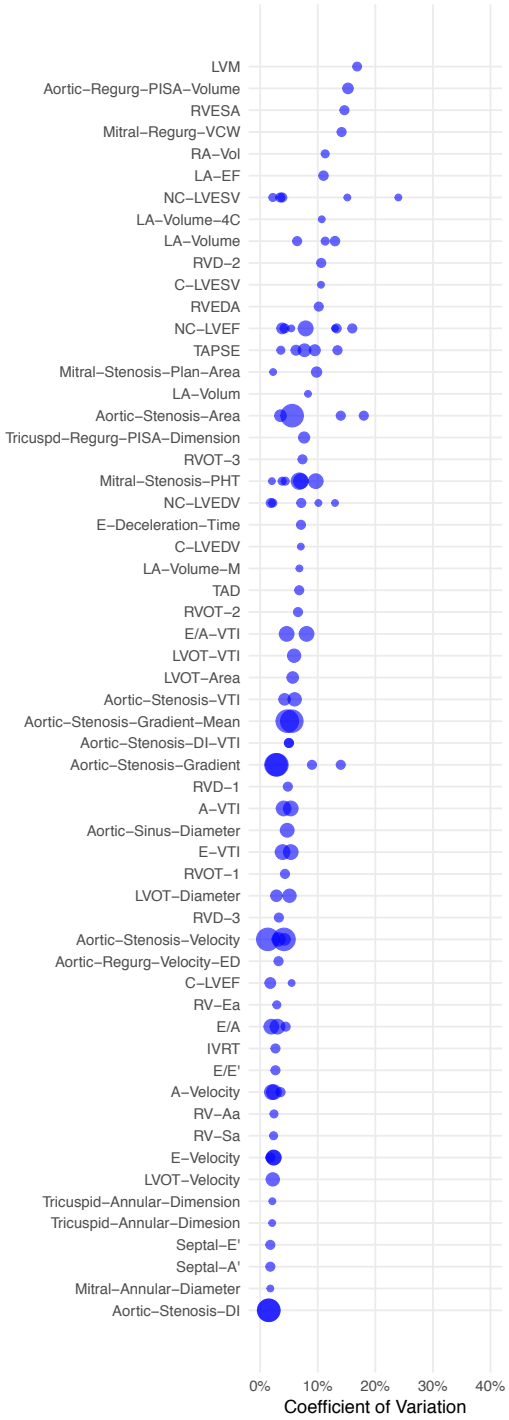


Figure 2-1 Individual intra-reader coefficients of variation from reproducibility studies

Intra-reader coefficients of variation, calculated from the measurements made by the same reader, re-measuring the same set of acquired images. Each dot represents one study. The size of the dot is proportional to the size of the study.

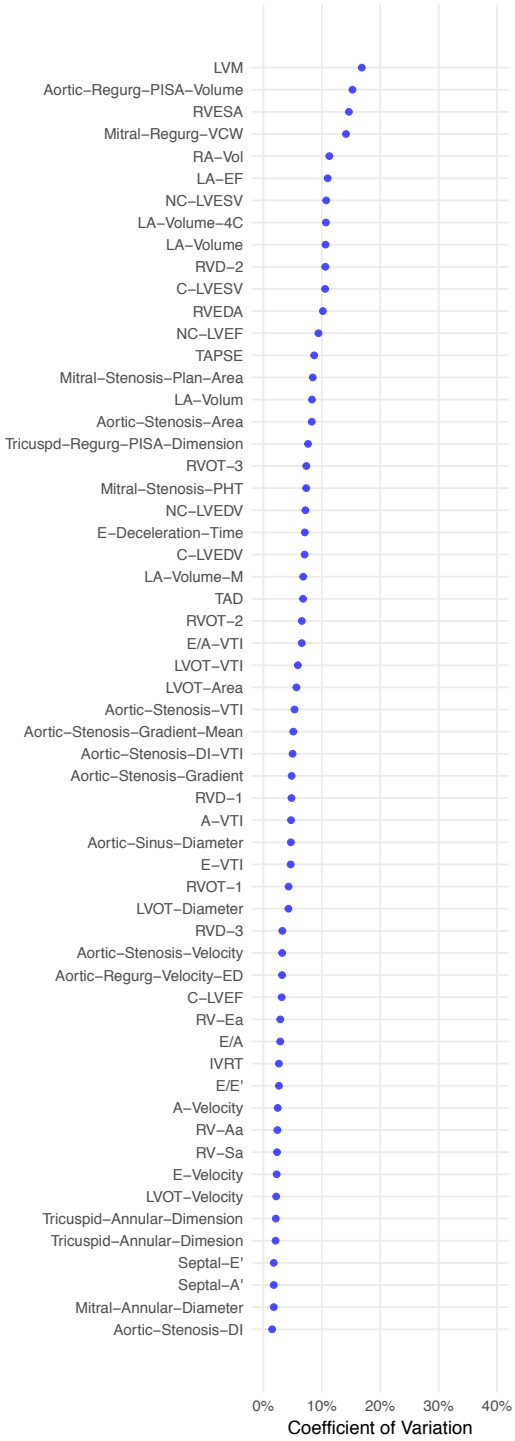


Figure 2-2 Summarised intra-reader coefficients of variation from reproducibility studies

Intra-reader coefficients of variation, calculated from the measurements made by the same reader, re-measuring the same set of acquired images. Each dot represents the weighted root mean square of the individual coefficients of variation.

2.4.3 Inter-reader reproducibility of echocardiographic measurements

51 Studies provided 116 assessments of intra-reader reproducibility across 61 unique parameters. The inter-reader CV ranged from 1.6% for the peak jet velocity of aortic stenosis to 18.9% for the left ventricular mass. The inter-reader CV for a non-contrast left ventricular ejection fraction was 9.3% based on data from 10 studies including 1574 patients. The individual study and summarised results are shown in Figure 2-3 and Figure 2-4 respectively.

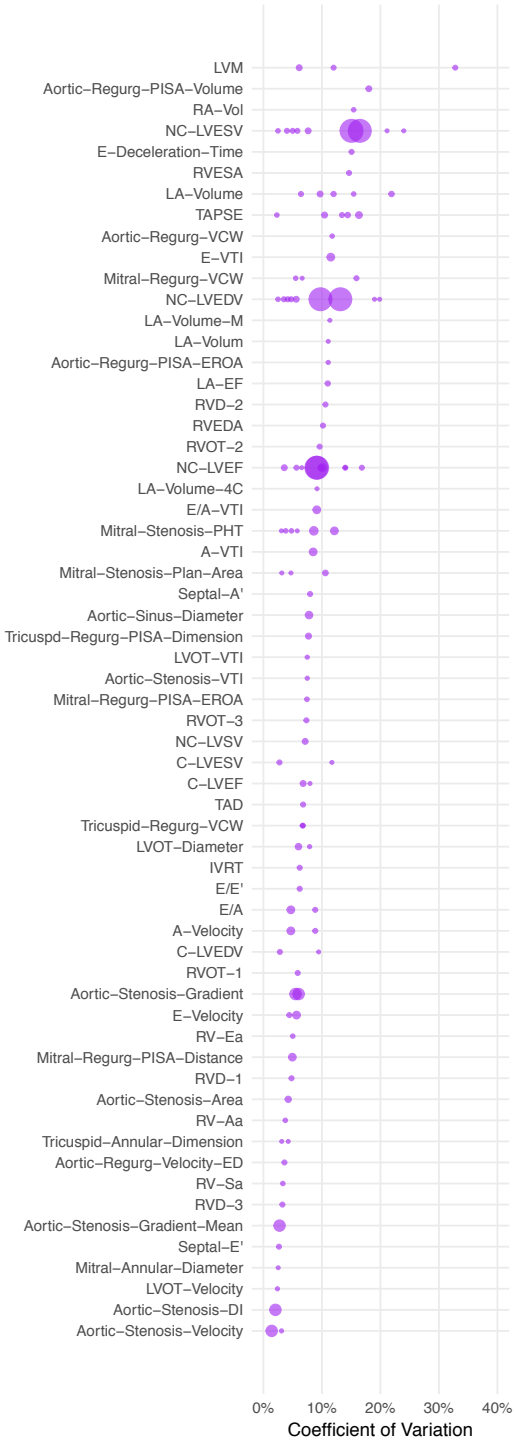


Figure 2-3 Individual inter-reader coefficients of variation from reproducibility studies

Inter-reader coefficients of variation, calculated from the measurements made by a different reader, re-measuring the same set of acquired images. Each dot represents one study. The size of the dot is proportional to the size of the study.



Figure 2-4 Weighted mean inter-reader coefficients of variation

Inter-reader coefficients of variation, calculated from the measurements made by a different reader, re-measuring the same set of acquired images. Each dot represents the weighted root mean square of the individual coefficients of variation.

2.4.4 Test-retest variability of echocardiographic measurements

17 studies provided 71 assessments of the test-retest reproducibility across 44 unique parameters. The claimed intra-reader CV ranged from 3.0% for the right ventricular apex-to-base distance (RVD-3) to 19.7% for the measurement of the contract LV end-systolic volume. The intra-reader CV for a non-contrast left ventricular ejection fraction was 6.8% based on data from 6 studies including 228 patients. The individual study and summarised results are shown in Figure 2-5 and Figure 2-6 respectively.

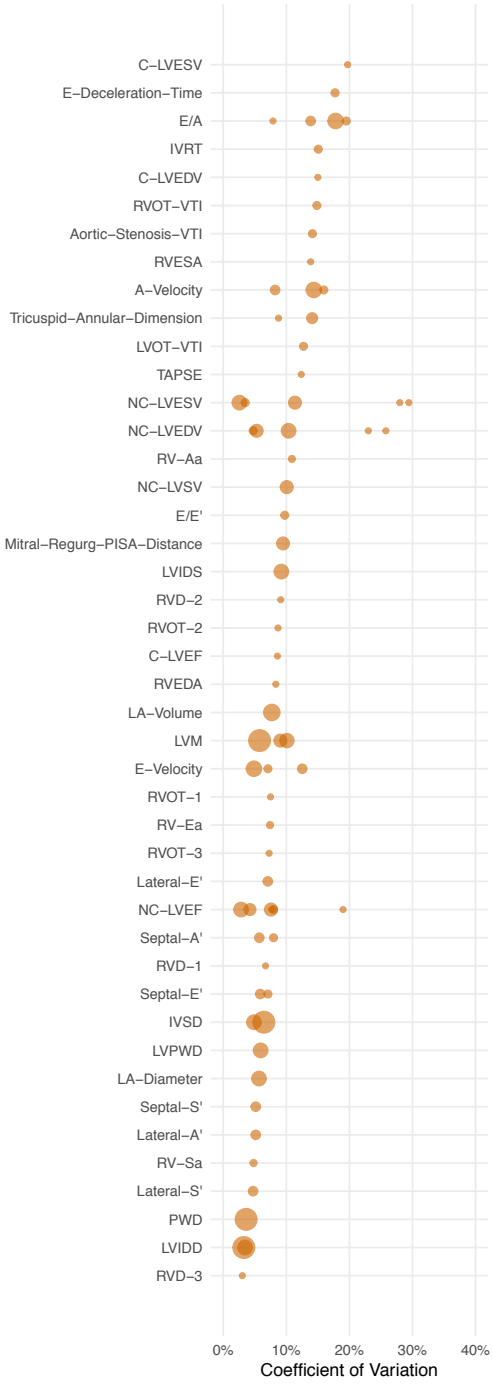


Figure 2-5 Individual test-retest coefficients of variation from reproducibility studies

Test-retest coefficients of variation, calculated from measurements made from two separate image acquisitions. Each dot represents one study. The size of the dot is proportional to the size of the study.

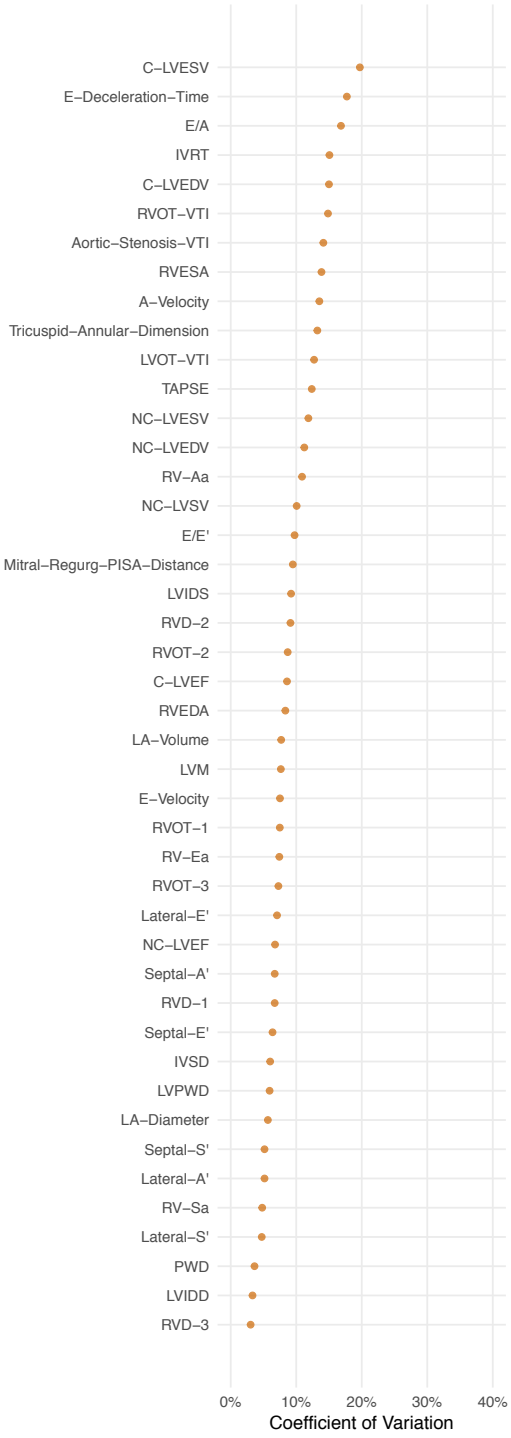


Figure 2-6 Weighted mean test-retest coefficients of variation from reproducibility studies

Test-retest coefficients of variation, calculated from measurements made from two separate image acquisitions. Each dot represents the weighted root mean square of the individual coefficients of variation.

2.4.5 Test-retest variability from randomised controlled trials

The placebo-arm of 37 randomised controlled trials (including 4071 patients) provided 144 assessments of the test-retest reproducibility across 39 unique parameters. The intra-reader CV ranged from 1.7% for the left atrial width, to 42.9% for the measurement of the PISA derived EROA in mitral regurgitation. The placebo-arm derived test-retest CV for a non-contrast left ventricular ejection fraction was 19.2% based on data from 29 end-points from 20 trials, including 2156 patients. The individual study and summarised results are shown in Figure 2-7 and Figure 2-8.

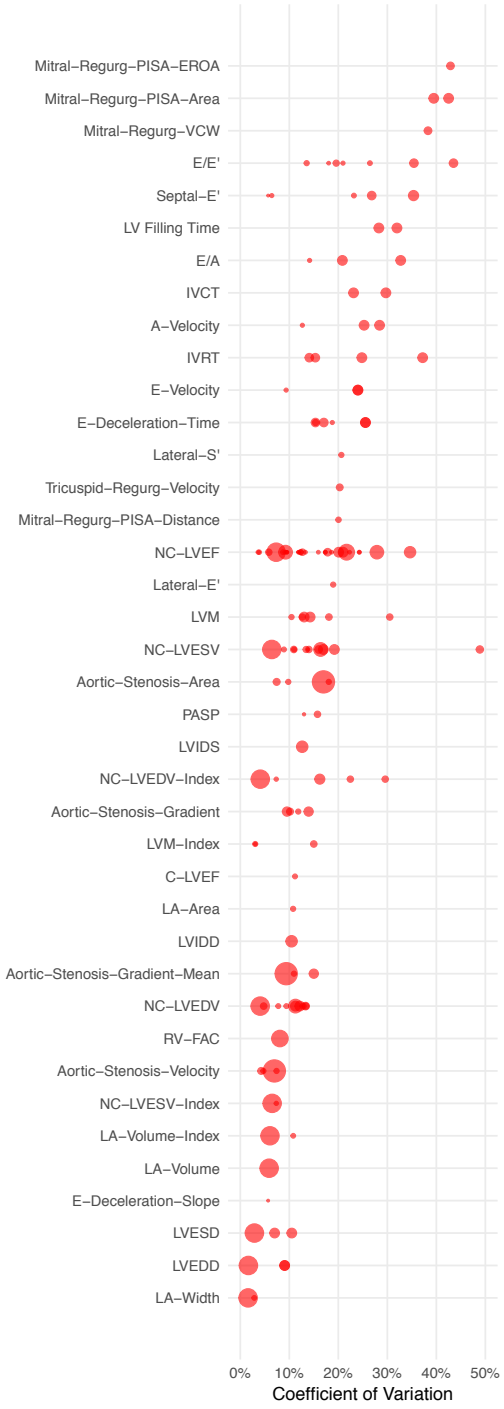


Figure 2-7 Individual test-retest coefficients of variation from the placebo arms of RCTs

Test-retest coefficients of variation, calculated from repeated measurements made from the placebo arm of RCTs. Each dot represents one study. The size of the dot is proportional to the size of the study.

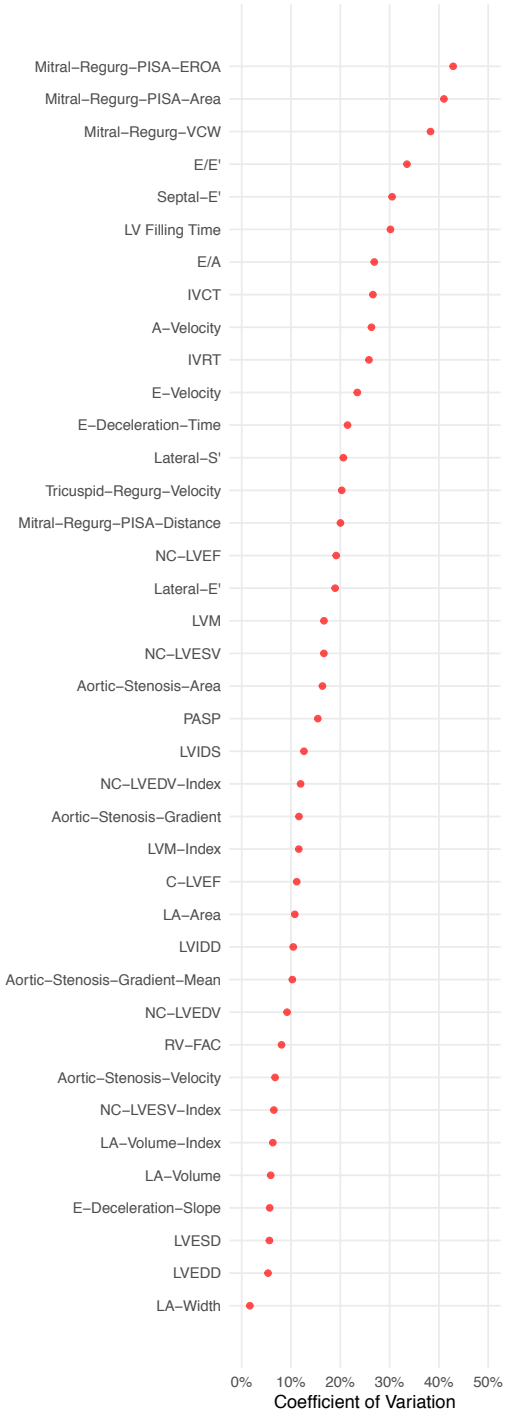


Figure 2-8 Weighted mean test-retest coefficients of variation from the placebo arms of RCTs

Test-retest coefficients of variation, calculated from repeated measurements made from the placebo arm of RCTs. Each dot represents the weighted root mean square of the individual coefficients of variation.

2.4.6 Comparison between intra-reader, inter-reader, and test-retest variability derived from reproducibility studies and randomised controlled trials

18 parameters had test-retest data from both reproducibility study and RCT data. Of these in 16/18 (89%) the reproducibility data provided much larger estimates of the coefficient of variation than the dedicated reproducibility studies (p=0.001, Table 2.1, Figure 2-9, Figure 2-10). Across these 16 parameters, the ratio by which the CV was underestimated ranged from 4.78 for the Septal E' to 1.21 for E deceleration time (Table 2.1).

On average, the reproducibility studies under-estimated the CV by 2.3-fold. As the required sample sizes for clinical trials are proportional to the square of the standard deviation (or CV), then this underestimation would lead to trials requiring 1.47 to 22.82-times more patients than might be expected based on reproducibility studies (assuming no treatment variation). The median time interval between scans was 1 day for the test-retest studies, and 6 months for the RCTs.

There was no significant correlation between the CV obtained from RCTs and reproducibility studies (rho=0.31, p=0.2).

Table 2.1 Test-retest coefficients of variation for echocardiographic biomarkers for which data from both the placebo-arm of RCTs and dedicated reproducibility studies were available.

	Number of included patients		**Test-retest Coefficient of variation**		**Ratio of underestimation of CV from reproducibility study compared to RCT**	**Ratio of underestimation of sample size (assuming no treatment variation)**
	RCT	Repro Study	RCT	Repro Study		
Septal-E'	328	50	31%	6%	4.78	22.82
Lateral-S'	32	30	21%	5%	4.35	18.95
E/E'	363	20	34%	10%	3.44	11.82
LVIDD	219	261	10%	3%	3.14	9.88
E-Velocity	319	138	23%	8%	3.12	9.71
Non-contrast LVEF	2726	228	19%	7%	2.83	8.02
Lateral-E'	32	30	19%	7%	2.68	7.19
LV Mass	501	317	17%	8%	2.18	4.75
Mitral Regurgitation PISA Distance	37	59	20%	9%	2.11	4.46
A-Velocity	319	138	26%	14%	1.95	3.78
IVRT	526	20	26%	15%	1.71	2.93
E/A	319	148	27%	17%	1.60	2.55
Non-contrast LVESV	1658	178	17%	12%	1.41	1.98
LVIDS	219	78	13%	9%	1.37	1.88
Contrast LVEF	28	10	11%	9%	1.30	1.69
E Deceleration Time	594	20	21%	18%	1.21	1.47
NC-LVEDV	1561	178	9%	11%	0.82	0.67
LA-Volume	620	100	6%	8%	0.76	0.58

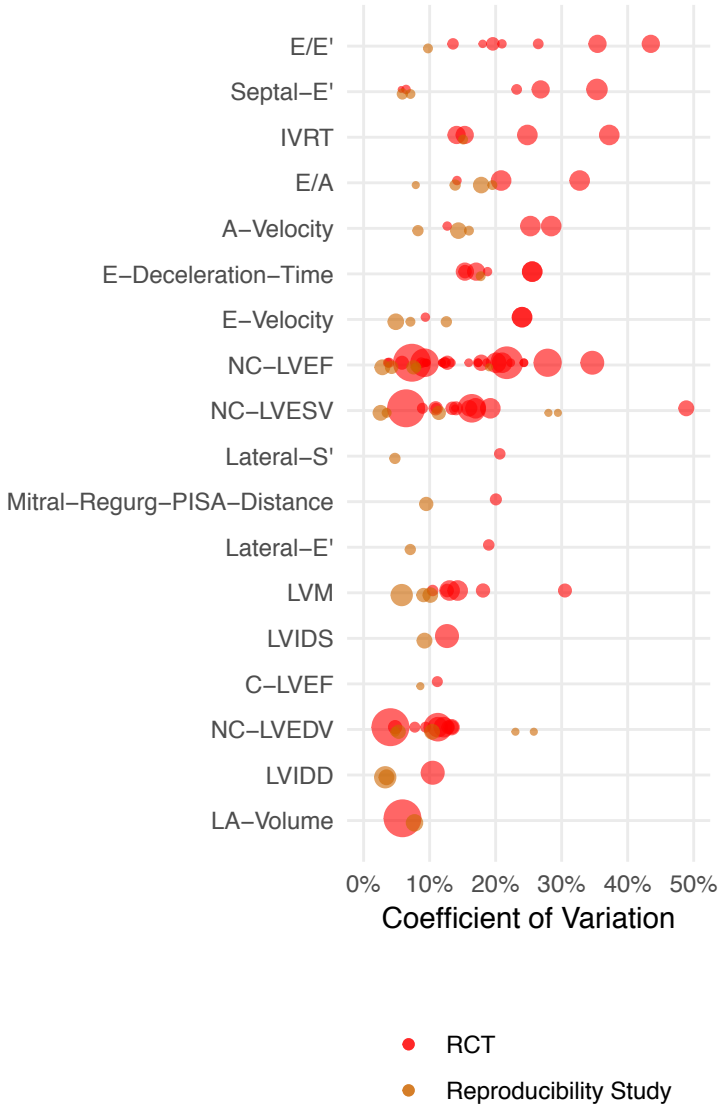


Figure 2-9 Individual test-retest CV of measurements for which data was available for both reproducibility studies and the placebo arms of RCTs

Test-retest coefficients of variation, calculated from separately acquired images from reproducibility studies, and from repeated measurements made from the placebo arm of RCTs. The area of each dot is proportional to the size of each study.

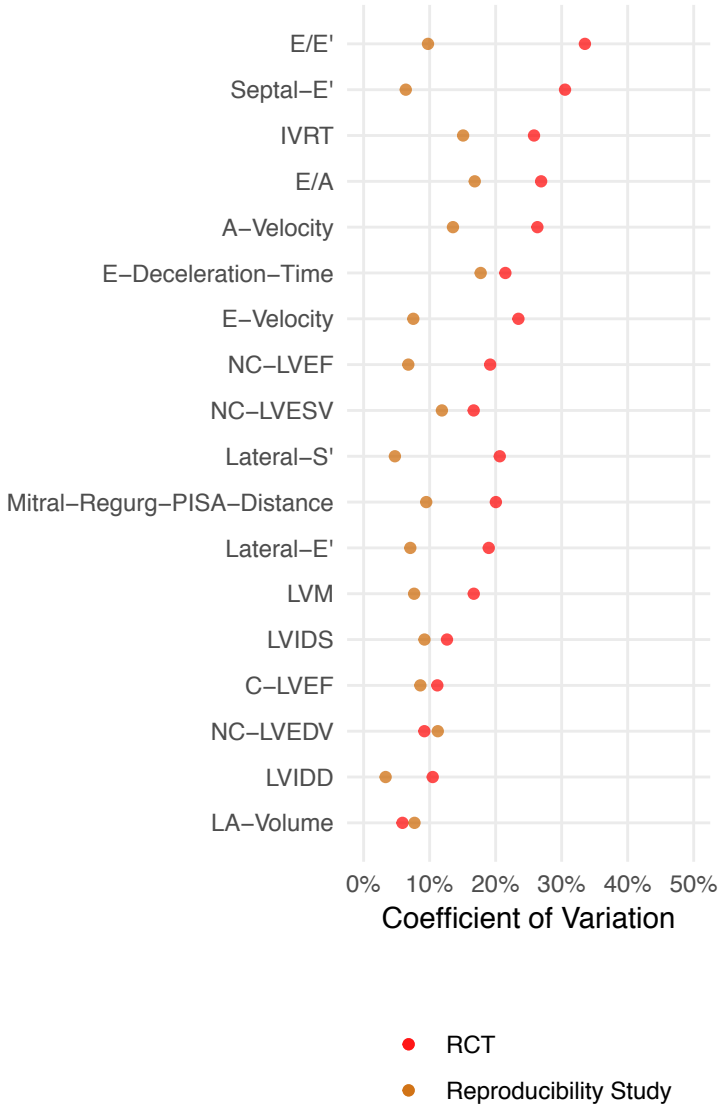


Figure 2-10 Summary test-retest CV of measurements which data was available for both reproducibility studies and the placebo arms of RCTs

Test-retest coefficients of variation, calculated from separately acquired images from reproducibility studies, and from repeated measurements made from the placebo arm of RCTs. Each dot represents the weighted root mean square of the individual coefficients of variation.

2.4.7 Comparison of RCT data to reproducibility study data for specific parameters

Overall, of the 90 echocardiographic parameters for which data were available, reproducibility studies provided data on 75 (83%) of them. Of these 75, the ESC guidelines only referenced studies providing test-retest data on 44 (59%) of them. Data for all parameters is summarised in Figure 2-11.

2.4.7.1 Aortic stenosis

For aortic stenosis parameters, the ESC guidelines referenced no studies that quoted useable test-retest data. Only intra-reader and inter-reader variability were presented in referenced studies. The RCT derived test-retest coefficient of variation ranged from 16% for the aortic valve area to 7% for the aortic peak velocity. Diligent readers using ESC guidelines and associated references, would therefore have to base their estimate on the inter-reader CV, which ranged from 4% for aortic valve area to only 2% for aortic peak velocity. This represents a 3.86 to 4.23-fold under-estimation of the true CV (Table 2.2, Figure 2-12, Figure 2-13).

Table 2.2 Reproducibility of aortic stenosis parameters

Aortic stenosis parameter	Number of included patients			Coefficient of variation			Ratio of CV from RCT to inter-reader data from reproducibility studies
	RCT	Inter-reader	Intra-reader	RCT	Inter-reader	Intra-reader	
Aortic valve area	1062	36	226	16%	4%	8%	3.86
Aortic peak gradient	371	300	340	12%	6%	5%	2.01
Aortic mean gradient	1096	150	300	10%	3%	5%	3.37
Aortic peak velocity	1062	162	383	7%	2%	3%	4.23

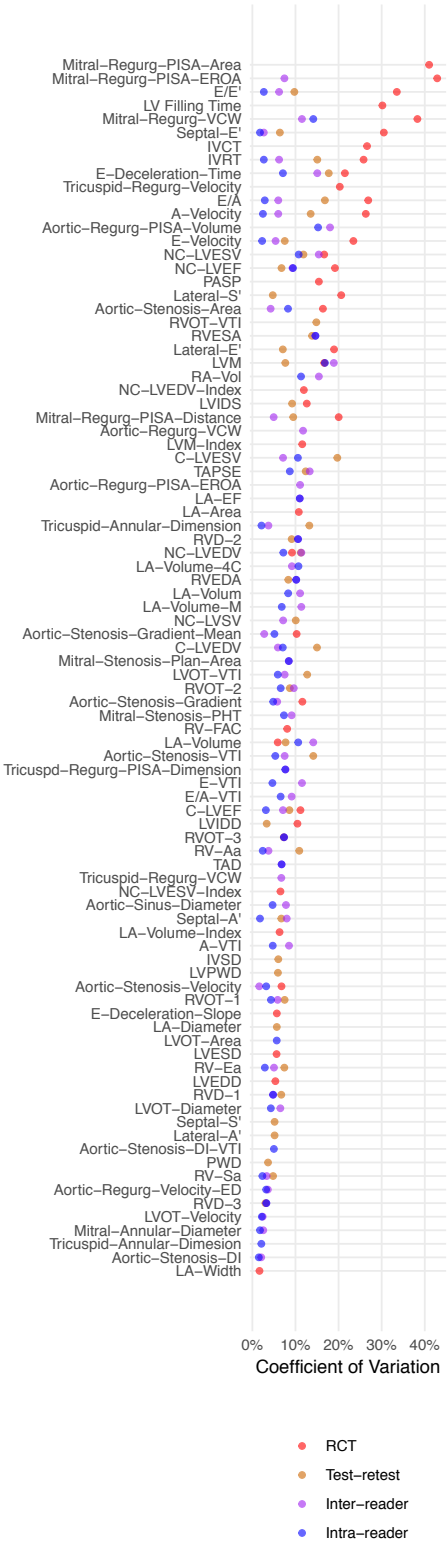


Figure 2-11 Intra-reader, inter-reader, and test-retest coefficients of variation from reproducibility studies and RCTs

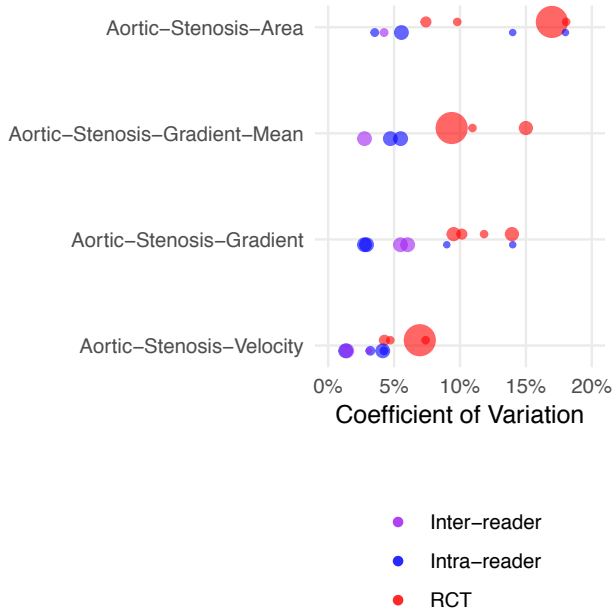


Figure 2-12 Individual CV of aortic stenosis measurements from reproducibility studies and the placebo arms of RCTs

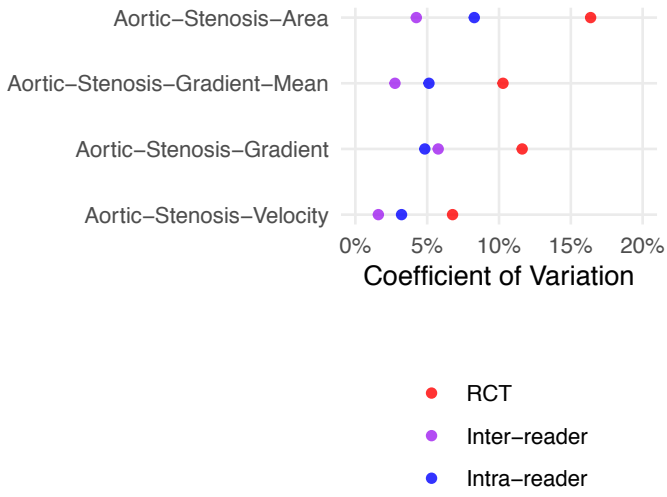


Figure 2-13 Summary CV of aortic stenosis measurements from reproducibility studies and the placebo arms of RCTs

2.4.7.2 Left-ventricular Ejection Fraction

The reproducibility studies referenced in the ESC guidelines provided information on the test-retest, inter-reader and intra-reader variability for both contrast-enhanced and non-contrast biplane LVEF assessments. As compared to the placebo-arm of RCTs, the reproducibility studies underestimated the test-retest CV for both contrast LVEF and non-contrast LVEF (11.2% vs 8.6%, and 19.2% vs 6.8% respectively). This is a 1.3-fold and 2.84-fold underestimation respectively. RCTs would require a sample size 1.7-fold and 8.0-fold larger respectively than might be expected based on reproducibility studies (Table 2.3, Figure 2-14, Figure 2-15).

Table 2.3 Reproducibility of LVEF measurements

	Number of included patients				**Coefficient of variation**				**Ratio of CV from RCT to inter-reader data from reproducibility studies**
	RCT	Test-retest	Inter-reader	Intra-reader	RCT	Test-retest	Inter-reader	Intra-reader	
Contrast LVEF	2726	228	1574	171	11.2%	8.6%	7.1%	3.1%	1.30
Non-contrast LVEF	28	10	40	40	19.2%	6.8%	9.3%	9.4%	2.84

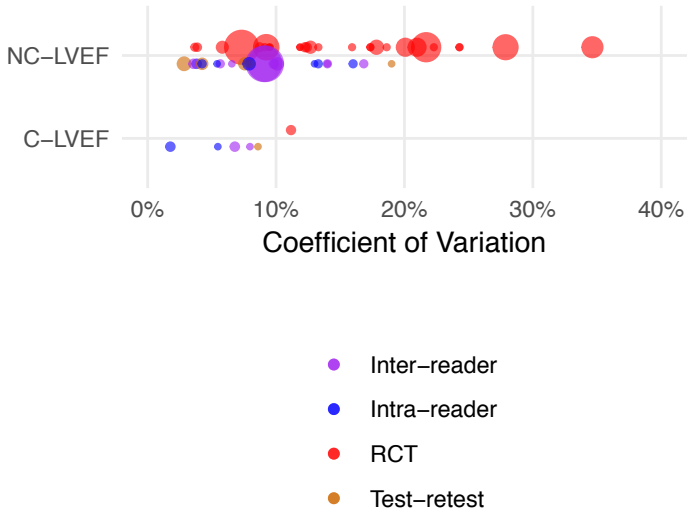


Figure 2-14 Individual coefficients of variation for contract (C-) and non-contrast (NC-) LVEF from reproducibility studies and the placebo arms of RCTs

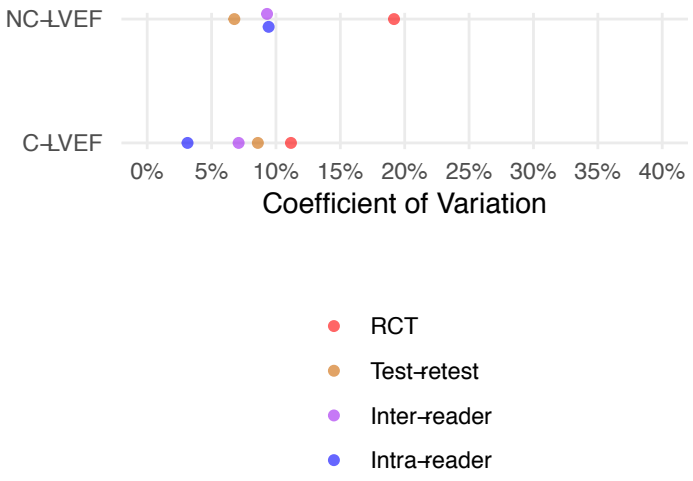


Figure 2-15 Summary coefficients of variation of contrast (C-) and non-contrast (NC-) LVEF form reproducibility studies and the placebo arms of RCTs

2.5 Discussion

In this Chapter I have provided, for the first time, a compendium of the test-retest variability of common echocardiographic biomarkers with data derived from two sources – dedicated reproducibility studies, and the placebo-arm of RCTs. I also have provided data on the less clinically useful (but more often reported) *inter*- and *intra*-reader variability.

I found that the CV of these biomarkers are much larger than may been reported. As compared to the placebo-arm of RCTs, dedicated reproducibility studies underestimated the test-retest coefficient of variation by 2.3-fold. This ranged from 4.8-fold for septal-E' to 1.2-fold for E-wave deceleration time. Sample sizes for RCTs using continuous variables as outcomes are proportional to the square of the standard deviation or coefficient of variation. Therefore, a triallist deriving information from a reproducibility study is liable to underestimate the sample size on average approximately 5.3-fold.

2.5.1 Biomarker features associated with large coefficient of variation

The primary driver of large coefficients of variation was high dimensionality to the biomarker, i.e., assessment of a three-dimensional volume generally has a larger CV than a two-dimensional measure such as a VTI, and in turn both were in general worse than the assessment of a one-dimensional marker such as a dimension.

More complex echocardiographic markers have larger coefficients of variation. Perhaps the most complex measurement in that is in common use is the EROA

calculated using PISA. Within clinical trials, this has a CV in excess of 40%. Of intermediate complexity, the aortic valve area, has a CV of 16%. Finally, a simple measure such as aortic peak velocity has a CV of 7%.

Based on these data within this Chapter, triallists would be recommended to carefully consider if the complexity of "more accurate" markers (such as aortic valve area) are worth requiring a sample size $(16/7)^2$ = 5-fold it over the simpler measures (such as aortic peak velocity).

Some of the biomarkers with a higher coefficient of variation may be less in clinical use that those with a lower coefficient of variation. This would minimise the practical implications of reported biomarkers with a large coefficient of variation. However, non-contrast LVEF and E/E' have a coefficient of variation in the RCTs of 19% and 34% respectively, in contrast with the claimed 10% and 7% in reproducibility studies.

2.5.1.1 Signal-to-noise ratio

Whilst some echocardiographic biomarkers have much greater coefficient of variation than others (e.g. the 7% of aortic valve peak velocity vs the 40% of mitral regurgitation EROA) and hence greater noise, this must be put into the context of the expected signal. Within clinical diagnosis an appropriate comparison is that of the spread of values in the population. For example, for peak aortic velocity, values typically fall between 1 m/s and 4m/s – a 4-fold range. However, for mitral regurgitation EROA, values may fall from 5 mm^2 to 40 mm^2 an 8-fold range, and 2-

fold larger than that of peak aortic velocity. As I will discuss in Chapter 5 this can be formally contextualised by calculating the intraclass correlation coefficient (ICC).

For clinical surveillance, or in a clinical trial, the comparative signal is much smaller – that of the expected change in the value – and therefore the reproducibility that much more important.

2.5.2 Limitations

2.5.2.1 Unknown source of the discordance between RCTs and reproducibility studies

This study was not designed to ascertain the underlying cause for the discordance between RCTs and reproducibility studies. However, there were some differences between the studies.

2.5.2.1.1 RCT measurements were conducted over a more clinically relevant time span

One possible cause of the discordant results is that the time interval between test and retest was larger for RCTs than in the dedicated reproducibility studies (median interval 6 months vs 1 day respectively). Over greater periods of time, biological variability will accumulate. Studies that enrolled patients that were likely to result in significant disease progress or regression (such as post-MI or with an active control) were excluded. As clinicians will typically reassess patients at an interval closer to 6 months than 1 day, the data from the RCTs is more likely to be useful for clinical practice (and of course for future RCTs).

2.5.2.1.2 RCTs may not be performing echocardiography to "gold-standard' levels

Other differences between dedicated reproducibility studies and RCTs may exist. It could be argued that the dedicated reproducibility studies undertook special measures that were not performed in the randomised trials to ensure reproducibility. However, as this discordance was consistent over many trials and over 88% of biomarkers, this would represent a significant misunderstanding of the practicalities of using echocardiography by the global community of triallists.

Only some trials used echocardiographic core-labs where through special measures, better reproducibility may be achieved. However, as I will find in Chapter 3, the majority of measurement variability in echocardiography arises at the acquisition stage. Therefore, even with improved interpretative processes, there is a natural limit to the reproducibility that can be achieved.

2.5.2.1.3 RCTs have bias resistant features that reproducibility studies may lack

RCTs both undertake and document their significant effort expended in ensuring and maintaining blinding during trials. Curiously, these bias-resistant features are not typically documented for reproducibility studies.

Overly optimistic assessments of reproducibility may arise in dedicated reproducibility studies through the introduction of bias that would not arise within well-conducted RCTs. This bias could arise from the removal of subject data either:

- Before acquisition. This would occur, for example, if a patient was not recruited as it was noted that they had a high BMI and may therefore be expected to have poor echocardiographic windows.

- After acquisition. This would occur, for example, if data from an individual was excluded after seeing it in the context of other individual's data.

The former is more acceptable than the latter, but both would lead to optimistic calculations of the reproducibility of biomarkers.

2.5.2.2 Reproducibility studies sourced from guidelines

In this study I limited the dedicated reproducibility study search to studies referenced by recent ESC echocardiographic guidelines. The studies included within the guidelines may have been selected for their inclusion for their low variability. However, an interested clinician or researcher would be likely to first go to established guidelines such as these and base their opinion on studies referenced within.

2.5.2.3 Biases introduced by data abstraction

In this study, I expressed the within patient standard deviation as a coefficient of variation so that the relative variability between different markers could be compared. Calculating this by dividing the overall standard deviation by the overall mean whilst not mathematically correct, is a reasonable approximation (Bland 2006). More strictly, the coefficient of variation should be used if the error is proportional

to the mean, whilst the standard deviation if it is not. Indeed, the purpose of Bland-Altman plots is to show the relationship between the errors and the mean, rather than a graphical demonstration that around 95% of the data-points will lie between the 95% limits of agreement. Unfortunately, within most reproducibility studies, and all RCTs included, the relationship between the differences and the mean was not described.

On average, these approximations would not be expected to bias the results consistently in one direction to a substantial degree.

2.5.2.4 Lack of 3d- or strain-based biomarkers

In this study I did not assess 3d-, or strain-based biomarkers as they are not yet considered ubiquitous. They (like many new biomarkers) have been reported as providing increased accuracy and precision (Medvedofsky et al. 2017). However, if the additional variability comes from true intrinsic biological variability rather than analytical variability, then these more advanced markers will be just as susceptible.

2.5.3 Clinical implications

Clinicians should be aware that the real-world test-retest variability of common echocardiographic markers over a typical clinical time period is larger than might be presumed. This irreproducibility may cause clinician and patient uncertainty and result in the repetition or re-analysis of studies. As patients approach criteria for intervention, great anxiety may be generated as their results will oscillate either side of the decision threshold.

2.5.4 Implications for research

Clinical triallists, when considering using an imaging biomarker as a trial end-points should:

(1) Derive sample size calculations from data obtained from previous RCTs rather than reproducibility studies. If a novel biomarker for which no previous RCT data is available is selected as an end-point, it may be prudent to compensate for the average underestimation by multiplying the calculated sample size by 5.3. Their sample size calculation will also need to consider treatment heterogeneity.

(2) For studies, such as pharmacological therapies to slow the progression of aortic stenosis, triallists can use the data presented within to select a biomarker with the lowest test-retest reproducibility. For example, the peak aortic velocity has a test-retest variability over 2-fold less that of the aortic valve area, resulting in a 5-fold smaller required sample size.

Research into understanding the sources of test-retest variability of common echocardiographic markers and how to minimise them is to resolve these discrepancies.

2.6 Conclusion

In this Chapter I have provided a summary of the expected test-retest variability over clinically relevant time-periods (median 6 months) of standard 2D

echocardiographic parameters. Furthermore, I found that dedicated reproducibility studies provide gross underestimates of the true variability as compared to the placebo-arm of blinded RCTs.

Measurement precision is desirable for both clinicians and researchers. True improvements can only come when the sources of irreproducibility are elucidated and targeted. I attack both these problems in the following Chapters.

3 The causes of irreproducibility in Doppler measurements of aortic stenosis

3.1 Abstract

3.1.1 Introduction

Quantification of aortic stenosis suffers from surprisingly poor test-retest reproducibility when done under blinded conditions. This is sometimes attributed to inadequate training, or operator variability, but whether this is actually true has never been properly tested.

If two sonographers do separate scans and come to different conclusions, this is not necessarily due to differences between the operators. In this chapter I explain what I mean by this.

I then develop a method that can be applied to decompose the sources of variability in any clinical measurement. The method is pragmatic, in that it does not need a rigid grid of complete data. It provides a graphical display of the sources of variability and recognises that some apparent sources actually contain other sources that can be separated out.

I use this method to separate the sources of variability in aortic stenosis variability so that I can identify which aspect are worth targeting with improvements.

Hypothesis: I will be able to break-down the sources of variability that contribute to the doppler assessment of aortic stenosis so that they can be specifically targeted in subsequent Chapters.

3.1.2 Methods

Traditional approaches to separating the contributions to variability in clinical measurements require a complete grid of N patients by M sonographers, perhaps multiplied in another dimension by P readers. Putting together a rigid dataset of $N \times M \times P$ observations is impracticably complex. Therefore, most studies of reproducibility set M and P to very small values, such as 1 or 2.

Furthermore, it is not traditional to attempt to extract all the quantitative contributions, because this would make the analysis unimaginably complex.

In this Chapter I adapt a method to solve both of these problems. I take a method (a random-effects model), commonly used in other measurement sciences, but not previously applied to cardiac imaging, and apply it to assess the individual sources of variability in aortic stenosis quantification.

In total, 117 different patients with aortic stenosis were scanned by multiple operators (1 to 4). Each operator made multiple replicate acquisitions. These were then analysed in some cases by multiple observers.

I adapted a random-effects model to enable separation of the sources of variability. I developed a display method that could shows in an intuitive format, exposing both the Pythagorean behaviours of standard deviations and the hierarchical nature the contributions to variability.

3.1.3 Results

For continuous wave doppler, my findings were as follows:

Variability observed between tracings was 3.3%

Variability *observed* between beats was 6.0%. However, after accounting for the 3.3% automatically arising from the act of tracing, the amount *genuinely* caused by the change in beat was 4.9%, which is approximately $\sqrt{6.0^2 - 3.3^2}$.

Variability *observed* between placements was 8.9%. However, after accounting for the 4.9% arising from change in beat and the 3.3% arising from the act of tracing, the amount *genuinely* caused by the subtle changes in probe placement was 6.5%, which is approximately $\sqrt{8.0^2 - 4.9^2 - 3.3^2}$.

Variability observed between operator was 11%. However, after accounting for the 4.9% arising from change in beat, the 3.3% arising from the act of tracing, and the 6.5% arising from probe placement, the amount *genuinely* caused by the different operators was $\sqrt{11\%^2 - 6.5^2 - 4.9^2 - 3.3^2}$

I continued this process for change of operator, change of reader, and change of visit. I also repeated this entire process for the pulsed wave data.

The most remarkable finding was that the great majority of the variability arose at probe placement level itself (11.3% out of the total 14.1% for pulsed-wave) but not for continuous wave (6.5% vs 13.0%).

3.1.4 Conclusions

I was able to adapt for use in cardiac imaging, a random-effects model that can be applied to clinically realistic datasets, without requiring a complete grid of every patient assessed by every observer on a fixed number of occasions.

I was also successful in describing the components thus extracted in a "peel-away" variance model so that echocardiographs can truly understand the source of variability, and not assume it is an inherent disagreement between observers.

With my approach, I was able to find that the majority of the $(11.0\%)^2$ variance in measurement of continuous wave velocities *observed* between two operators occurs not because of the change of operator whose contribution is only $(6.1\%)^2$, but because in variability arising at the levels of probe placement, change of beat, and episode of tracing, whose contributions total $(8.9\%)^2$.

For pulsed wave doppler, the distinction was even more startling. Of the $(14.1\%)^2$ *observed* variation between operators, virtually all of it was explained by the $(13.7\%)^2$ arising at the levels of probe placement, change of beat, and episode of tracing.

In this chapter discuss the implications of these findings for how we can improve on quantification of the aortic valve Doppler velocities. I develop these themes throughout this thesis.

3.2 Introduction

As previously discussed in Section 2.4, whilst some studies report excellent reproducibility of echocardiographic Doppler derived measurements, within clinical trials the realised variability can be much higher (e.g. the test-retest coefficient of variation for aortic valve area is 16%). Furthermore, our group has published that within clinical practice the test-retest variation to be very much wider than commonly realised, with a coefficient of variation (CV) for the dimensionless index of 12% at best (Finegold et al. 2013).

Before developing an automated method to improve the reproducibility of aortic stenosis assessment, an understanding of the causes of the irreproducibility of aortic stenosis measurements.

3.2.1 What are the causes of irreproducibility in the Doppler assessment of aortic stenosis?

Guidelines do give specific instructions to improve reproducibility of measurement of aortic stenosis. Recommendations for follow-up scans of aortic stenosis are to use the same operator and same machine, which is difficult to achieve. Strangely, the advice is also to unblind operators to prior values, which would not be expected in a reliable measurement methodology (Douglas et al. 2009; Evangelista et al. 2008).

"Operator dependent" is a frequent criticism of echocardiographic valve quantification, but there is no good evidence that the variability is caused by change in operator rather than spontaneous variation.

If measured independently on separate visits, indices of aortic stenosis severity can show fluctuations that do not necessarily represent genuine biological changes. Given the magnitude of the decision being made (procedural intervention versus continuing surveillance), we should try to quantify the different sources of these fluctuations so that efforts to reduce the fluctuations can be correctly designed to target the sources that are contributing the most.

While there are studies of sources of variability, these typically focus on only one or two elements (Chapter 2). In this study, we attempt to determine all the elements simultaneously.

Second, studies rarely address the fact that measured variances are in-fact nested. For example, if two operators disagree, we sometimes incorrectly assume that this is because of differences between operators. In reality, we should remember each operator may disagree with himself or herself by as much as they disagree with another operator, and if this is the case, we should not incorrectly state that there are differences between the operators.

3.2.2 Aims of this study

In this study, I will present a method that quantitatively extracts each of the many nested layers of sources of variance.

I designed it to extract the *true* variability for each the following components:

- **Tracing:** Variation when same reader looks at same stored beat on second occasion, blinded. An alternative term is ***intra***-reader.

- **Beat:** Conceptually, variation between one beat and another, everything else kept the same. Practical measurements unavoidably include tracing variability which must therefore be subtracted.

- **Placement:** Conceptually, variation when probe is removed and put back on chest, all else kept same. (Practical measurements unavoidably include tracing and beat variability which must be subtracted). An alternative term is ***intra***-operator variability

- **Operator**: Conceptually, variation when operator is changed. (Practical measurements unavoidably include tracing, beat, and placement variabilities which must be subtracted). Also known as ***inter***-operator variability.

- **Reader:** Conceptually, variation when different reader looks at same stored beat. (Practical measurements unavoidably include tracing variability which must therefore be subtracted). An alternative term is ***inter-***reader variability.

- **Visit:** Conceptually, variation when a patient returns on a different day (but not so long that significant progression of disease may have occurred). An alternative term is visit-to-visit variability.

It was also designed to be compatible with data that could be obtained with minimal disturbance to clinical practice by:

- Not requiring a fixed set of operators to assess each patient, and,
- allowing for a different amount of data to be collected from each patient.

I will then use apply this method to understand the sources of variation in the assessment of aortic stenosis.

This Chapter combines methods adapted from other measurement sciences and for the first time applies it to a cardiac imaging biomarker.

Hypothesis: I will be able to break-down the sources of variability that contribute to the doppler assessment of aortic stenosis so that they can be specifically targeted in subsequent Chapters.

3.3 Methods

3.3.1 Distinguishing the sources of variability even though they can only be measured nested

There are more sources of variability than commonly distinguished. For example, if two echocardiographers independently assess the same patient they will almost always obtain numerically different values. This should not be attributed to differences between echocardiographers because each echocardiographer, if they reassessed the same patient under blinded conditions, would themselves almost always obtain numerically different results. Thus, the true incremental contribution of changing the operator, and only be assessed by subtracting the variance arising from the same operator assessing the patient twice.

The variability within the same operator can itself be broken down into components.

For example, even though the operator is skilled and intending to position the probe correctly on both occasions, the two placements will inevitably be subtly different. Nevertheless, it is wrong to automatically attribute any differences in the measurements between the two occasions on probe placement, because every beat is itself different. Since two probe placements will always be addressing different sets of beats, the beat variability must be subtracted because it inevitably contributes to the apparent impact of subtle-differences in probe placement (and between operators).

But there is still more. Every time an acquired beat is read, even by the same reader, a subtly different value is obtained. This within-reader variability in reading must be subtracted because it inevitably contributes to the apparent differences between heart beats (and those due to probe placement, and those due to operator).

Furthermore, even examining the same acquired beat, different readers may obtain different values. This should not be attributed to differences between readers unless we first subtract the variability from the same reader analysing the same beat twice independently.

Finally, there are biological variability on a visit-to-visit basis (such as changes in fluid balance). Different values may be obtained on different days. This should not be attributed differences between different visits unless we first subtract all the underlying variability first.

3.3.2 Quantifying the sources of variability

In principle the independent contributions to variance could be calculated easily if each patient underwent independent scans by each operator, on each occasion, placing the probe multiple times and with each placement recording multiple beats and all beats subsequently analysed multiple times by multiple readers. However, if there were just 5 patients, 5 operators, 5 placings, 5 beats per placing, 5 readers, and 5 tracings per reader, with patients attending for 2 separate visits, this would necessitate $2 \times 5^6 = 31{,}250$ tracings in total. The enormity of this might have

discouraged such an endeavour previously. Even if this amount of effort could be invested it might be difficult.

To make this analysis feasible two things are needed:

- **Targeted data collection**. Some patients will have multiple visits, but single operators, other patients will come for one visit, but be seen by multiple operators.

- **Contemporary statistical techniques**. Random-effects models allow for realistically achievable designs such as this (which may be termed *unbalanced* in the statistical literature).

3.3.3 Patients

Patients were recruited from patients with mild, moderate, or severe aortic stenosis of native or tissue valves from "Valve Clinics" at my institution.

It was essential to capture all the different sources of variability. There are many sources of variability, and each has to be captured in order to quantify it. The reason no previous study has fully separated the sources is that it is too difficult to complete a full matrix of all sources of variability on a single patient cohort, because this would involve all patients, attending for several very long days of study.

In this chapter I apply contemporary statistical techniques to permit separate cohorts to be used to extract overlapping sources of variability, and then the results are

combined in a statistically valid manner to obtain all the sources. Figure 3-1 shows how the two cohorts covered all the sources of variability.

Two cohorts of patients were recruited.

- **Cohort A:** An initial cohort of 27 patients who were scanned by multiple BSE accredited operators (2 to 4), from a pool of 17. This data was collected by a colleague and formed part of a departmental quality improvement project.
- **Cohort B:** A further cohort of 90 patients who were scanned by a single BSE accredited operator (myself), of whom 25 were brought back for a second visit within 12 months.

The statistical methods allowed both sets of data to be combined to simultaneously estimated the constituent variances.

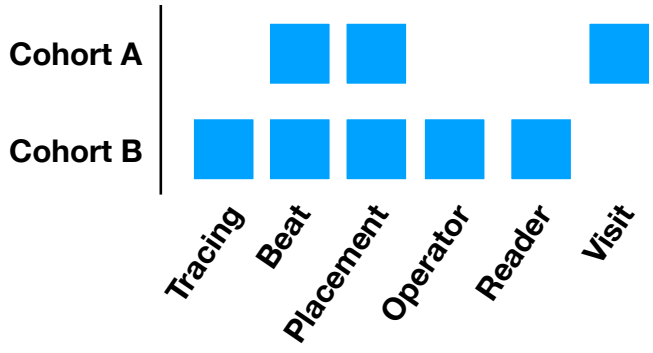


Figure 3-1 Two cohorts of patients were used to provide data for every source of variability. Cohort A contributed data at the beat, probe placement, and visit level of variability. Cohort B contributed data at the tracing, beat, placement, operator, and reader level.

3.3.4 Data collection

3.3.4.1 Cohort A

Patients attending for a routine echocardiographic appointment for surveillance of their aortic stenosis had additional images collected as part of a departmental quality improvement project. Data from this project is included in this study.

In addition to the standard echocardiographic images, the lead echocardiographer acquired a run of multiple-beats (up to 10 seconds worth) of continuous wave through the aortic valve. They then acquired a multiple-beat run of pulsed wave in the LVOT. The probe was then removed from then re-placed on the chest and the process repeated at least two further times.

After the lead echocardiographer had acquired at least 3 multiple-beat runs of both continuous wave and pulsed wave data another echocardiographer repeated the process on the same patient. Up to 3 additional (4 in total) echocardiographers acquired data in this manner.

3.3.4.2 Cohort B

This data was collected as part of the validation data required for Chapter 4. Patients attending the echocardiography department for a routine echocardiographic appointment for surveillance of their aortic stenosis were recruited (REC reference 14/LO/1520).

As with patients in Cohort A, I acquired a run of multiple-beats (up to 10 seconds worth) of continuous wave through the aortic valve. I then acquired a multiple-beat

run of pulsed wave in the LVOT. The probe was then removed from then re-placed on the chest and the process repeated at least two further times.

Selected patients were re-scanned at an interval of 6- to 12-months.

3.3.4.3 Technical considerations for acquiring multiple-beat runs

Each patient required multiple continuous wave Doppler strips, and multiple pulsed wave Doppler strips through the aortic valve and LVOT in the 5-chamber view.

The continuous wave and pulsed wave data were acquired in a manner consistent with BSE guidelines (Wharton et al. 2015). The novel techniques developed in the following Chapter (Chapter 4) were not used in this study.

Each strip contained many beats: many more beats than would be captured for a standard clinical echocardiogram. Simply capturing a freeze-frame image would only acquire approximately 3 beats. If the sweep speed were reduced it would be possible to capture many more beats, but the image resolution of each beat's Doppler signal would be corresponding reduced which is undesirable.

I therefore adapted the techniques originally developed in the Chapter 4 during the creation of the software, for use in routine clinical measurements. I made available three methods of capturing the Doppler spectrum:

(1) Acquiring 10 beats or 10 seconds worth of Doppler spectrum without "freezing" the image. The video is "unravelled" into a long strip using software I developed.

(2) Using a frame-grabber plugged into the video output port of the echocardiographic machine.

(3) Using the freeze-frame DICOM files.

For the initial cohort of patients, I had not developed these techniques therefore long runs of data were acquired by carefully piecing together sequentially acquired "frozen" images.

For each patient, at least three continuous wave and pulsed wave acquisitions from separate probe placements were obtained for analysis.

3.3.5 Measurement of the echocardiographic data

3.3.5.1 Cohort A

Images were exported as anonymised, deidentified TIFFs. For each beat an image was generated that removed all other beats, leaving only it and the calibration markings. The operator, position, beat number, along with calibration data were entered into a database. Images were then screened for quality. No whole patient, operator, or position were excluded from the database, but 16/1880 images (<1%) were considered to be of such poor quality as to prevent further analysis. Images were analysed using ImageJ to select the maximum velocity on a 15-inch MacBook Pro (Apple Inc., California, USA) in a darkened room using the imaging software Fiji, Version 1.0 (Schindelin et al. 2012). Sonographers were given the choice of using a mouse or trackball.

One physician selected the peak velocity in every image. Multiple sonographers remeasured the peak velocities in a subset of 123 beats, randomly chosen by automatic algorithm with an even distribution across patients and operators.

Therefore, for each patient, every beat, for every probe placement, for every operator was traced by a single reader. This reader along with multiple other readers traced a subset of these beats (123 beats) in duplicate.

3.3.5.2 Cohort B

This much larger cohort of 90 patients contributed to the estimation of the visit, beat, and placement variability. For this cohort, all tracing was done automatically using the software developed in Chapter 4. This prevented these data from contributing to the estimation of tracing or reader variability. This was done for two reasons:

1. The vast amount of data precluded manual assessment
2. By doing this, I targeted providing data on the visit, beat, and placement variability.

For every patients the automated tracing was used to calculate the peak velocity on the continuous wave and pulsed wave tracing for every beat in a run (e.g. 10 beats) , of every probe placement (at least 3), for every visit (1 or 2).

My analytical method was designed to allow datasets form different cohorts to be combined and analysed validly together.

3.3.6 Data analysis

All the measurements were combined in a database, with each measurement tagged with the patient code, operator, reader, visit number, probe placement number, and beat number.

In statistical terminology a distinction is drawn between the type of variation where the one cares about the identity of the individual instances (fixed-effects) and the type of variation where one does not care about the identity of the individual instances, only their overall distribution (random-effects). For example:

- In a comparison weights between men and women, gender is a fixed effect because once the analysis is complete, we care about whether the men or women were the heavier group.

- In a comparison of replicate measurements of weights, in a group of 20 men, each weighted on 10 different days, we may be trying to establish how much weights vary over time, and whether they vary similarly in different men. The purpose of this study to make a general statement about variability, and not about any particular day or any particular man. Therefore, both "day" and "man" are random-effects. This means we are only interested in the distribution formed by different days, and not about any day in particular. Similarly, we are interested in distribution formed by different men, and not about any man in particular.

- Mixed-models allow us to address both fixed and random effects in the same model. In the case of aortic stenosis measurements, we do want in the end, to be able to in the end identify the patient "John Smith" as having mild aortic stenosis and patient "Jane Bloggs" as having severe aortic stenosis. However, we do not care whether it is the third or seventh beat that is the highest velocity: we only care about the distribution of velocities.

In my analysis, I treated the patient as a fixed effect, as I wanted to be able to later describe the severity in individual cases. All other factors I considered to be random-effects.

I represented the hierarchical nature of the variabilities by nested random effects. For example, suppose CW strip A contained 10 beats, CW strip B contained 12 beats, and CW strip C contained 8 beats:

- The choice of strip (A, B, C) is a random-effect as I am only interested in how much the strips vary from each other, not on their individual values, as there is no way I will ever be able to recreate the individual strips.

- The beat number (1, 2, 3 etc) is a random-effect as I am only interested in how much the beats vary from each other, not on their individual values, as there is no way I will ever be able to recreate that beat.

- The various beats of strip A belong with each other, those of strip B belong with each other, etc.

- However, the various “beat 1”s from the three strips (A, B, C) do not belong with each other. They have no reason to be more like each other than any other beat in that patient.

There is therefore a hierarchy as follows:

- Each patient has several CW strips.
- Each CW strip has several beats.

I represented this structure in the model as a nested random effect. The overall structure I used was as follows:

- There were many patients (fixed effect). Each patient had…
- One or more ***visits*** (random effect). The Doppler data of each visit was viewed by…
- One or more ***readers*** (random effect). Each reader was reviewing data acquired by…
- One or more ***operators*** (random effect). Each operator always acquired…
- More than one strip. Each strip was a separate probe ***placement*** (random effect) on the chest. From each placement (strip) there were…
- More than one ***beat*** (random effect). Each beat was read…
- One or more times which we called ***tracings*** (random effect).

A mixed effect model with patients as a fixed effect, and the arbitrary beat-number nested under the arbitrary placement-number, which itself was nested under operator and visit. Estimates of the variance components with associated confidence (or credible) intervals of these components were generated using MCMC techniques by the MCMCglmm (Hadfield 2010) package in R (using default priors, 100000 iterations, burn-in of 15000 iterations, thinning factor of 50). The measurements were

log-transformed before processing to allow the variance components to be easily converted to coefficients of variation (Bland 2006).

These estimates were plotted individually as bar charts, but also as a series of nested squares (as it is the variances that are additive) in a "peel-away" model I developed.

3.4 Results

3.4.1 Baseline characteristics

Overall, across Cohort A and Cohort B, 117 patients were recruited with aortic stenosis of a native or tissue aortic valve. With the multiple patients, readers, operators, probe placements, and beats, the dataset consisted of 5,654 continuous wave observations, and 4,419 pulsed wave observations.

Within Cohort A, 27 patients with native valve aortic stenosis with and a CW mean peak velocity of 2.95 m/s (range 1.43 to 4.76 m/s) and PW mean peak velocity of 0.93 m/s (range 0.61 to 1.43) were scanned. Each patient was scanned by a median of 3 operators (range 2 to 4). On average 3.1 (range 1 to 7) and 2.8 (range 1 to 7) acquisitions from separate probe placements were performed in each patient respectively, with a mean of 14.8 (range 1 to 51) beats in each acquisition.

Within Cohort B, 90 patients with aortic stenosis in a native or tissue valve aortic stenosis were recruited. Each patient was scanned by a single operator (myself). For there was a mean of 5.9 (range 1 to 15) and 4.8 (range 1 to 15) acquisitions from different probe placements for continuous wave and pulsed wave respectively. The mean velocities were 3.02 m/s (range 1.70 to 4.72) and 1.18 m/s (range 0.65 to 2.13) respectively.

3.4.2 Display of variability components in clinically appreciable manner

To help clinicians and researchers understand the meaning of the results I have presented them graphically in a way that lets makes clearer how the components can

be quantified by effectively subtracting one observed variance from the observed variance within it is nested.

In clinical practice we often talk about variation in terms of standard deviation because this conveniently has the same physical unit as the measurement itself. We use this advantageously to calculate a coefficient of variation which is the ratio of standard deviation to a typical value, i.e. the mean.

The way in which two independent sources of variability combine however, is that the *square* of the resulting standard deviation is equal to the *square* of the two contributing standard deviations.

3.4.2.1 Example: Two columns of random numbers added to form a third.

Table 3.1 shows an example for illustrative purposes. In each column, the numbers are drawn from a distribution that has mean 100 and a standard deviation of 20 for column A and 10 for column B. The right-hand column is calculated as the sum of A + B. The reader can see that the standard deviation of the third column is not the sum of the standard deviations of the first two columns. However, the variance of the third column does approximately corresponds to the sum of the first two columns.

Table 3.1 For the sum of two variables it is the variances rather than standard deviations that summate

	A	B	A+B
	122	102	224
	130	88	218
	75	94	168
	81	97	177
	145	100	246
	79	124	203
	88	79	167
	124	110	234
	109	100	210
	85	75	159
	102	110	212
	139	91	230
	112	84	196
	103	133	236
	79	106	186
Standard Deviation	23	16	28
Variance	550	254	769

For this reason, when there are two sources of error adding together, and each is contributing an SD (SD_A and SD_B), the resulting SD of the overall measure is not SD_A + SD_B, but instead $\sqrt{{SD_A}^2 + {SD_B}^2}$.

Likewise, when a small source of error SD_B is removed from a large error SD_A , the resulting is not SD_A - SD_B, but $\sqrt{{SD_A}^2 - {SD_B}^2}$.

One option would have been to represent the variances directly, but clinicians find these difficult to grasp, as their units are the square of those of the underlying measure. For example, the variance of a velocity is measured in m^2s^{-2}, which does

not have an immediate physical representation. Moreover, the variance equivalent of a coefficient of variation is would have the units of %% i.e. per 10,000. This is not intuitive to most readers.

I therefore developed a way of showing the conventional coefficient of variation (percent) and simultaneously show a value practically equivalent to the variance in a manner that makes the nesting and subtraction process relatively obvious.

My method is to the coefficient of variation on the x-axis, but to draw each value as a square, whose area is an index of its variance. This has several favourable properties.

1. The clinician can instantly perceive the size of the variability from the familiar value of the coefficient of variation on the x-axis.

2. The clinician readily recognises the relative magnitude of two sources of variability since from the areas of the respective squares. For example, below I show one source of variability with standard deviation of 5 units and another with standard deviation of 15 units. Examining only the numbers the clinician might guess the latter was three times more important than the former. But on viewing the picture, it is immediately obvious that the right-hand source is an order of magnitude important (the ratio of variances is $\left(\frac{15}{5}\right)^2 = 9$, Figure 3-2).



Figure 3-2 Relative sizes of two components contributing to reproducibility

3. In the example above, looking at just the SDs a clinician might guess that the combination of the two SDs would give an SD of 20 units. However, if the clinician looks at the picture and realises that it is the two areas that combine, it is apparent that the combination comes to an SD of less than 20 units. This is shown below in Figure 3-3. In fact, the addition of the two CV generates a CV of only $\sqrt[2]{5^2 + 15^2} = 15.8\,units$. In other words:

"Once you have a big noise component, adding smaller noise components makes the outcome only a little worse".

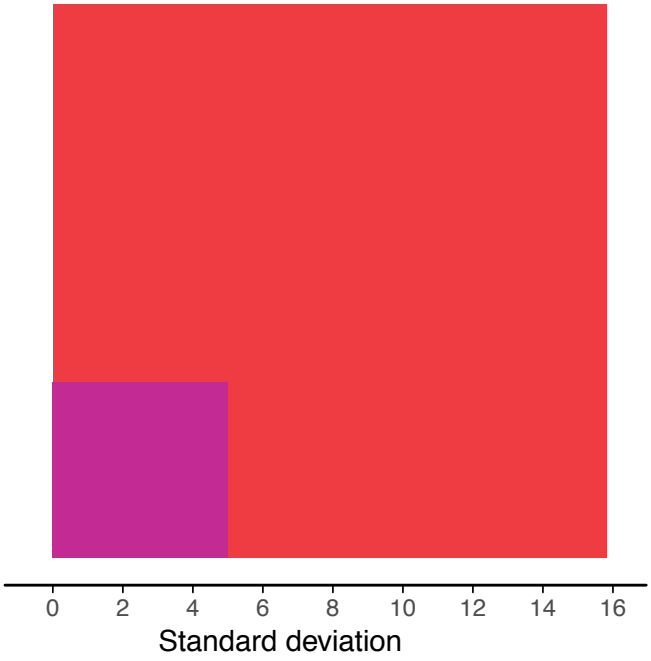


Figure 3-3 The addition of two components of variation

4. From a practical point of view, we want to estimate the benefit of elimination particular sources of noise. Each source costs considerable effort to eliminate. We therefore want to estimate how useful it will be in advance of putting in the effort. The way to do this is to look at the variability including the source, but subtract the variability from one level below the source, i.e. containing the variability from all nested levels below it. For example, if we want to know the effect of changing the probe placement (i.e. lifting it off the chest from a correct position and putting it back on the chest in a correct position), we should calculate the following:

$$SD_{Caused\ by\ changing\ probe\ placement}{}^2$$

$$= SD_{Observed\ on\ changeing\ probe\ placement}{}^2 - SD_{Observed\ between\ beats}{}^2$$

The important observation here is that if there is a large amount of variability, removing a small amount of variability actually has surprisingly small effect. The pithy epigram I propose for this is:

"In noisy data, removing small amounts of noise makes virtually no difference".

This can be depicted graphically in Figure 3 3 below. Here two operators are, for example, measuring EF in 50 patients. The researcher has calculated the signed difference between the EF values reported by the two operators and has found this to have a standard deviation of 9 EF units. The researcher knows from previous research that change of operator introduces a noise of 2 EF units. The researcher then re-runs the study with a single operator doing two scans per patient, expecting find an SD of only 7 EF units (since 9-2 = 7). Unfortunately, the researcher the SD of the second study to still be almost 9 EF units. The reason for this is that the calculation for the expected SD is as follows:

$$SD_{expected} = \sqrt[2]{SD_{with\ noise}{}^2 - SD_{of\ noise}{}^2}$$

$$SD_{expected} = \sqrt[2]{9^2 - 2^2} = 8.8$$

This can be visualised as subtracting a small square from a large square and

reshaping the area into a new square. If the side of the small square (representing its SD or CV) is much smaller than the side of the big square, then the area of the small square is *very* much smaller than the area of the big square. As a result, only a very small proportion of the area is removed, and the residual area, once reshaped into a square has a side similar to that of the original square. This is depicted below for an outer square of side 9, and an inner square of side 2. It is visually obvious, once looked at this way, that the side of the square composed of the residual area is not much less than the original 9 (Figure 3-4).

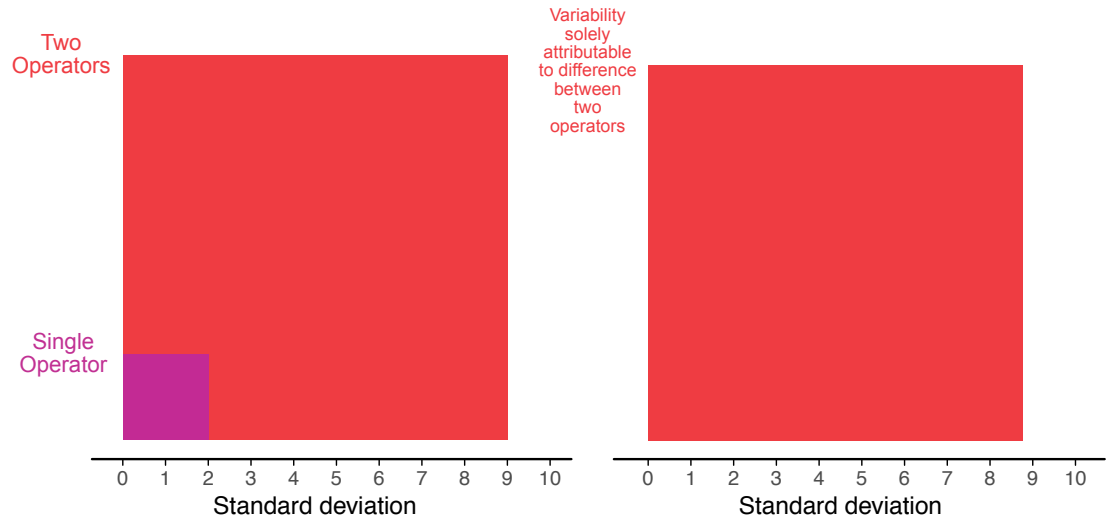


Figure 3-4 Subtraction of one source of variability.

This propagation of errors is valid for the addition and subtraction of standard deviations. For these experiments, we will use the coefficient of variation. As the coefficient of variation is equivalent to the standard deviation divided by the mean

value, and as these CVs refer to the same mean value. The addition and subtraction of these CVs will behave similarly.

3.4.3 Application of display method to the peak aortic velocity data

The variability between measurements of peak aortic stenosis velocity in a patient is broken down by the statistical method, and the resulting areas are numerically as follows (also shown in graphical form in Figure 3-7):

3.4.3.1 Tracing variability

Because the process of tracing does not have any subcomponents in the model, all the variability observed at the tracing level is considered to arise at the tracing level itself. It should be remembered I use the term tracing to bring to mind to the concept of a human reader drawing around an envelope, even though in the case of peak aortic velocity, the reader is only selecting a vertical level of the peak. I use this terminology because it is easy to remember what it means.

The tracing variability of aortic peak velocity had a coefficient of variation of 3.3%. This is shown as a purple square whose side is 3.3%. This both the variability observer at the tracing level, and the variability contributed by the tracing process.

3.4.3.2 Variability observed, and arising, at beat level

In the experiment of selecting different beats a total variability of 6.0% is *observed*. However, we know that tracing the same beat repeatedly would have a variability of 3.3%. Therefore, if we want to know how much variability arises at the beat level, we need to specifically calculate it as follows:

$$CV_{\textbf{\textit{arising}}} \text{ at beat level} = \sqrt{CV_{\textbf{\textit{observed}}} \text{ at beat level}^2 - CV_{\textbf{\textit{observed}}} \text{ at level } \textbf{\textit{below}}^2}$$

This is not 2.7% but $\sqrt{6.0^2 - 3.3^2} = 4.9\%$.

This can be visualised graphically (Figure 3-5) as a square of side 3.3% being subtracted from a coloured square of side 6.0%. The remaining coloured area, once reshaped into a square, has side of 4.9% (not 2.7%).

The conclusion from this analysis that the change in beat makes a larger contribution to variability than the act of tracing does.

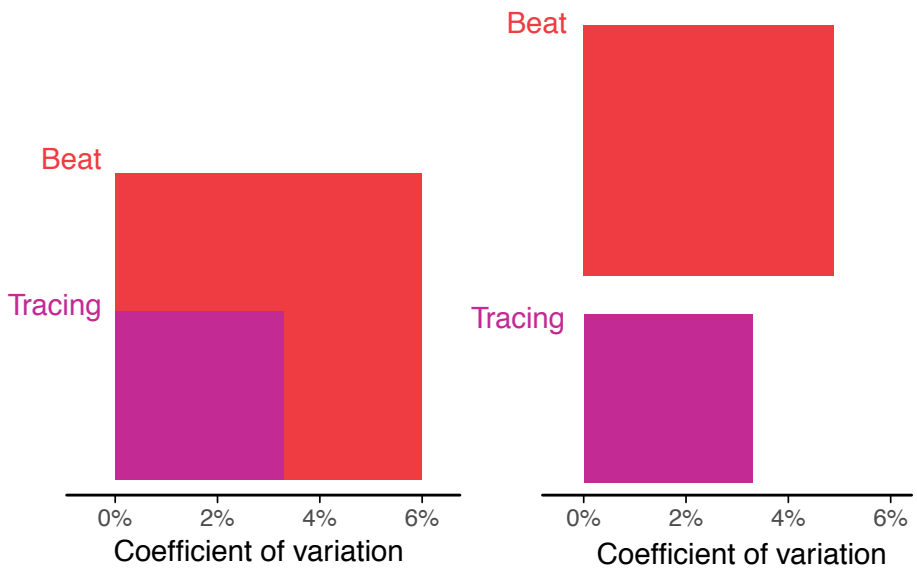


Figure 3-5 The "tracing" and "beat" level components of variation

3.4.3.3 Variability observed, and arising, at probe placement level

In the experiment of lifting the probe off and placing it back onto the chest, a total variability of 8.9% is *observed*. However, this process inevitably means looking at a

different beat, and we already know that looking at a different beat produces an observed variability of 6.0%.

Therefore, if we want to know how much variability arises at the probe placement level, we need to specifically calculate it as follows:

$$CV_{\boldsymbol{arising}} \text{ at placement level} = \sqrt{CV_{\boldsymbol{observed}} \text{ at placement level}^2 - CV_{\boldsymbol{observed}} \text{ at level } \boldsymbol{below}^2}$$

This is $\sqrt{8.9^2 - 6.0^2}$ = 6.5%.

This means that the miniscule differences in probe placement make a contribution larger than that of beat-to-beat variability and the act of tracing.

3.4.3.4 Variability observed, and arising, at the operator level

In the experiment of having a different operator acquire a different set of images, a total variability of 11.0% is *observed*. However, this process inevitably means the probe placement has inevitably changed, and we already know that changing probe placement produces an observed variability of 8.9%.

Therefore, if we want to know how much variability arises at the operator level, we need to specifically calculate it as follows:

$$CV_{\boldsymbol{arising}} \text{ at operator level} = \sqrt{CV_{\boldsymbol{observed}} \text{ at operator level}^2 - CV_{\boldsymbol{observed}} \text{ at level } \boldsymbol{below}^2}$$

This is $\sqrt{11.0^2 - 8.9^2}$ = 6.1%.

This means that, for peak aortic velocity, changing the operator makes slightly less difference than the same operator lifting and replacing the probe on the chest again.

3.4.3.5 Variability observed, and arising, at the reader level

In the experiment of having a different reader (who may be different from the operator) review a set of acquired images, a total variability of 11.2% is *observed*. However, in this experimental set up, this process included the variability from different operators acquiring different images which has an observed variability of 11.0%.

Therefore, if we want to know how much variability arises at the reader level, we need to specifically calculate it as follows:

$$CV_{\boldsymbol{arising}\text{ at reader level}} = \sqrt{CV_{\boldsymbol{observed}\text{ at reader level}}{}^2 - CV_{\boldsymbol{observed}\text{ at level }\boldsymbol{below}}{}^2}$$

This is $\sqrt{11.2^2 - 11.0^2}$ = 2.1%.

This means that the choice of reader makes only a minuscule contribution to the overall variability. It should be noted that all these readers were accredited in echocardiography.

3.4.3.6 Variability observed, and arising, at the visit level

In the experiment of the patient returning on a different scan for a repeat date, a total variability of 13.0% is *observed*. However, in this experimental set up, this process

included the variability from different readers reporting imaging from different operators.

Therefore, if we want to know how much variability arises at the visit level, we need to specifically calculate it as follows:

$$CV_{\boldsymbol{arising}} \text{ at visit level} = \sqrt{CV_{\boldsymbol{observed}} \text{ at visit level}^2 - CV_{\boldsymbol{observed}} \text{ at level } \boldsymbol{below}^2}$$

This is $\sqrt{13.0^2 - 11.2^2}$ = 6.3%.

3.4.3.7 Showing all the components

Of the variance components shown in Figure 3-6, only the value at the tracing level can be directly observed. All other direct observations of variability include not only the source being targeted, but also the tracing variably and all levels in between.

The simplest display of this process of sequential composition is a Russian doll sequence of square. This takes the areas shown in the left panel of Figure 3-6, reshapes them as in the right panel of Figure 3-6 and then rearranges them to be nested so that the effect of the different components in common can be visualised as in Figure 3-7. The value of this representation is that one can see the non-linearity of the combination of different coefficients of variation. The underlying data is shown in Table 3.2.

Table 3.2 Continuous Wave: Sources of variability and their component and observed coefficient of variation (CV)

Source of variability	**Continuous Wave**		
	Component CV	**95% Credible interval**	**Observed CV**
Tracing	3.3%	(3.2% to 3.6%)	3.3%
Beat	4.9%	(4.6% to 5.3%)	6.0%
Placement	6.5%	(6.1% to 7.0%)	8.9%
Operator	6.1%	(2.4% to 11.9%)	11.0%
Reader	2.1%	(1.4% to 3.5%)	11.2%
Visit	6.3%	(4.8 to 11.1%)	13.0%

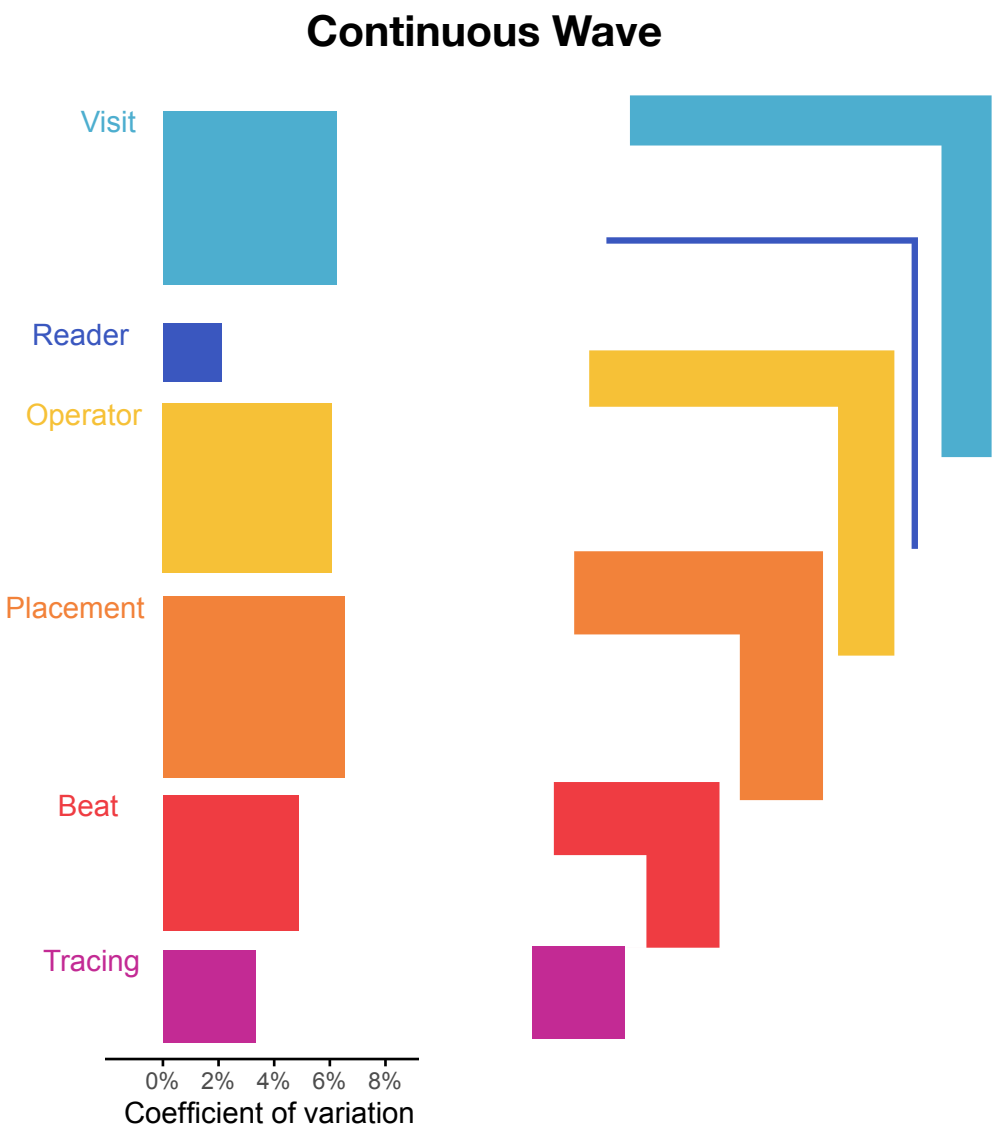


*Figure 3-6 The **Component view** of the **peel-away-**variance **model** of the sources of variability contributing to the echocardiograph measurement of the continuous wave LVOT peak velocity*

Component view: Each component is represented by a square whose area shows its variance and whose side shows its coefficient of variation (left). The components are "reshaped" so they can be nested (right). They are then combined in Figure 3-7.

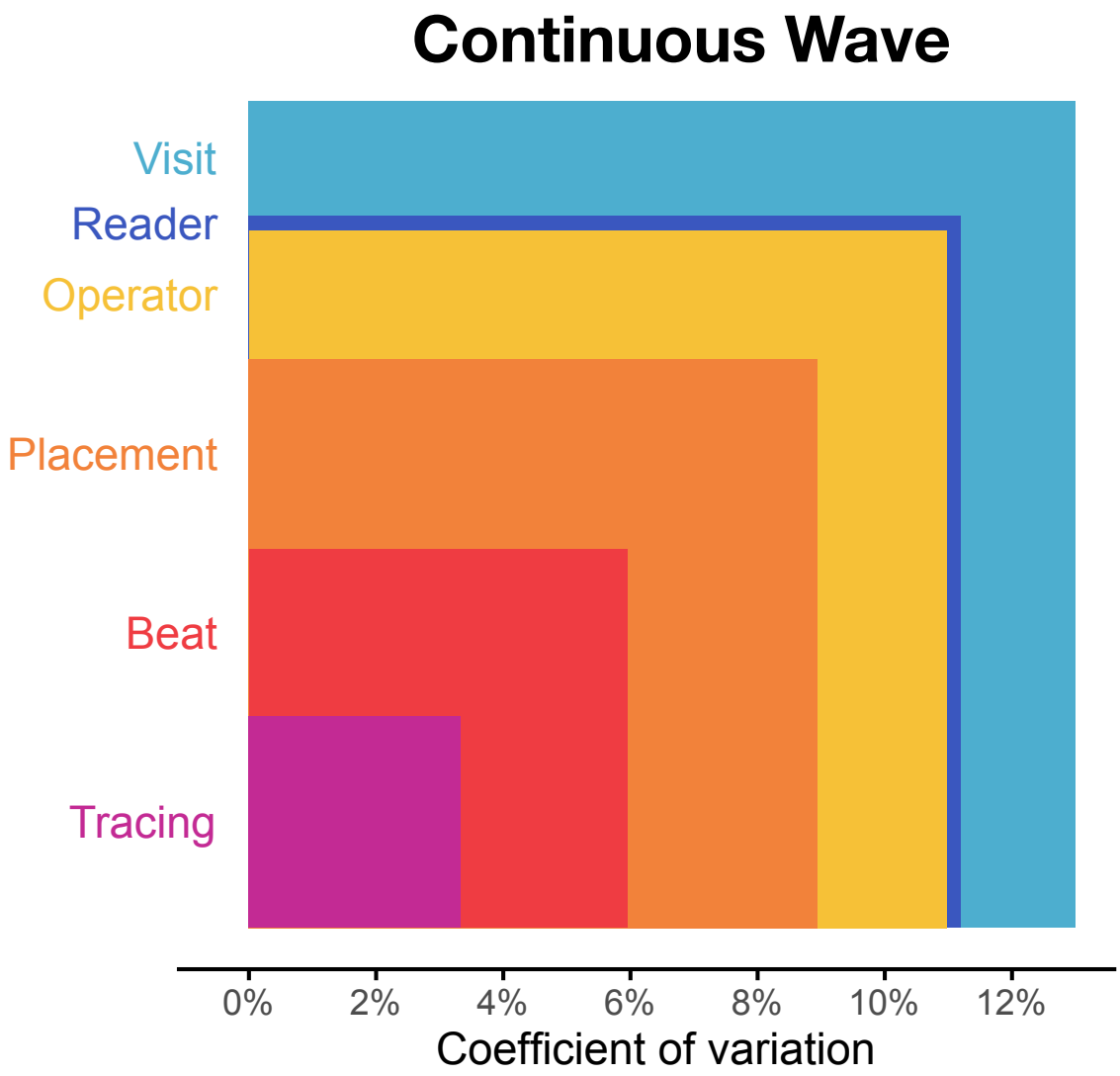


*Figure 3-7 The **Peel-away-variance representation** of the **peel-away-variance** model of the sources of variability contributing to the echocardiograph measurement of the continuous wave peak trans-aortic velocity*

Peel-away-variance representation: It is the arithmetical total of the variances (areas) that determine total variance, as shown in this view. The side of the overall square represents the overall coefficient of variation.

3.4.4 Application to LVOT peak velocity

The process described above I also applied to the LVOT data. These are shown in tabular form in Table 3.3. The components are shown in Figure 3-8, and the corresponding "peel-away variance representations" are shown in Figure 3-9.

A notable feature is that the variably contributed by change in operator and change of visit are tiny compared to the remaining variability. In fact, probe placement (which includes sample volume placement) contributes an overwhelmingly dominant component to the overall variability.

This is important because it means that being careful to choose the same operator, the same reader, and averaging multiple beats will have virtually no beneficial effect on the overall variability.

Table 3.3 Pulsed Wave: Sources of variability and their component and observed coefficient of variation (CV)

	Pulsed Wave		
Source of variability	**Component CV**	**95% Credible Interval**	**Observed CV**
Tracing	4.5%	(4.3% to 4.7%)	4.5%
Beat	5.9%	(5.6% to 6.2%)	7.4%
Placement	11.3%	(10.2% to 12%)	13.7%
Operator	0.0%	(0.0% to 13.6%)	13.7%
Reader	2.9%	(2.0% to 5.2%)	14.1%
Visit	0.3%	(0.0% to 9.9%)	14.1%

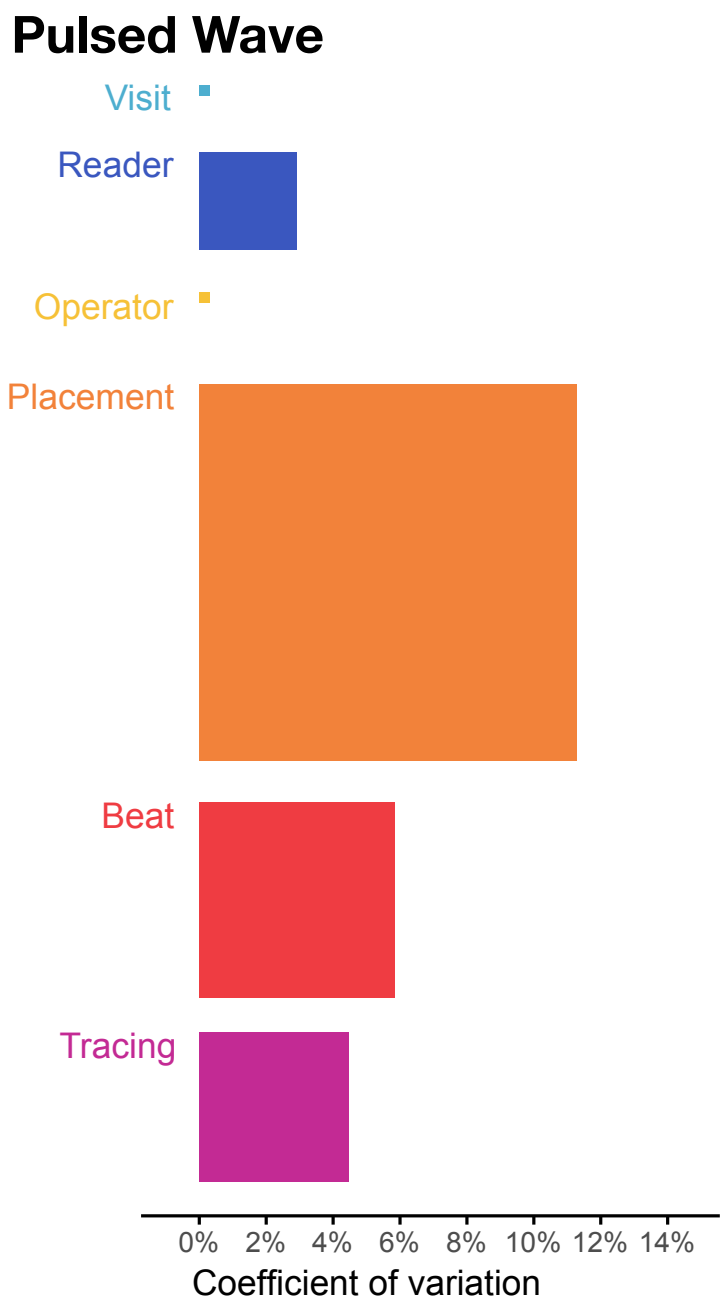


*Figure 3-8 The **Component view** of the **peel-away-variance model** of the sources of variability contributing to the echocardiograph measurement of the pulsed wave LVOT peak velocity*

Component view: Each noise component is represented by a square whose area shows its variance and whose side shows its coefficient of variation

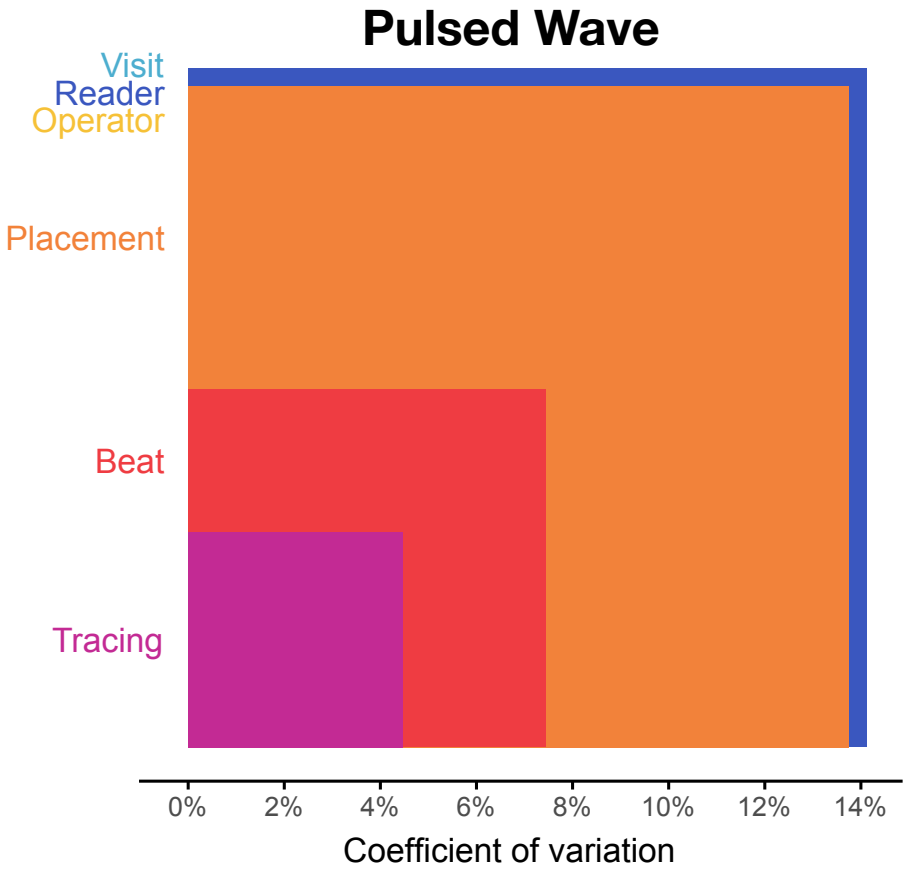


*Figure 3-9 The **Peel-away-variance representation** of the **peel-away-variance model** of the sources of variability contributing to the echocardiograph measurement of the pulsed wave LVOT peak velocity*

Peel-away-variance view**:** It is the arithmetical total of the variances (areas) that determine total variance, as shown in this view. The side of the overall square represents the overall coefficient of variation.

3.4.5 Components of variability

Overall, the total *observed* variably over visits was 13% for continuous wave, and 14.1% for pulsed wave.

The *observed* CV between different operators was 11.0% for continuous wave peak velocity. However, the majority (8.9%) of this variance was due to variability arising at the levels of probe placement, change of beat, and episode of tracing. The *true* variance between two operators is only 6.1%.

For pulsed wave doppler, this distinction was even more startling. Of the 14.1% *observed* variation between operators, virtually all of it was explained by the 13.7% arising at the levels of probe placement, change of beat, and episode of tracing.

Overall, the inevitable subtle differences in probe placement were the greatest source of error for both continuous wave (coefficient of variation 6.1% and 11.3% respectively), with less variability with CW as compared to PW ($p < 0.0001$).

Tracing and reader variability are readily amenable to automation. Their coefficients of variation were 3.3% and 2.1% for CW and 4.5% and 2.9% for PW respectively. They were smaller for CW than PW ($p < 0.0001$, $p=0.19$ respectively). In aggregate their combined CV were 4.0% (CW) and 5.4% (PW).

All components are shown in Figure 3-10, with their confidence intervals shown in

Table 3.2 and Table 3.3.

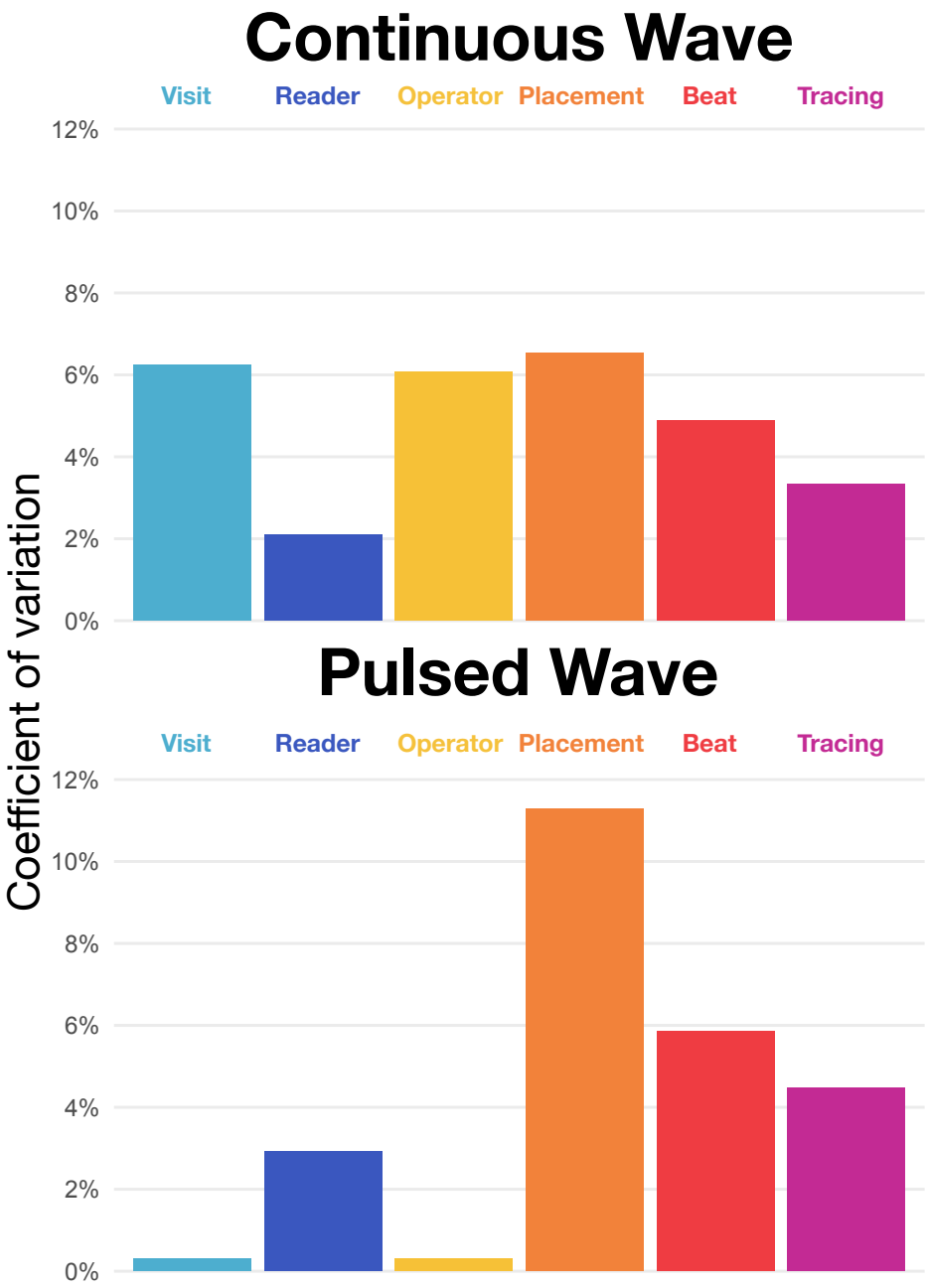


Figure 3-10 The individual components that contribute to the variability of both continuous wave and pulsed wave Doppler.

For both continuous wave and pulsed wave Doppler probe placement was the greatest contributor to the variability of measurements.

3.5 Discussion

For the first time, this study has shown that the individual components that make up the variability of aortic stenosis measurements can be distinguished and separately quantified without carrying out an impossibly complex complete combinatorial grid. This is achieved using a mixed effects model, using MCMC to solve it. This has never been done before to assess sources of variability in cardiovascular imaging.

During this work I also developed a unique method to display the results to make it obvious to clinical readers to show how the sources combine with each other to create overall variability, and the Pythagorean relationship that means that only moderately larger sources of variability are, in fact, overwhelmingly dominant in determining the overall coefficient of variation.

I had several specific findings:

3.5.1 Surprising dominance of probe placement component

I could find no report in the literature that makes clear that the single biggest contributor to variability between measurements of aortic stenosis velocities was the miniscule, unnoticed differences between probe positions and sample volume placements that seem clinically perfect.

This may be because there has been never been a study that specifically extracted this component. I suspect that this is the case and the underlying reason is that the statistics are considered too difficult and the interpretation of the results too abstract.

My adaptation of relatively straightforward Markov-Chain Monte-Carlo methods and development of a simple visualisation may resolve these problems.

There are two practical impacts of discovering that subtle variation in probe placement (within the range of clinically ideal placements) has the dominant influence on measurement variability.

First, the bad news is that even if an enormous effort was put in to resolving the other problems, and that effort was 100% successful, there would be a disappointingly small reduction in the overall variability. For example, the coefficient of variation for PW peak velocity would fall from 14.1% only as far as 11.3%, despite removing components whose individual sizes are 4.5%, 5.9%, 2.9%, 0.3%.

Second, the good news is that we now know the most fruitful component to target, namely the impact of subtle, unnoticed, probe placement and sample volume variability.

3.5.2 Opportunities for reducing variability originating at probe placement level

There are two approaches to reducing variability from a source. The first is universally effective. It is to make multiple samples of the variability (i.e. do that thing several times), specifically not trying to make them all the *same* as each other, but rather making them *independent*, but all individually good.

In the case of measuring aortic stenosis velocities, this would involve repeatedly taking the probe off the chest, attempting to achieve a good position each time, but without automatically reproducing the previous position.

The second approach is to do something to directly reduce that source of variably. This can be difficult.

In the case of aortic stenosis, for pulse wave measurement, the variance contribution of probe placement is visibly the dominant contributor (Figure 3-8). There are many reasons for this. The flow field in the LVOT is complex and does not conform to the plug flow shape when teaching the continuity principle. The complexity means that in slightly different sample volume positions, the velocity may be different to an important degree. This is particularly so because LVOT velocities are relatively low. Also, there is no particular landmark to aim for. Moreover, the LVOT is a moving target. Finally, the flow-field need not be cylindrical, nor indeed the LVOT itself.

My proposal is three-fold:

(1) Modify the acquisition protocol to require multiple independently positioned, but each clinically ideal, probe placements (i.e. with removal of the probe in between each episode). This applies the first approach described above.

(2) Assist the operator by partially automating the process of finding the correct sample volume. This applies the second approach described above.

(3) I will incorporate further improvements that make use of correlated noise components in different variables, that can be made to cancel out by enforcing simultaneity of measurement.

To deliver (1) and (2) will require automation of image analysis. This is because the greatest labour in measuring a beat of aortic stenosis is the human element of manually marking the velocity. While an operator will certainly be happy to acquire more strips of data, they will not be happy to analyse these extra strips. This is evident even on teaching documents of how to make echo measurements, where, despite the advice to measure multiple beats and take the average, almost universally the example shown is of a single beat being measured, and that single value being used in all downstream calculations.

To deliver (3) I will show how it is possible to simultaneously measure the LVOT and aortic valve velocities. This is so that the variation *caused* by the difference in beats is reduced.

In Chapter 4, I will develop a method of automation that provides a basis for all three advances, (1), (2), and (3).

3.5.3 These data suggest that current advice given to clinicians and researchers is wrong and harmful

3.5.3.1 Wrong recommendation 1: Keep same operator and reader

One step commonly recommended to minimise variability is to arrange for a consistent member of staff to be scanning the patient on each occasion, and perhaps a consistent reader (who might be the same person) performing the measurements.

Unfortunately, my research shows that these steps will not work. The variability contained in the choice of operator and the choice of reader is negligible in comparison to the other sources (Figure 3-11).

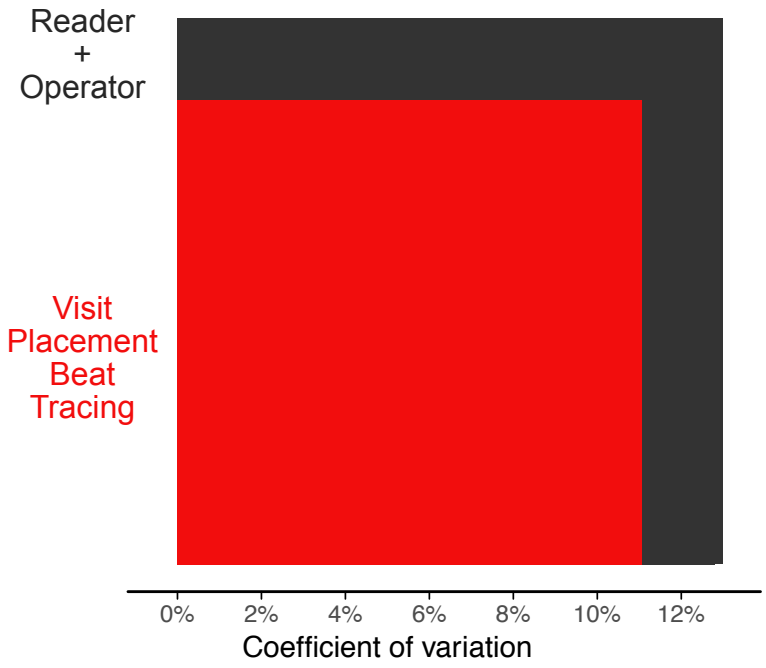


Figure 3-11 The removal of reader and operator variability only has a modest effect on overall variability (continuous wave)

The by utilising the same operator and reader for each acquisition, the reader and operator variability can be eliminated. However, even though

individually, their CVs are 2.1% and 6.1%, their elimination only reduces the total CV from 13.0% to 11.1%.

The advice is almost useless, and I consider it harmful as it is damaging to the workflow of the department to arrange a consistent member of staff to be available for rescanning that patient. Either one member of staff does the aortic stenosis measurements for all patients of the trust, or each patient is allocated a personal sonographer who stays consistently with that patient throughout their life. Both of these create scheduling inflexibility, and special chaos when there is absence due to sickness or other duties.

The energy expended by a hospital trying to deliver a fixed or personalised sonographer and reader is wasted. This is harmful because that energy and attention could instead have been placed on something (anything) more useful to patient's welfare, and/or hospital finances.

3.5.3.2 Wrong recommendation 2: "Need more training"

When there are problems of reproducibility, we frequently just to the assumption that this is due to inadequate training. However, in my experience this is almost never backed up by a blinded study showing that people who have undergone a training process deliver better blinded reproducibly than those with the same baseline training randomised to not have the training process. I suspect this is because we all assume that more training always worthwhile. It might not be.

In my study, all operators were accredited and experienced, and even still there was an overwhelming contribution from unnoticeable probe and sample volume placement variability. Perhaps, unskilled operators would have shown more variability at the operator level, but there is no doubt that these operators were experts, and despite this there was a substantial segment of variability persisting.

3.5.4 Targeting variability

The variability of a measurements can be captured by combining multiple readings. Indeed, guideline bodies such as the ESC and ASE recommend averaging across three to five (Baumgartner et al. 2009). However, they neither provide a rational for this exact number, nor an estimate of what level of reproducibility that would be obtained before or after averaging. Furthermore, averaging over a few consecutive beats will only reduce some of the components – tracing and beat. Sources such as placement, reader, and variability associated over longer time scales will not be reduced.

For example, considering just the variability associated at a single visit, if three consecutive beats are averaged the coefficient of variation will fall from 11.2% to 10.0% (CW), and 14.1% to 12.6% (PW). However, if they were taken from separate acquisitions, then they would fall further to 8.3% (CW) and 8.3% (PW).

Within a research study serially measuring aortic stenosis severity, a CV of variation of 8.3% for a single value is equivalent to a CV of 11.9% (CW) between two values.

3.5.5 Effect of variability on management

3.5.5.1 Variability compromises ability to assess progression

Accuracy in measuring aortic stenosis severity is crucial in a patient's management. Rapid progression over serial echocardiograms has been used as an indicator for intervention, with a threshold of 0.3m/s/year. Of a patient with moderate stenosis (say 3m/s) this represents a change of 10%. If the coefficient of variation of the change is over 10% then the 95% confidence interval for change ranges from -17% to 21% (the interval is symmetric on a log scale), then this criterion would often be met even if there was no *true* change. Even with the three repeated acquisition, the differences between two measurements would range from -13% to +14%, again leading to many false reports of rapid progression.

Using current techniques, it is very difficult to measure their progression over a year with enough precision to reliably use this criterion. Fortunately, it has been de-emphasised in recent guidelines.

For clinical researchers, the larger the variance, the greater the number of patients that need to be recruited.

3.5.5.2 Variability complicates assessing proximity to a threshold

Whilst it is exceptionally unlikely that a single arbitrary threshold such as 4 m/s exactly demarcates which patients would benefit from intervention, this nevertheless is the "golden criterion" in both practice and as recommended by guidelines.

Even for a simple measure such as aortic peak velocity, if there is noise in the measurement it is possible for the measurement containing the noise measurement to be on the wrong side of the decision threshold. The chance of this happening rises as the underlying value approaches the action threshold, and is approximately 50% in that region. What I mean by this is if the true velocity is 3.95m/s, about half of the measured values would below this, and half above, i.e. they would be 4.0 or above. Similarly, if the true value was 4.05 m/s, about half the time the measured value would be below 4.0 m/s.

3.5.5.3 Variability opens the door to unintentional clinician bias

When serial echo data is analysed unblinded their sensitivity to changes decreases as operators bias results towards the mean (Gosse et al. 2004). Clinicians, when confronted with serial measurements of a parameter they expect to at least stay either the same or progress, may be reluctant to report a value below the previous one. Measurements that are higher due to chance will not be detected by this process and may be interpreted as evidence of genuine deterioration. As a result, both clinically, and in observational studies, there will be a constant bias towards the exaggeration to the rate of progression of disease.

3.5.5.4 Variability is compounded in composite measures

In general, when two measured values are multiplied or divided to produce a third, calculated, value, the noise percentages in the original measurement add together to

produce the noise percentage in the third value. This general statistical principle applies whenever the noises in the two measured values are uncorrelated.

The consequence of this, for the dimensionless index (the ratio between the aortic valve and LVOT velocities), is catastrophic. Aortic valve peak velocity has a large coefficient of variation (11.2% in my data) and LVOT velocity has an even larger one (14.1%). Calculating the ratio between the two creates a value with of coefficient of variation of 18% This means a 95% uncertainty range of, i.e. +1.95 SD of approximately -28% to +38%.

In practice, clinical sonographers reduce this by taking out outlying values, achieving a real-world coefficient of variation of 12.6% (Finegold et al. 2013). Yet, even this is not satisfactory, because:

- This means that if DI is measured to be X, if it was remeasured blinded to the first measurement, in 95% of cases, the value would be between X-21% and X+26%. That is to say, a dimensionless index of 3.3 on the immediately successive measurement would be in the range 2.6 to 4.2, 95% of the time.

- The sonographer has to identify the value as an outlier, as reference to the visual appearance or the previous scan result. In research practice this would be misconduct.

The reproducibility of the aortic valve area, which incorporated a third measure, that of the diameter of the LVOT squared, will be even worse.

3.5.6 Variability limits the maximum agreement between two measurement techniques

In general, a measurement technique cannot agree with another more than it agrees with itself (Spearman 1904). A lack or test-retest reproducibility limits how much a variable can correlate with itself, let along another noisy variable.

For example, the high correlation of simultaneous in vivo Doppler and dual catheter measurements of aortic gradients established echocardiography as the "gold-standard" technique for the quantitative assessment of aortic stenosis (Currie et al. 1985). However, the authors' enthusiasm was tempered by the significantly worse concordant they found with non-simultaneous echocardiographic measurements (such as those from the referring clinic). This is not unexpected, the addition of biological variability over non-simultaneous measurements would naturally have been expected to limit this agreement.

Improvements in precision of a measurement will allow for both greater agreement with other measures and for predicting outcomes.

3.5.7 Limitations

This study did not investigate the impact of equipment model or make. For other, more complex echo measurements, this has been shown to significantly influence measurements. However, in that case it was due to the complex algorithms related to speckle tracking.

Whilst this study utilised 17 operators and 13 different readers, it did not recruit from multiple centres. Therefore, it is possible that different local protocols and policies may have led to results that may be better (or worse) than other centres.

All operators and readers were accredited in echocardiography. It is possible that the coefficients of variation would be larger in those still in training. We did however, include sonographers with a range of post-accreditation experience, who worked at multiple sites and hospitals. I therefore expect the results to be generalisable.

I did encounter some statistical issues. Despite the large dataset (over 10,000 observations), the 95% credible interval on the estimates on some of the components (operator and visit) has significant uncertainty. The random-effects model and computational environment provided unstable estimates with larger coefficients of variation as it was a partially crossed factor. Other authors have run into similar issues (Schroeder et al. 2004).

3.5.8 Comparison with other literature

Analyses like this have not been done in cardiovascular imaging. However, they have been done in other fields of cardiology (Schroeder et al. 2004), and are very common in other scientific disciplines. Furthermore, very few studies recognise the fact that two different sonographers may disagree almost as much as one disagrees with themselves, and then use this to calculate the true component contributed by the change in sonographer.

Failure to recognise that the observed differences between two sonographers are not caused by the change in sonographer has caused us as a discipline to give the wrong advice, such as sticking with the same sonographer rather than exploring to find the real cause. Furthermore, it has hindered the development of genuinely effective solutions.

A second problem with some of the literature is that it presents not standard deviations, coefficients of variations, or Bland-Altman Limits of Agreement, but correlation coefficients. Interpreting these can be fraught with hazard, as the correlation coefficient is highly dependent on the underlying distribution of the values. A study that includes a wide range of values (perhaps mild to severe aortic stenosis) would report a larger correlation than a study only including a limited range (say just moderate aortic stenosis).

However, the Bland-Altman limits of agreement technique does have some limitations. It was intended to be used to compare two different measurement methods rather that two different operators in the same method. It therefore leads to an experimental set-up limited to two specific observers, readers, or operators. This provides a good estimate for the agreement between the two specified operators, as if they each were a method, but may be poorly generalisable to multiple operators. Just as if you had a study to estimate the range of heights in a population you would not only measure two people, if you want to estimate the range of results different sonographers get, you should not only assess two sonographers.

My study therefore uniquely provides estimates for the individual contribution of many of the components that contribute to the variability of measurements of aortic stenosis. These estimates are comparable with our previous retrospective analysis (Finegold et al. 2013), but larger than previous estimates (Ramirez & Wong 1985; Siostrzonek et al. 1988) which reporting a coefficient of variation for Vmax ranging from 5% to 7%.

3.6 Conclusion

In this Chapter I have shown that "peeling away" elements to understand the components of measurement variability is feasible.

In our example, the variability mainly arises not, as traditionally believed, from differences between operators or readers, but rather the unavoidable impossibility of placing the probe in the exact same place each time. Beat-to-beat biological and tracing variability also make large contributions

This approach allows researchers, such as myself in this thesis, to rationally target the contributory components to reduce measurement variability.

4 The development and validation of the “instantaneous Dimensionless Index”: a novel tool to more reliably follow-up aortic stenosis

4.1 Abstract

4.1.1 Introduction

In this Chapter I explain the rational behind, the development of, and the clinical validation of the "instantaneous Dimensionless Index" (iDI). This provides automated quantification of the Dimensionless Index (DI) using the CW Doppler spectrum alone. It is vendor-independent and uses novel automated assessment techniques.

Hypothesis: The automated instantaneous Dimensionless Index, generated from just continuous wave Doppler spectra, produces values equivalent to the Dimensionless Index but with greater precision.

4.1.2 Methods

I developed the instantaneous Dimensionless Index. The iDI improves the reproducibility of the assessment of aortic stenosis though three techniques. Firstly, automated thresholding, tracing, quality control, and measurement eliminates tracing and reader variability. Secondly, mandated collection of multiple-beats, enabled through automation can reduce variability through averaging. Thirdly, it uses a custom-designed ensembling routine to simultaneous extract LVOT and aortic valve data using conservation of mass principle. This eliminates beat-to-beat and most probe position variability when calculating the dimensionless index. A description and rational of the design is provided.

I then prospectively recruited patients with native or tissue valve aortic stenosis from my hospital and its community clinics (Patient Group A). As part a pilot to verify that my technique can be carried out easily by other echocardiographers, a colleague at a different institution also captured data from patients drawn from her own institution (Patient Group B).

Multiple recordings of multiple-beats of continuous and pulsed wave Doppler were made in every patient. To validate that the 'inner' profile of the continuous wave trace reflects the LVOT velocity, these were calculated across the multiple acquisitions in each patient, averaged, and then compared. To validate the iDI, we calculated both it and the conventional DI (based on peak velocities), with values averaged across multiple acquisitions and then compared. Finally, a cohort of patients were re-assessed at 6 months to calculate both the same-day and test-retest reproducibility.

4.1.3 Results

I recruited 90 patients with aortic stenosis (DI 1.14 to 5.04) at my institution (Patient Group A). The mean difference between LVOT velocity extracted from the CW Doppler and the LVOT velocity from PW Doppler was minimal (-0.02 m/s, -0.05 to 0.02, p=0.3). This resulted in minimal bias between iDI and DI (-0.08 units, -0.14 to -0.02, p=0.005), with a high correlation between the two measures (rho=0.93, p<0.0001). As compared to the DI, the iDI significantly reduced the coefficient of

variation on same-visit test-retest (5.8% vs 12.9%, p<0.001), and at follow-up (9.1% vs 16.6%, p<0.001).

My colleague recruited 10 patients with aortic stenosis (DI 2.70 to 4.71) at her institution (Patient Group B). Similarly, there was minimal, statistically non-significant bias between the iDI and DI, with iDI values being, on average -0.04 units less than DI (95% CI -0.10 to 0.02, p=0.194). Furthermore, the same-visit coefficient of variation for iDI was significantly less than the coefficient of variation for DI (3.8% vs. 17.0%, p=0.004).

4.1.4 Conclusions

First, iDI doubles the precision of aortic stenosis quantification. Using the iDI in a clinical trial with aortic stenosis severity as an endpoint would require 3-to 4-fold fewer patients.

Second, iDI values can be used as a direct replacement for traditional Dimensionless Index, because they have no clinically significant bias against, and strong correlation with, traditional DI.

Third, an independent clinician, with no involvement in the origination of the iDI software are able to acquire data for automated calculation of iDI.

4.2 Introduction

The motivation for developing the "instantaneous Dimensionless Index" (iDI) was to provide clinicians and researchers with a reproducible biomarker that fulfils five requirements:

1. Feasible in the majority of patients with minimal operator training.
2. Provides a measure that can be immediately interpreted with minimal training.
3. Improves precision, with a target of four-fold reduction in the relative variance attributable to noise.
4. Takes no more time than the current clinical approach.
5. Does not depend on particular make or model of echocardiograph.

As discussed in Chapter 3, multiple sources of variability during both acquisition and analysis contribute to the irreproducibility of the traditional dimensionless index (DI).

During acquisition:

- **Non-simultaneity of the continuous wave and pulsed wave measurements**: The continuous wave and pulsed wave recordings are not simultaneously acquired. Consequently, errors in both measurements are compounded when divided.

- **Beat-to-beat variability**: Biological variability due to spontaneous fluctuations in stroke volume.

- **Position variability**: Variation due to un-noticed subtle differences in probe position.

- **Operator variability**: Additional variation arising from a change of operator.

During analysis:

- **Tracing**: Variation when the same reader looks at the same stored beat on a second occasion.

- **Reader**: Variation when different readers review the same trace.

The *instantaneous Dimensionless Index* algorithm incorporates the following elements to eliminate most of these sources of variability (Figure 4-1).

- Automated thresholding, tracing, quality control, and measurement eliminates tracing and reader variability.

- Algorithm-mandated collection of multiple beats (6 to 8) worth of data, enabled through automation, reduces variability through averaging.

- A custom-designed ensembling routine to simultaneous extract LVOT and aortic valve data using conservation of mass principle. This eliminates beat-

to-beat and most probe position variability when calculating the dimensionless index.

My implementation can calculate the ***instantaneous Dimensionless Index*** in either real-time from video output of echocardiographic machines, or post-process DICOM files. It works across echocardiography machines from multiple manufacturers and requires no additional acquisition time.

4.2.1 Non-simultaneity of a large element of error.

As I discussed in Chapter 3, both the aortic and LVOT velocity profile are subject to random variation on a beat-by-beat basis from changes in cardiac contractility, pre-, and after-load. Furthermore, by obtaining them on separate acquisitions, there will be subtle differences in probe-position and sample volume placement. As the dimensionless index is calculated by taking the ratio of the aortic and LVOT velocity, these uncorrelated errors become compounded and therefore increase variability. Therefore, ironically, whilst the aim of the DI is to adjust for large differences in flow by comparing the aortic to the LVOT velocity and hence reduce variability; the smaller beat-by-beat differences are in fact compounded and increase variability.

The fundamental problem is lack of simultaneity. There are fluctuations in circumstances disturbing one variable, but a different set of fluctuations disturbing the other variable. If the two variables could be measured simultaneously, there is, in principle, an opportunity to greatly reduce the variability of the ratio.

4.2.2 Overwhelming advantages of simultaneous measurement.

In principle, simultaneous recording of the aortic and LVOT velocities can allow the ratio to be calculated with smaller error than either of the constituent variables (Figure 4-2).

It is this concept that I explored. It has been noted that the LVOT profile is visible within the continuous wave acquisition (page 135 Maus et al. 2016). However, it is usually faint, indistinct, and not suitable for accurate measurement (Figure 4-1, left panel).

In my thesis I developed software and algorithms, and associated hardware, to make it possible for ordinary clinical staff to acquire the ratios of the double profiles from individual beats. I called this the *"instantaneous Dimensionless Index"* because it can calculate a ratio from a single heart-beat.

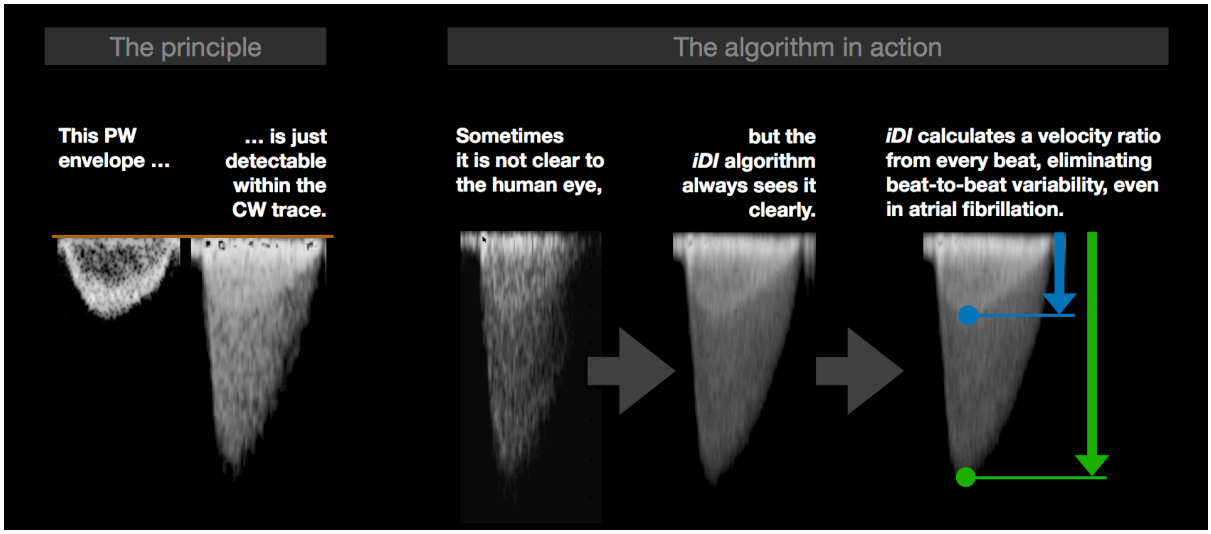


Figure 4-1 The principle of the instantaneous Dimensionless Index

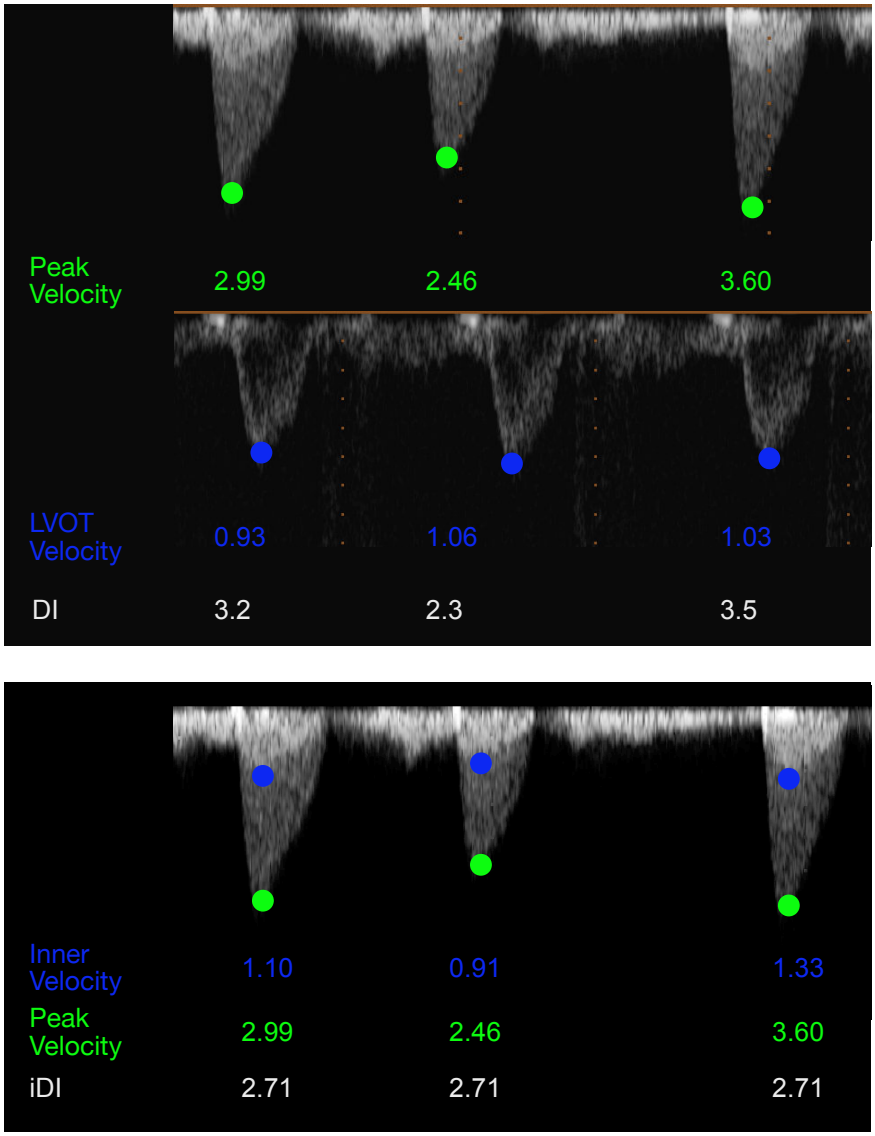


Figure 4-2 The simultaneous measurement of the LVOT and peak velocity can reduce measurement variability

(Top panel) In this patient with an irregular RR interval, the peak continuous wave velocity varies from beat-to-beat. The non-simultaneously measured pulsed-wave LVOT velocity similarly varies from beat-to-beat. The ratio of these values (the dimensionless index) therefore will vary substantially (in this example from 2.3 to 3.5) depending on which beats are selected.

(Bottom panel) The same continuous wave strip as above is shown. The LVOT profile is visible within this trace. Measuring both simultaneously and calculating the ratio (the "instantaneous Dimensionless Index") produces a much more stable result with far less variability.

4.2.3 Steps to the solution

I realised that if I was able to "enhance" the Doppler tracing it may make simultaneous extraction of these components possible with a resultant reduction in variability. I make this possible by employing the principle that the ratio between the two should be consistent between one-beat and another.

My approach is illustrated in Figure 4-1. The pulsed-wave envelope (first of the five profiles) is just about detectable as an inner profile within the continuous-wave trace (second profile).

In typical cases, however, the inner profile is relatively indistinct (third profile). My algorithm is to ensemble multiple-beats to produce a single signal-averaged profile (fourth profile). This step is complicated because not all the individual beats being ensembled are of the same height and duration. However, by normalising each beat to the same height, they can be combined and to produce a "ensembled" consensus beat. This ensembled beat can then have the inner and outer velocities measured using the knowledge that the two profiles should be proportional to provide additional noise resistance. The instantaneous dimensionless index can then be calculated.

4.2.4 Validation of the instantaneous Dimensionless Index.

The iDI provides an automated value for the Dimensionless Index. However, just obtaining a value is not satisfactory.

- We need to know in what proportion of day-to-day clinical cases of aortic stenosis, iDI can be calculated.

- We need to know how its precision compares to the traditional DI.

- We need to know how the values it obtains compare to the traditional DI.

- We need to establish whether the technique is straightforward enough to be used by third parties, rather than mysteriously working in the hands of the inventor, as sometimes happens with medical technology (Zaman et al. 2014).

I therefore also set out to design and conduct a study to test this across a wide spectrum of patients at my own institution, complemented by a further validation by an independent researcher at a different institution.

4.3 Methods

4.3.1 Development of the instantaneous Dimensionless Index

4.3.1.1 Steps of processing

The calculation of the "instantaneous dimensionless index" is a multi-stage process which includes the following sections:

- **Data acquisition**: Runs of continuous wave Doppler through the aortic valve are acquired and assembled into a long strip of beats.
- **Image cleaning**: Artefacts such as calibration marks are removed, and the image clipped to the baseline before being converted to grey-scale
- **Thresholding:** Using a custom algorithm, an appropriate grey-scale threshold is selected to binarize the image.
- **Tracing:** The outline of the Doppler waveform is traced.
- **Peak identification:** A robust method is used to identify the true peaks.
- **Removal of poor-quality beats**: As the heart moves during the respiratory cycle, the quality of the tracing will vary during an acquisition. Poor quality beats are removed.
- **Beat ensembling:** The individual beats that pass quality control as then ensembled.

- **Extraction of both peaks of the dual profile:** The inner peak (which represents the LVOT peak velocity) and the outer peak are extracted and a ratio taken to generate the instantaneous Dimensionless Index.

4.3.1.2 Software

All software was developed by myself in Python programming language (Python Software Foundation 2017). The final product is an installable application that can be used on Windows and macOS. As previously described, it can process DICOM files exported form echocardiographic machines or imaging system, or video obtained using a frame-grabber plugged into the video output an echocardiograph.

4.3.1.3 Data acquisition

The algorithms for the automated, reproducible assessment of aortic stenosis (and mitral regurgitation in Chapter 5) requires the acquisition of long recordings of either continuous wave Doppler through the LVOT and aortic valve, or colour M-mode Doppler through the mitral valve.

Whilst the developed algorithms can operate on data from a few beats (say 3) such as those typically captured on a "freeze-frame" image, longer runs of 5 to 10 beats (or more) are preferable.

To allow for the greatest utility three techniques were developed:

(1) Acquisition of DICOM files.

(2) Acquisition from the video-out port of an echocardiographic machine.

(3) Live acquisition and processing from a video-out port of an echocardiographic machine.

4.3.1.4 Acquisition via DICOM files

On most echocardiographic machines, rather than acquiring a freeze-frame of Doppler data, a multi-beat movie can be acquired – with the sliding-bar looping across the screen. This movie of Doppler data can be extracted from the DICOM file, the sliding-bar detected, and a long strip generated (Figure 4-3). The DICOM file also encodes the scale and sweep speed which is used in later processing.

Some echocardiographic machines, for example GE, save the raw spectrum data in private, proprietary portions of the DICOM file. The spectrum is stored at the fastest sweep speed; allowing for later display at a reduced sweep speed (so that more beats can fit on the screen). I was unable to find any freely available documentation on the open internet, or open source software that was able to decode them. I was however able to reverse engineer the storage format to allow me to extract the full resolution Doppler acquisition.

4.3.1.4.1 Acquisition from the video-out port of an echocardiographic machine.

The majority of echocardiographic systems have a video-out port, which may be analogue (VGA), or digital (DVI or DisplayPort). Using a video-capture device and a laptop, the image on the monitor can be captured live. This can be stored for later off-line analysis. The algorithm that follows in this Chapter is able to run in real-time; constantly processing the last 10s worth of data and displaying the result.

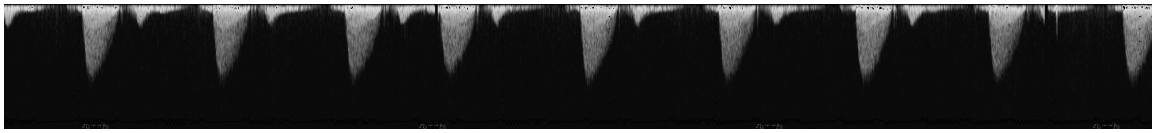


Figure 4-3 A continuous wave Doppler acquisition from a patient with aortic stenosis

4.3.1.5 Image cleaning

Image sanitisation consists of three stages:

1) Localisation of the baseline and removing the Doppler data in the alternate direction.

2) Removal of the superimposed ECG.

3) Removal of the white dots and other labels superimposed over the spectrum by the echocardiographic machine.

Typically, the spectrum is stored in RGB (red-green-blue) format within the image. The spectral data within it is grey-scale. To detect the baseline, which for example, is brown on Philips echocardiographic machines, the image is transformed from RGB to the Hue-Saturation-Value (HSV) representation. The algorithm then scans the rows, searching for the row that has, on average, the greatest number of pixels closest to the brown hue used. Once identified this the spectrum on the alternate side is removed (e.g. if the spectrum was acquired in the apical 4-chamber position, the

data above the line is removed). Again, using the HSV representation the green ECG is removed, along with other labels detected based on their colour.

The image is then converted to single greyscale channel consisting of the sanitised spectral data.

4.3.1.6 Thresholding

Tracing the image requires selecting an appropriate threshold to differentiate the foreground pixels from the background noise. Depending on the background noise level, gain settings, and the quality of the image the optimal threshold will vary from image-to-image.

Global thresholding typically involves transforming a greyscale image (for example Figure 4-4) into histogram where the number of pixels of each level of grey. A threshold is then selected turning the greyscale image into black and white.

There are multiple existing methods for selecting a threshold. One very common technique is "Otsu's method" (Otsu 1979). This algorithm assumes that the image contains two classes of pixels following a bimodal distribution. It then attempts for find a threshold that maximises the inter-class variance. With our Doppler data this method often failed when the outer profile was faint compared to the inner profile; the threshold selected often led to clipping of the peaks (Figure 4-6).

In total 16 different existing algorithms were screened over a sample of traces. For the example shown in Figure 4-4, applying the 16 algorithms, produces the 16 respective outputs shown in Figure 4-7.

Similarly, the LVOT trace shown in Figure 4-8 when processed by the 16 algorithms produces the 16 respective outputs shown in Figure 4-9.

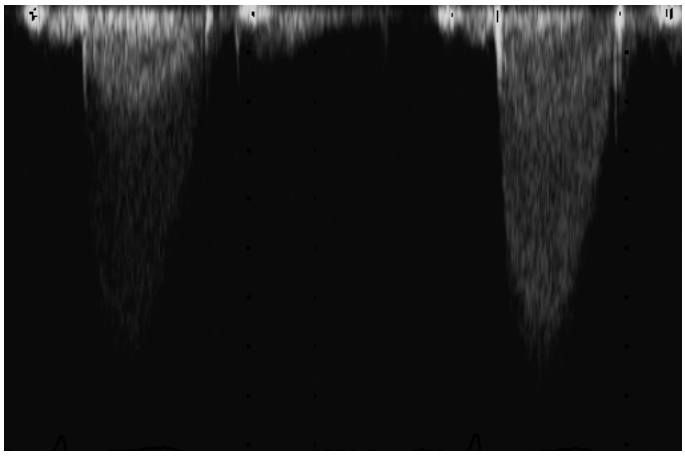


Figure 4-4 An example of a challenging continuous wave Doppler image for thresholding

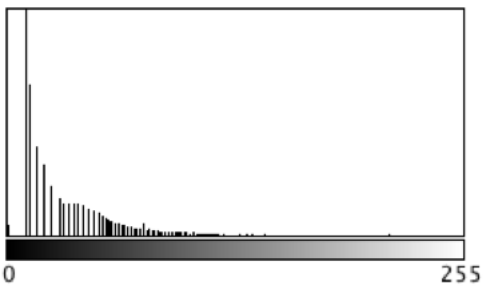


Figure 4-5 Histogram of above figure

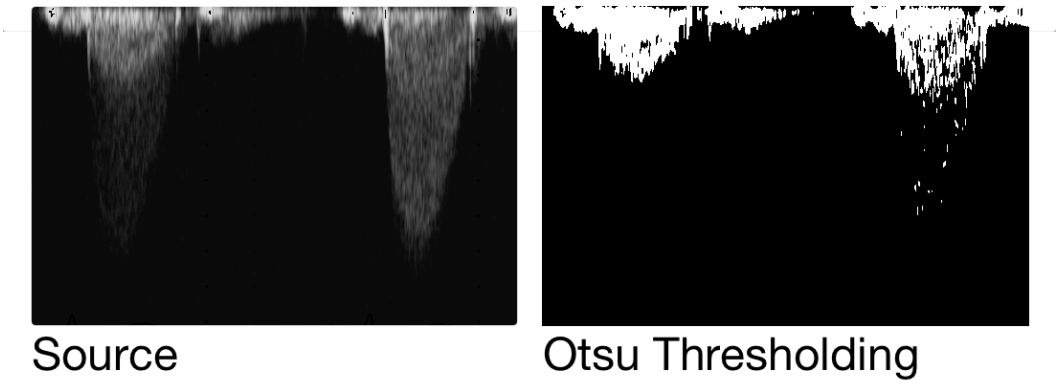


Figure 4-6 Thresholding using Otsu's method fails when the outer trace is faint

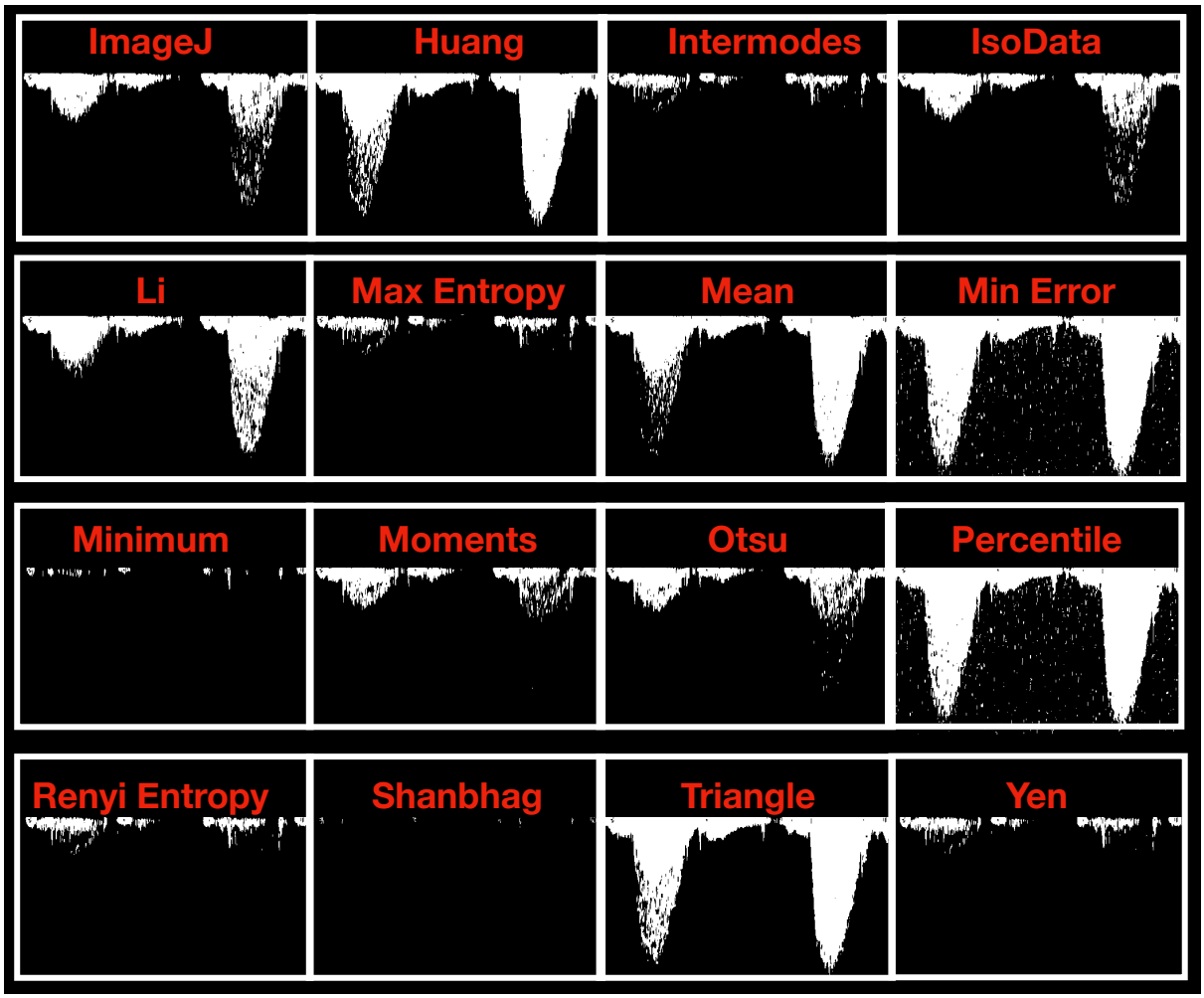


Figure 4-7 Results of 16 different threshold algorithms on the example continuous wave Doppler image

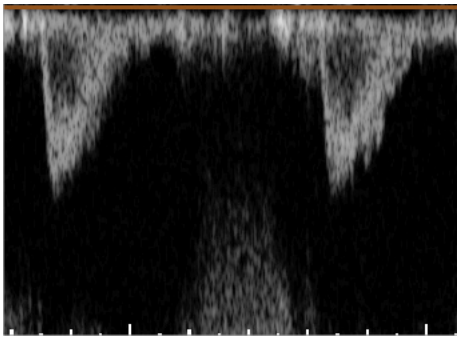


Figure 4-8 An example pulsed-wave Doppler image

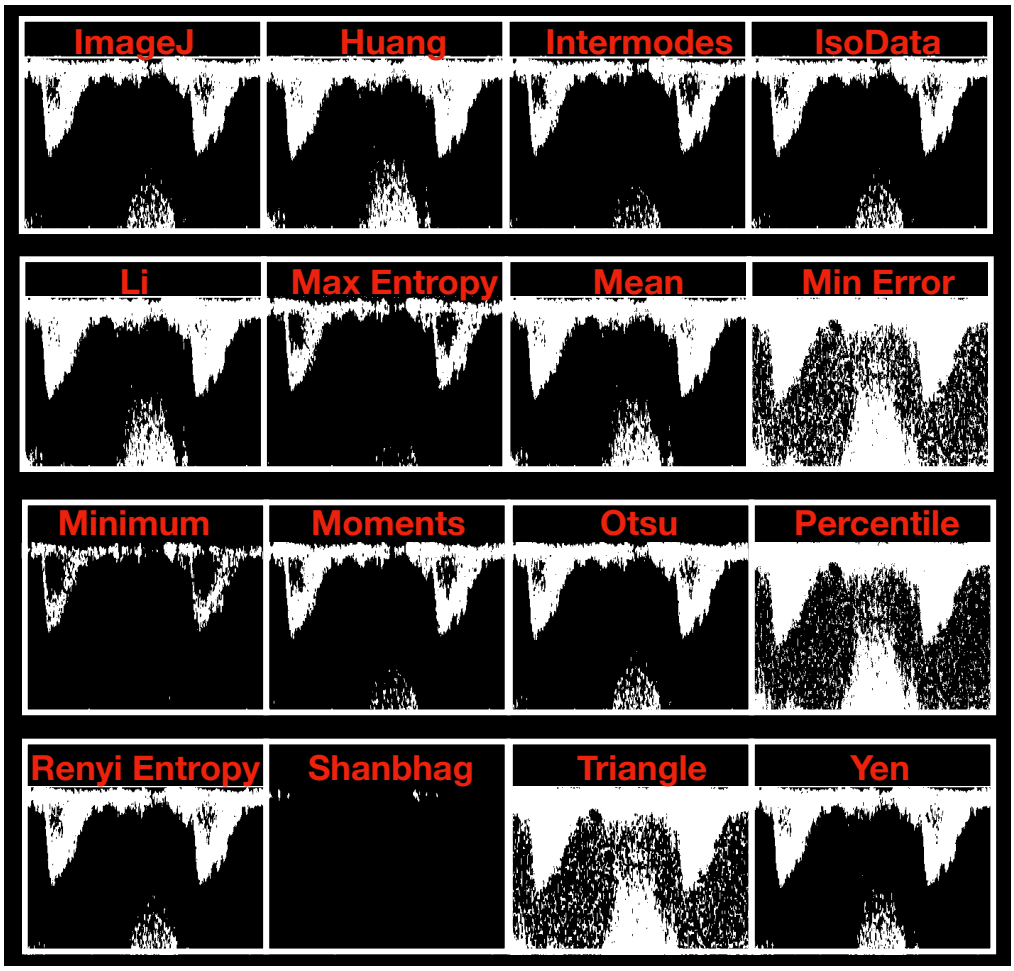


Figure 4-9 Results of 16 different threshold algorithms on the example pulsed-wave Doppler image

For the challenging continuous-wave Doppler example, most (13/16) of the algorithms caused at least one of the beats to be truncated. Even the three algorithms in which this did not occur (Min Error, Percentile, Triangle) failed on the noisy LVOT pulsed-wave example because they allowed too much background noise to pass.

I therefore developed my own algorithm based on domain specific knowledge (*Figure 4-10*). After thresholding, the image is then traced (described in Section 4.3.1.7). If the threshold is set too low (example in *Figure 4-10*, threshold = 10), then the trace of the spectrum will tend to be too shallow in the valleys between the peaks. Alternatively, when the threshold is set too high (example in *Figure 4-10*, threshold = 50), it will tend to clip off the peaks. Therefore, when the threshold is at the optimum, it would be expected that the amplitude of the oscillation of the trace at the frequency of the cardiac cycle will be at its maximum.

To find the optimum threshold the algorithm therefore iterates through all threshold in a range that would be expected to encompass reasonable choices (such as 5 to 50), binarizes the image, and traces the outline. It then performs a Fourier transform of the tracing (*Figure 4-10*, lower left panel) and searches for the highest peak within an appropriate frequency range (e.g. between 0.5Hz and 2.5Hz which corresponds to heart rates of 30bpm to 150bpm). A graph of the intensity of this peak against the tested threshold is generated (*Figure 4-10*, lower right panel), and the lowest

threshold that generates the highest intensity on the Fourier transform is selected as the optimum threshold.

This technique prevents the inadvertent clipping of the outer profile due to a too high threshold. This is illustrated in Figure 4-11 for the challenging example of Figure 4-4.

Instead of testing every threshold within a range, a more efficient search strategy could be used so that not all thresholds would need to be tested. However, as the process is quick, this was not found to be necessary.

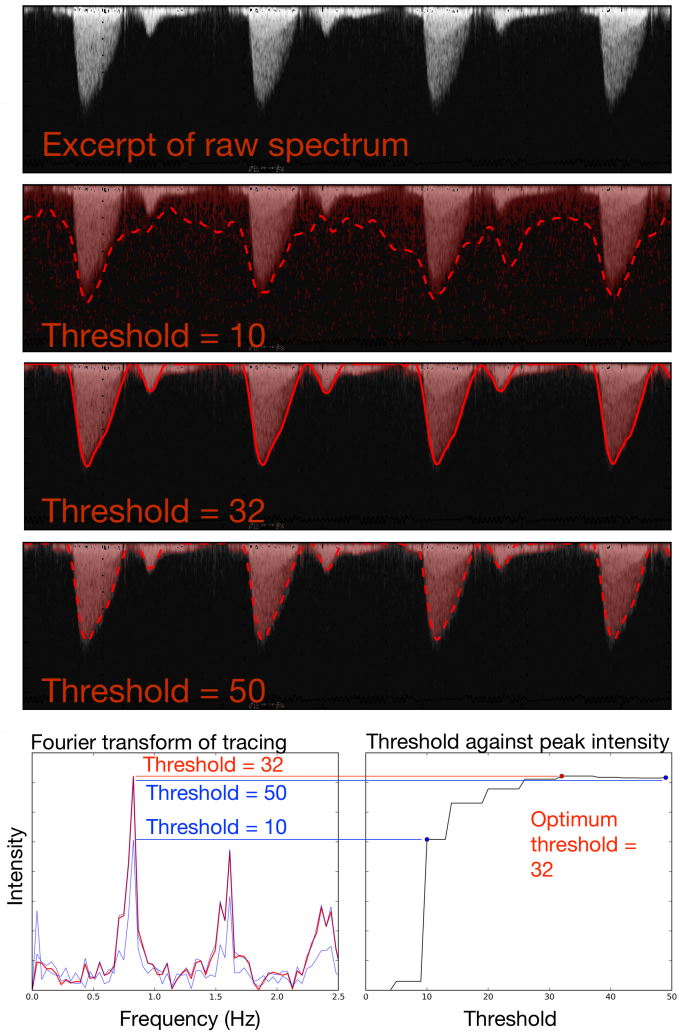


Figure 4-10 Identification of the optimum threshold to convert the grayscale image to a binary image for subsequent tracing

Selecting the optimum threshold to convert the grayscale image to a binary image relies on optimising the selected threshold to maximise the amplitude of the oscillation of the subsequent tracing within the plausible frequencies of heart rates (0.5Hz to 2.5Hz). Two example thresholds that are not the optimum, 10 and 50 are presented above, along with the calculated optimum threshold (32). When the threshold is 10, some of the background noise is classified as the foreground (red overlay). This causes the tracing to be too "shallow", and this is reflected by the lower peak of the Fourier transform of the tracing. At a higher threshold of 50, some of the foreground is classified as background, and a small portion of the peak is clipped off. Again, this is reflected in a reduced peak intensity of the Fourier transform. All thresholds between 5 and 50 are tested, with 32 producing the highest peak and hence was selected as the optimum for this example.

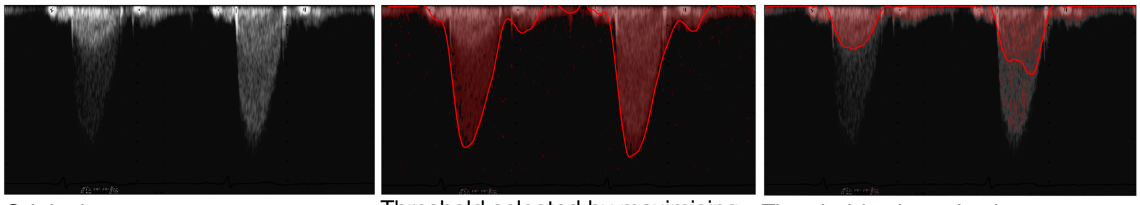


Figure 4-11 Improved thresholding performance of described method as compared to Otsu's

The raw spectrum (left panel) is a challenging example for thresholding algorithms. Otsu's method (right panel) classified much of the faint profile as background and hence the peaks are substantially clipped (red trace). My method, which aims to maximise the amplitude the trace accurately picks an appropriate threshold (middle panel).

4.3.1.7 Tracing

After selecting an appropriate threshold and subsequent binarization image is further sanitised by removing small connected clusters of less than 1000 pixels in size.

For each column of the image the longest contiguous segment of background is found of which the top point is marked as the raw trace (Figure 8, blue trace). This trace then undergoes a low-pass filter at 10Hz (Figure 8, yellow trace) to remove the high frequency noise present in the tracing. The inter-beat interval of the trace is estimated using autocorrelation. The filtered trace then is smoothed using a Savitzky-Golay filter (Savitzky & Golay 1964), (5^{th} order, window of 1/6 of the estimated inter-beat interval, Figure 8, red trace).

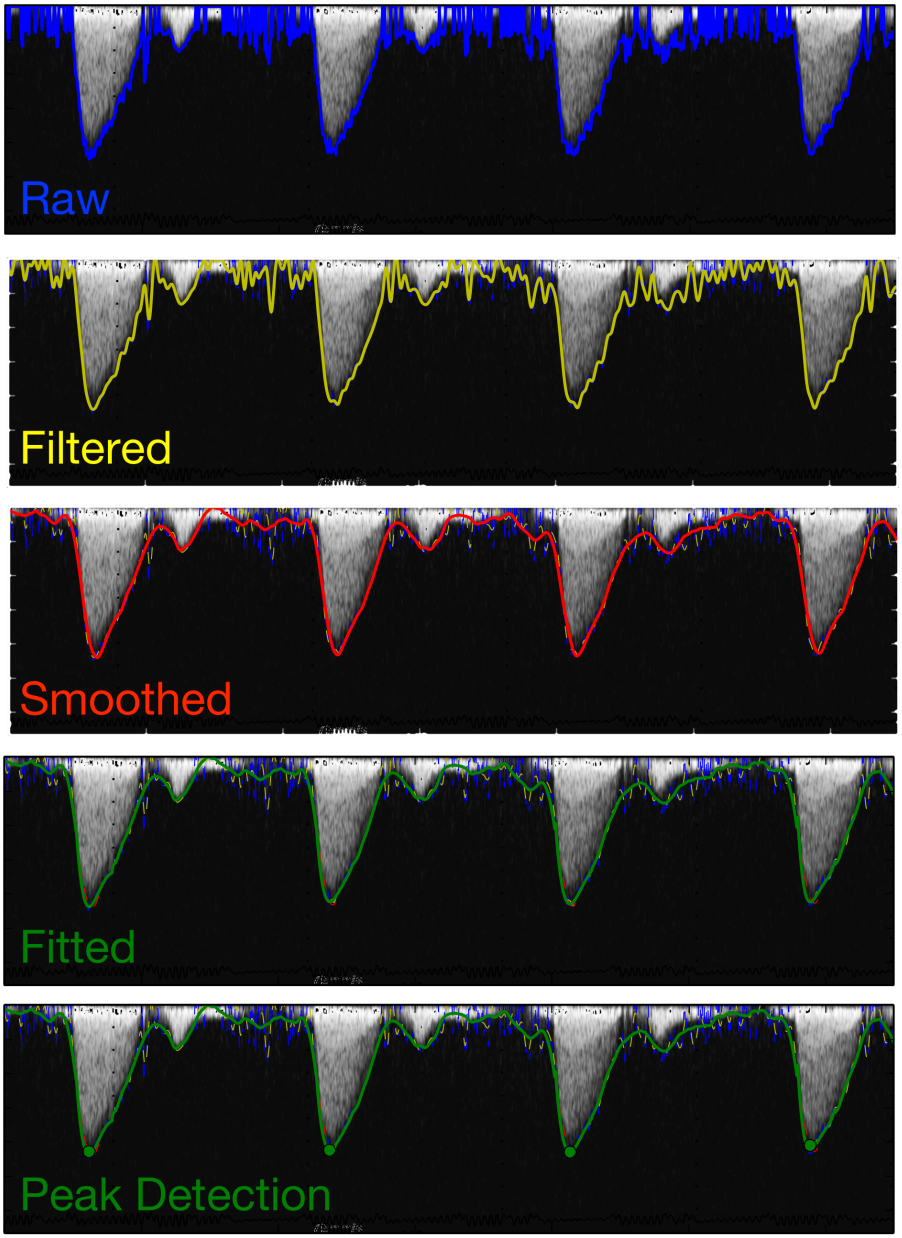


Figure 4-12 Stages of tracing and peak detection

After thresholding, the image is binarized and then traced by searching in each column for the greatest continuous segment of background. The pixel before this is then selected. This tracing is then filtered using a low-pass Butterworth filter before smoothed using a Savitzky-Golay filter. The tracing around the peaks is then fitted to a second-order Gaussian curve, and the peaks are identified.

4.3.1.8 Distinguishing genuine peaks from pseudo-peaks using the wavelet transform

Real-life traces include miscellany of small bursts of doppler of low velocity, which can easily be mistaken for peaks, by an algorithm that simply looks for values greater those in the two adjacent regions.

I developed an algorithm that strongly favours the real peaks due to aortic outflow and suppresses the distracting artificial peaks. The key step is performing a continuous wavelet transform (Du et al. 2006). In this, the wavelet algorithm quantifies the peakedness of every instant along the doppler strip. Since peaks can have different durations the wavelet transform looks for variety of possible scales of peak. The artificial peaks are always smaller than the genuine peaks and become suppressed by the genuine peaks during the wavelet transform process.

In Figure 4-13 the wavelet transform is shown in colour, with red indicating the location of peaks as ascertained by the wavelet transform. As can be seen, the strength of the effect of the genuine peaks is sufficiently large to prevent the artificial peaks producing visible disturbance on the wavelet transform.

4.3.1.8.1 Counteracting the template like behaviour of the wavelet transform

One disadvantage of using the wavelet transform is that it is in effect a form of template matching, and in this implementation is using a symmetrical wavelet, even though Doppler waveforms are asymmetrical, typically peaking earlier than the mid-point of the ejection phase.

I counteracted this weakness by refining these detected peaks to their local maxima.

Often the tip of the peak is slightly indistinct, due to additional noise, or a loss of signal. At this stage of the algorithm we want to ensure that this part of the trace is included. The algorithm therefore selects a portion of the trace around the peak, starting and finishing where the velocity is two-thirds of the peak, fit a second order Gaussian to the trace, before taking the new maximum as the final estimate of the height of the peak.

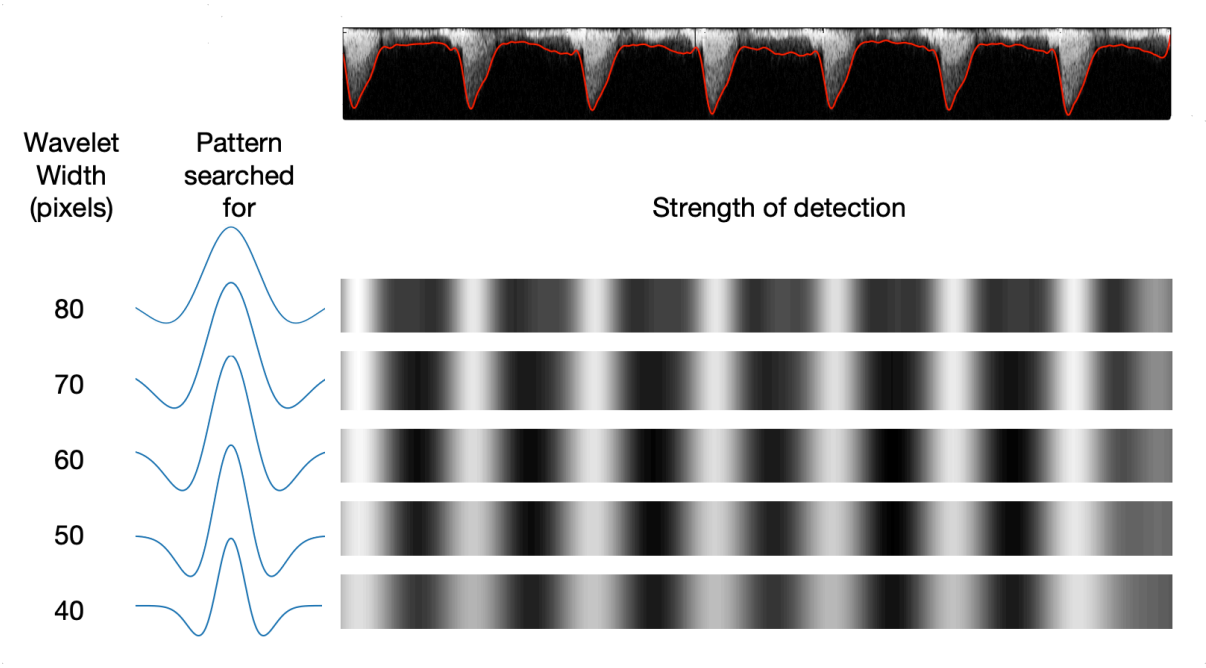


Figure 4-13 Detection of region of peak in each beat using continuous wavelet transform.

The smoothed trace (top right) is convolved with a Ricker wavelet of various widths (left). Correlation plots, indicating the 'strength of detection' are plotted below the spectrum. The ridges within this are then identified and filtered to produce the detected peaks. Because the underlying wavelet is symmetrical but the aortic waveform slightly asymmetrical, the point picked at this stage in the algorithm is slightly offset from the true peak. However, this process is good at ensuring that trivial artefacts are never mistaken for a peak, and the mis-alignment is easily fixed by seeking the nearest local maximum.

4.3.1.9 Automated quality control to reject unreliable beats

During long acquisitions the heart may move in and out of the imaging plane lowering the quality of the spectrum. The algorithm removes beats with a height lower than a multiple of the median absolute deviation from the median.

4.3.1.10 Pre-processing and ensembling

The individual beats are extracted and aligned in the horizontal direction based on their peaks. The ratio of LVOT and AV component around the peak is relatively stable between beats, even when different beats differ in height. I use this property advantageously by stretching the beats vertically so that the outer envelope (aortic velocity) is concordant between different beats. In principle, the inner envelope should now be concordant between the beats.

The individual beats are therefore vertically stretched so that all the peaks are of the same height. Then are then ensembled (merged) to form a final image. The resultant image has a significantly improved signal to noise ratio (Figure 4-14). This makes the inner and outer profile easier to distinguish.

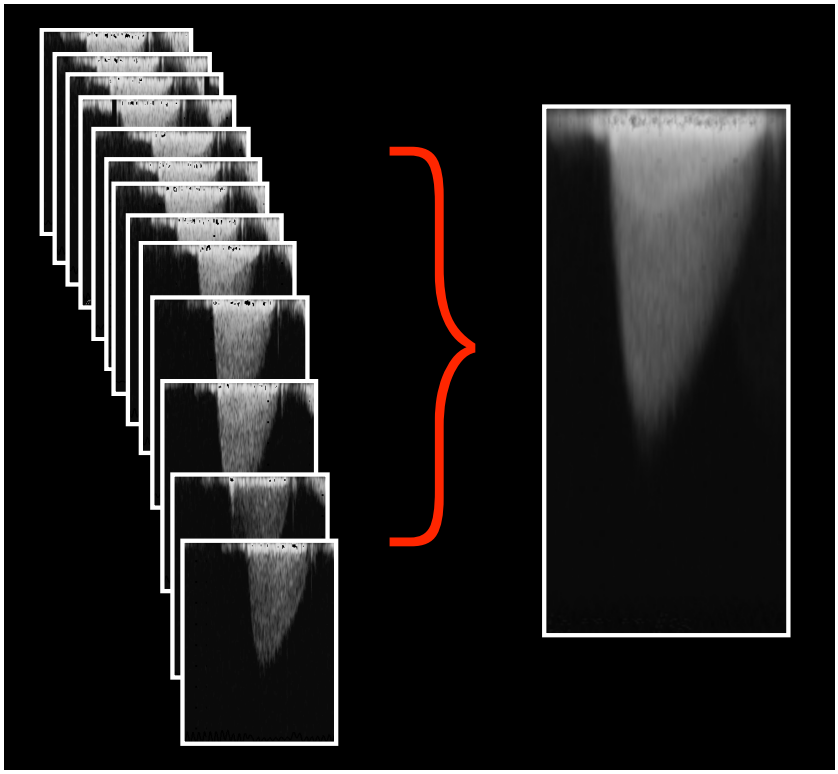


Figure 4-14 Beat averaging to produce an ensembled beat.

After detecting the individual beats from the Doppler trace, each beat is isolated, centred, and scaled to a uniform height. They are the averaged to produce an ensembled beat.

4.3.1.11 Special processing for the non-normal distribution of Doppler noise.

When merging the segmented a simple average (mean) of the data is used. However, the noise of ultrasound – both 2d and spectral Doppler – is not normally distributed. By eye, one can appreciate this as within the spectral doppler profile, where one would, without noise, would expect to see bright white, there are very dark pixels giving it a speckled appearance. The reflected ultrasound signal, when transformed into an electrical RF signal has noise that is approximately normally distributed. Through processing the signal is transformed and the of the noise approximately becomes Rayleigh distributed (Seabra & Sanches 2008). Finally, as the dynamic range of ultrasound images tends to be greater than the display, the range of values are compressed by log scaling them and the distribution of the noise approximates the Fisher-Tippett distribution (Slabaugh et al. 2009). In merging multiple beats the aim is to reduce noise though taking the average (mean). To maximise the information obtained from the data, the images are transformed so that it is normally distributed, the mean taken, and transformed back to the standard distribution.

4.3.1.12 Extraction of the inner and outer profile to calculate the instantaneous dimensionless index

Just as the original spectrum was thresholded, traced, and the peaks identified; so is the ensembled beat.

The difference between the two episodes of thresholding, tracing, and peak finding is that the first time it was only to cut the beats so that they could be assembled together, while now the purpose is to actually make the measurement of iDI.

To do this, the image is processed to identify potential ridges that signify the outer and inner envelopes. For both of these envelopes, just within the envelope, the Doppler intensity if higher than outside the envelope. Therefore, the step change in Doppler intensity is the indicator that the envelope is here.

My algorithm starts by calculating what is in effect the rate of change of doppler intensity at each vertical position on the doppler strip of the ensembled beat. A special transformation is applied so that adjacent vertical scanlines on the ensembled beat can be handled together, having accounted for the fact that away from the peak of the beat, the envelope is smaller.

As shown in Figure 4-15, the rate of change of Doppler intensity is plotted for each vertical position. The outer and inner envelopes are always prominent peaks within this rate-of-change map.

Naively, one could simply calculate the mean of the pixel values in the upper box and subtract it from the lower box as a measure of the differences between the two boxes. As the pixels in the boxes are not normally distributed, the Kullback-Leibler divergence, which is a measure of how one distribution differs from another, is used. The original, non-negative symmetric form is used (Kullback & Leibler 1951).

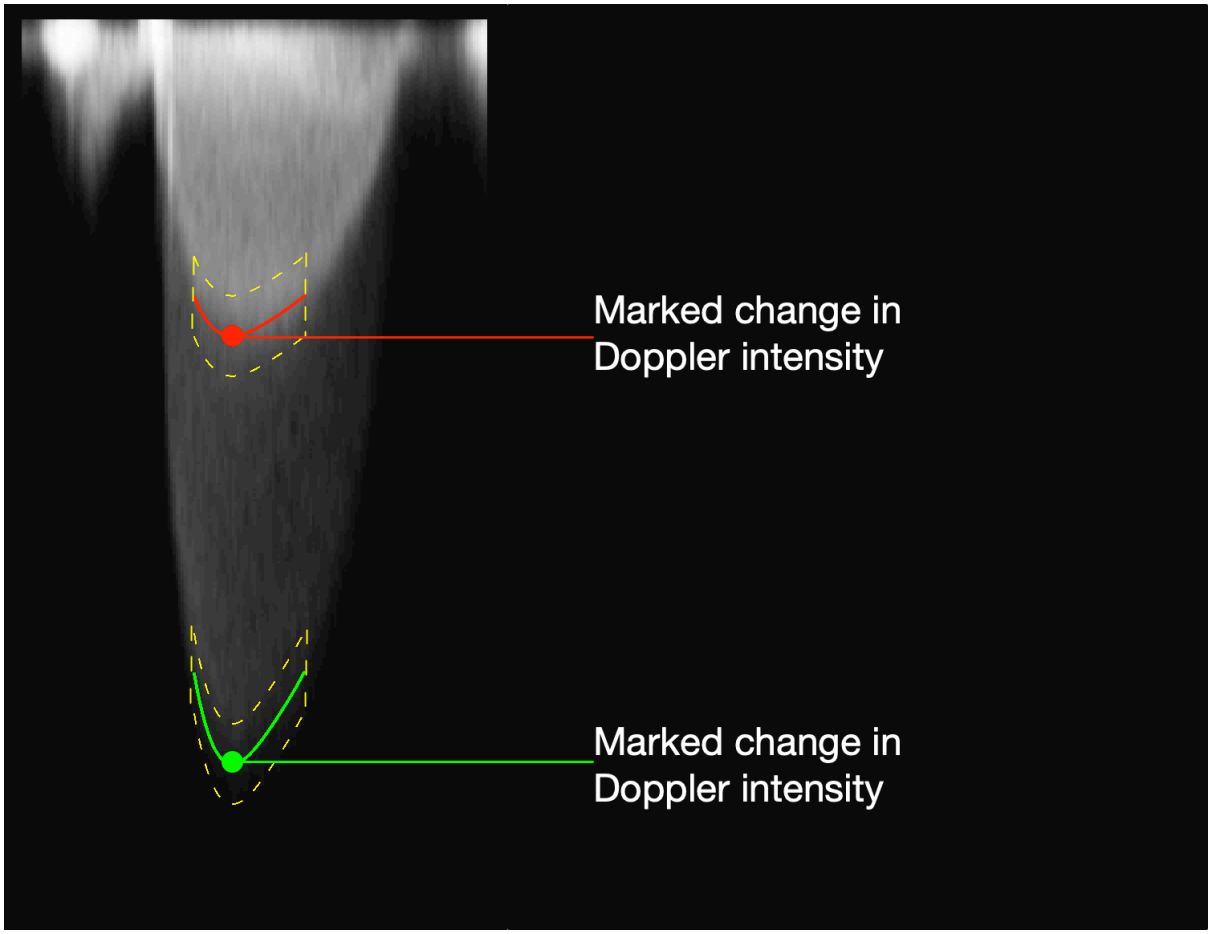


Figure 4-15 Detection of the inner and outer profile from the ensembled beat.

The ensembled beat undergoes automatic thresholding and tracing. A sliding window (yellow dotted lines) scans the ensembled beat from the peak to the base. Places where there is a marked change in the intensity of the pixels from upper to the lower portion of the box is detected. The two most intense changes correspond to the true outer and inner peak. The ratio of these is the dimensionless index.

4.3.1.13 Worked example uses of the instantaneous Dimensionless Index

As an example (Figure 4-16), I acquired 7 continuous wave recordings from a single patient. This produced 7 different values for the instantaneous dimensionless index. The coefficient of variation of these values reflects the short-term biological and probe placement variability. In the example patient this coefficient of variation was 4.5%.

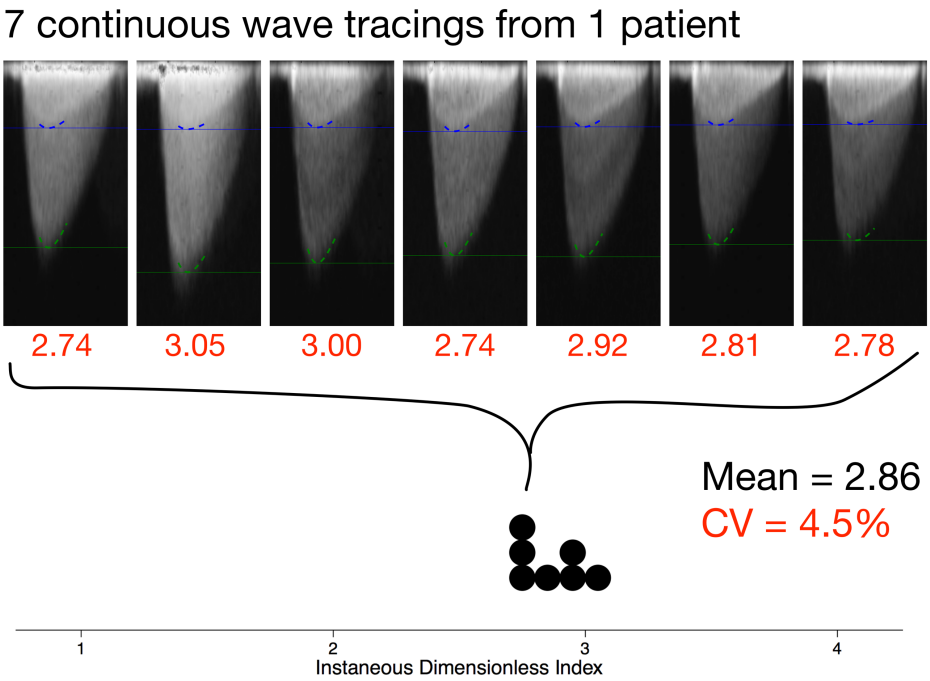


Figure 4-16 The instantaneous dimensionless index readings from 7 continuous wave Doppler recordings in one patient

After each continuous wave Doppler recording undergoes processing, and the inner and outer profile are detected, their ratio is calculated. These results are then visualised on a dot histogram where each dot represents a separate recording. The mean and coefficient of variation (CV) is calculated. In this patient the CV is 4.5%

4.3.1.14 Integration into user interface

I packaged the algorithms I developed into an easy to install and use piece of software that enables researchers and clinicians to view a folder of DICOMS (Figure 4-17), process them extract the full Doppler spectrum with peak identification and quality control (Figure 4-18), and then to form the ensembled beat to calculate the instantaneous Dimensionless Index (

Figure 4-19) and view the intensity plot (

Figure 4-20).

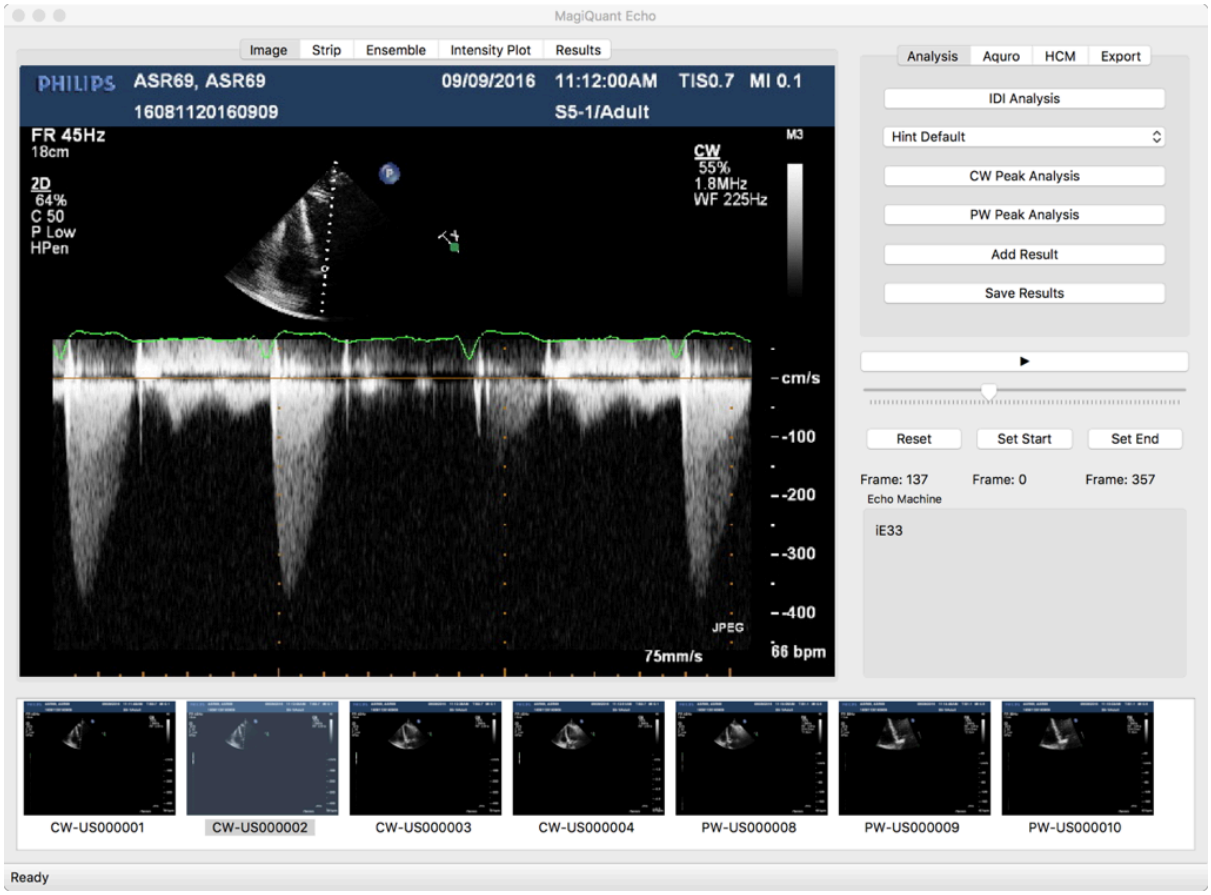


Figure 4-17 Viewing a folder of DICOM files

A folder of DICOM files is opened, with thumbnails on the bottom. Individual files can be selected and views. Processing is initiated by pressing the “IDI Analysis” button.

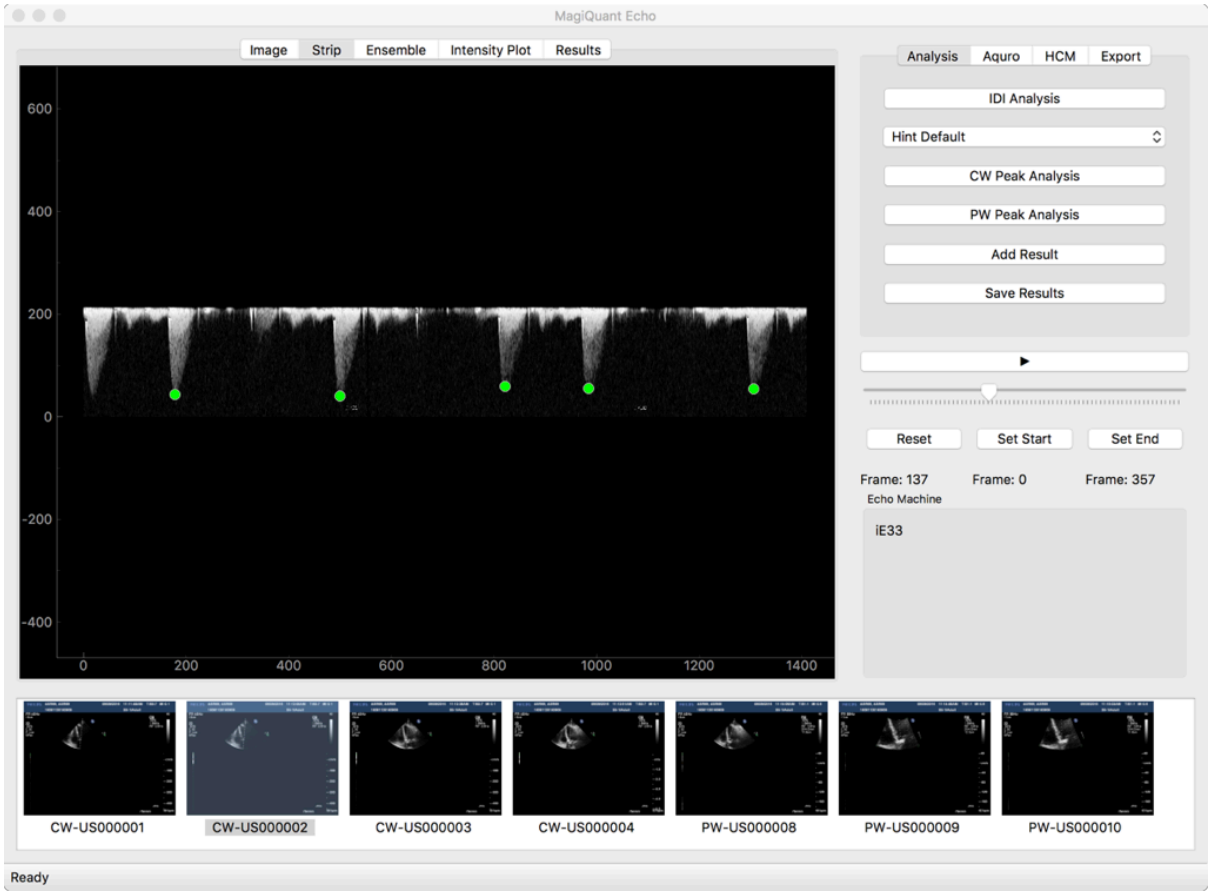


Figure 4-18 Extraction of the strip, peak identification, and quality control

The Doppler spectrum strip is extracted and displayed with the peaks marked on the beats used for ensembling. The first partial beat (as it is truncated) does not pass quality control and is discarded.

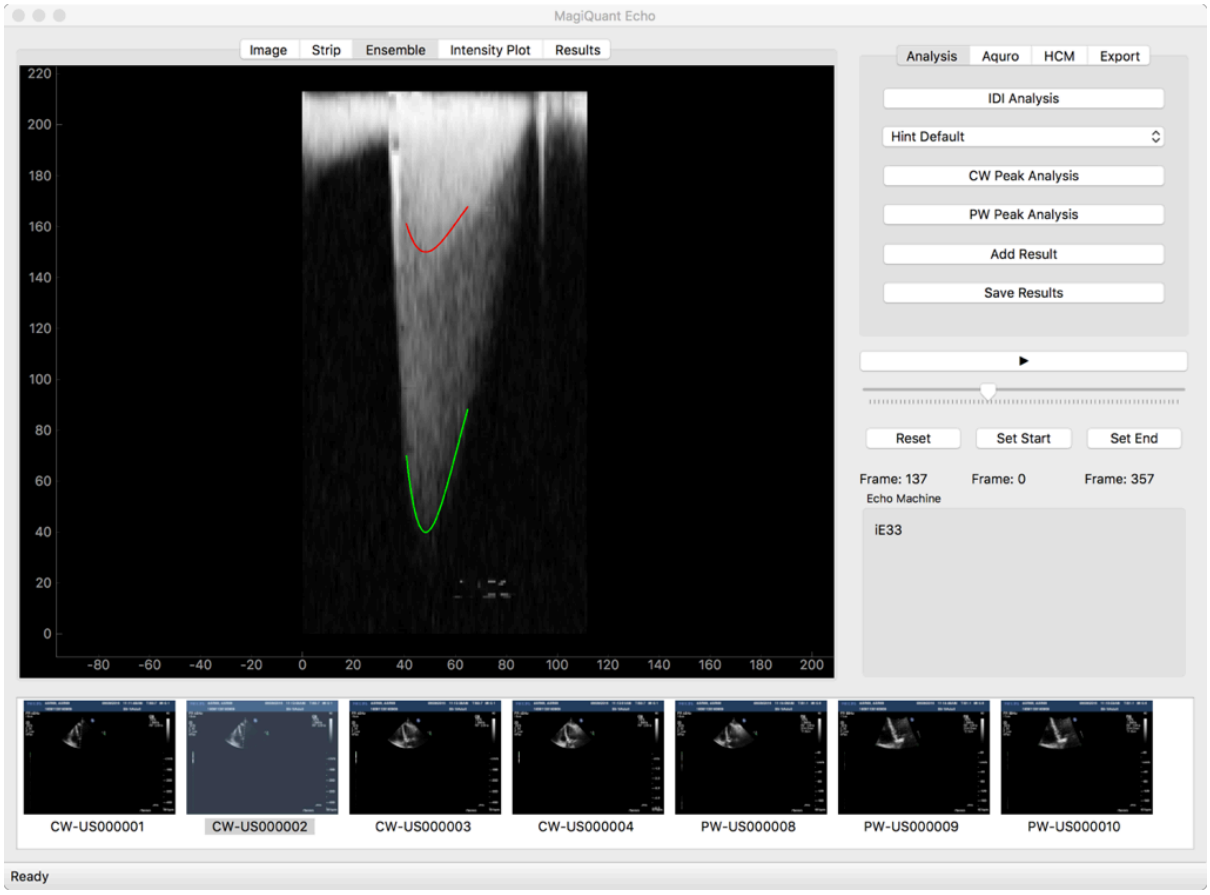


Figure 4-19 Viewing the ensembled beat with the inner and outer profile

The inner and outer profile are calculated and displayed allowing the operator to confirm that they have been correctly detected.

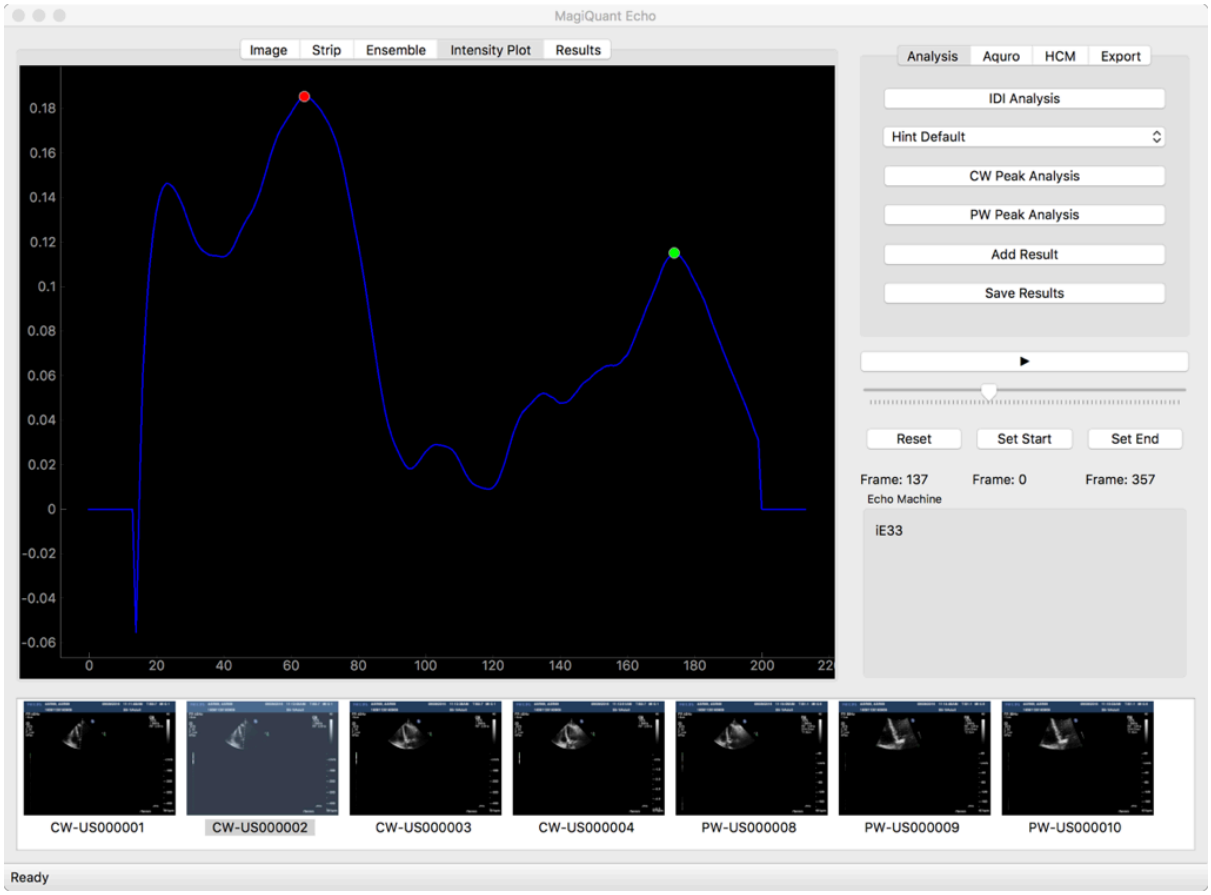


Figure 4-20 The intensity map

The intensity map with the inner and outer peaks is displayed allowing the operator to check correct detection and appreciate how confident the algorithm is in detecting the inner profile.

4.3.2 Study population

4.3.2.1 Patient Group A (Imperial College NHS Healthcare Trust, London)

For the initial validation study, patients at a single institution (Imperial College NHS Trust, London) with mild, moderate, or severe aortic stenosis in a native or bioprosthetic valve were recruited. Patients with mechanical valves were excluded. All participants gave written, informed consent to participate in this study which was approved by a local ethics committee (14/LO/1520). Aside from inability to give consent, there were no exclusion criteria for the initial validation study.

4.3.2.2 Patient Group B (Barts Health NHS Trust, London)

An independent clinical researcher (Dr. Sveeta Badiani) acquired images for iDI calculation from patients being enrolled in her Trust for a separate study of aortic stenosis. Those patients gave informed consent for aortic valve imaging within that study (17/LO/0929)

4.3.3 Doppler echocardiography

In Patient Group A, echocardiographic data were acquired using a Philips IE33, Philips Affinity 70, or Philips Epic 7. Data were processed off-line. Doppler interrogation of the aortic valve was achieved using the standard apical 5-chamber view (Evangelista et al. 2008).

4.3.4 Data acquisition

Recruited patients underwent a standard echocardiographic assessment. Data for this study was then specially acquired.

For each patient, multiple continuous wave Doppler strips, and multiple pulsed wave Doppler strips was acquired.

Each Doppler strip contained multiple beats: many more beats than would be captured for a standard clinical echocardiogram. At least 3 beats worth, and up to 10 seconds worth of continuous wave data was acquired. Similarly, a pulsed wave Doppler strip of multiple beats was obtained. The probe was then removed from the chest wall and then replaced before collecting another set of multiple beat continuous wave and pulsed wave Doppler strips. After each acquisition, the probe was removed and replaced on the chest wall. For each patient, at least three aortic and LVOT runs were obtained for analysis.

4.3.5 Automated image analysis

The LVOT PW and trans-aortic CW Doppler acquisitions underwent automated thresholding, tracing, peak-detection, and quality control as described above.

The peak velocity of each beat was recorded for further analysis.

As a reminder:

- Each patient had several CW and PW strips;
- Each strip had several beats.
- Each beat had its own peak measurement.

Therefore, each strip had its own average value of the peak measurements of its beats, and each patient their own average value across all CW strips, and separate average value across all PW strips.

I then calculated the conventional DI using two different approaches, and also the iDI.

(1) **"Typical DI":** The first method of calculating conventional DI is designed to be as similar as possible to standard clinical practice. In standard clinical practice the echocardiographer picks one beat for CW and one beat for PW and calculates the ratio. In my experimental protocol, there are many beats, for example 50 beats of CW and 50 beats of PW. There are therefore, 50x50=2500 possible combinations of single beat CW and single beat PW. Alternatively, I could have only combined the first beat of CW with the first beat of PW, and then the second with the second, and so on. Mathematically, both approaches have the same expected value, and the same standard deviation. However, the 50x50 approach is statistically more stable and therefore preferable for making a research comparison between methods.

(2) **"Idealised DI":** The second method of calculating DI is designed to achieve an artificially low amount of noise in the CW and PW measurements. For this I take the average CW velocity for a whole strip of beats, and the average PW velocity for a whole strip of beats. Because those average numbers have a

smaller percentage variability than the individual beat values, the CW/PW ratios thus calculated also have a smaller percentage variability. The reason for incorporating this second method it to give the conventional DI the best opportunity to shine when being compared to iDI.

4.3.6 Calculation of iDI

iDI was calculated using the same CW strips that were used for conventional (*typical* or *idealised*) DI.

4.3.7 Data analysis

In general, two different measures (such as iDI and DI, or inner profile vs LVOT PW) mathematically cannot match (or correlate) with each other better than they can match themselves. Therefore, to get into the position of finding equivalence between two measures, it is advisable to first maximise the precision of each measure. The simplest method of increasing precision is to make multiple acquisitions and calculate the mean, thus minimising the impact of noise.

This was the reason that I designed my protocol to acquire multiple traces for traditional DI and iDI.

4.3.7.1 Feasibility and data exclusion.

From each patient multiple CW and PW recordings were made. If there was enough data to calculate the traditional dimensionless index then that patient was included in the cohort. If the iDI was able to be calculated from at least one tracing, it was recording as being feasible in that patient, if not it was recorded as being not-

feasible. Patients were only excluded if they did not have data that was considered useable for the calculation of the traditional dimensionless index.

4.3.7.2 Does the iDI algorithm extract from the CW an inner profile that is suitably equivalent to the PW?

While the outer envelope used by the iDI is exactly the envelope used in clinical practice, there is no standard way of interpreting the inner envelope on a CW. Therefore, I decided to use the separately acquired PW as a gold standard. The challenge with this is that the PW itself is subjected to noise, and nothing can agree with it better than it can agree with itself.

I therefore tested the level of agreement:

a) Between one PW peak velocity and another PW peak velocity.

b) Between the mean PW peak velocity and the inner profile.

c) Between one inner profile and another inner profile.

For comparisons of things with themselves, i.e. (a) and (c), the only relevant aspect of agreement is the width of the 95% limits of agreement.

For the comparison between things, i.e. (b), both the width and the position of the 95% limits of agreement are relevant, since there is a possibility of bias between the things.

Data were visualised using a Bland-Altman (Bland & Altman 1986) and associated correlation plots.

4.3.7.3 Validation of the equivalence of the iDI and DI

Just as the PW and the inner profile peak velocity might be different not because of consistent differences between the two, but because each one is a noisy variable, the same can be said for DI and iDI. I therefore tested the level of agreement:

a) Between one typical DI and another.

b) Between the mean DI and the mean iDI.

c) Between one iDI and another.

As in Section 4.3.7.3 the 95% limits of agreement and mean bias were calculated, and Bland-Altman and correlations plots were made.

4.3.7.4 Comparative precision of iDI and DI

The precision of the DI and iDI, expressed as the coefficient of variation, was calculated for each patient. The coefficient of variation (CV) was calculated for each patient using the method of Bland (Bland 2006).

The noise in these echo measurements tends to be approximately proportional to the underlying value (the mean). For example, if CW values are varying between 2 and 2.2 m/s, then under situations of increased stroke volume, they would likely vary

between 4 and 4.4 m/s. Therefore, the geometric rather than the arithmetic standard deviation is used. The process is therefore as follows:

The natural logarithm was taken of each value, before the standard deviation calculated, and then the exponential taken. From this 1 was subtracted and then was multiplied by 100% as follows:

$$Coefficient\ Variation = \left(e^{sd(ln(value))} - 1\right) \times 100\%$$

Three coefficients of variation were calculated.

- First, the one for conventional DI in typical clinical practice, which employs one CW beat and one PW beat.
- Second, the one for conventional DI in theoretically idealised clinical practice, which employs a CW velocity calculated from the average of multiple beats, and likewise a PW velocity calculated from an average of multiple beats.
- Third, the one for the iDI.

For illustration of this calculation a worked example of a few acquisitions from a single patient will be presented. Furthermore, the raw data for each patient will also be presented as a dot-plot is summary statistics.

The difference between the coefficients of variations were calculated using a paired Mann-Whitney U test.

For comparison, the coefficient of variation was calculated for the peak CW, LVOT, and inner profile velocity on an acquisition-by-acquisition basis.

4.3.7.5 Clinical implication of the multiple repetitions

All three types of coefficient of variation I described above are based on a single CW acquisition (and for traditional DI a single PW acquisition), although this single acquisition could be a single beat or a single strip of beats.

The next question is whether there is an advantage to making more than one separate acquisition, of a beat or a strip of beats. By two separate acquisitions I mean one recording, and then ending the recording, lifting the probe off, and placing it back on the chest, and then making a second recording. The relevance of this process is that many very small differences in probe position will occur. These were already present within the first recording, and doing the process afresh allows another change of having these very small differences in another direction that might counteract any bias in the first.

Doing these extra steps would take time in clinical or research practice. Therefore, I wanted to test how much benefit they provide.

My approach to this was to actually acquire many strips of CW and many strips of PW, from each patient. I could then programme my software to combine two CW strips chosen at random from that patient and do similarly for PW. This would give

me an estimate of how much smaller the noise was when combining two independent strips.

I could then do this for 3 strips, 4 strips, etc.

4.3.7.6 Visit-to-visit coefficient of variation

The overall visit-to-visit coefficient of variation was calculated for the DI and iDI using a random-effects model and the natural logarithm of the DI and iDI. The results were transformed to percentages using the method described in Section 4.3.7.4.

4.3.8 Statistical analysis

Data was processed using Python (Python Software Foundation 2017), and statistical analysis using the statistical environment R (R Core Team 2016) with the "tidyverse" collection of packages (Wickham 2017).

4.4 Results

4.4.1 Patient Group A (Imperial College NHS Healthcare Trust, London)

4.4.1.1 Patient population

93 patients were recruited and had useable data for the calculation of the traditional dimensionless index. Of these patients, iDI was feasible in 90 of the 93 (97%).

The mean peak velocity was 3.17 m/s (SD 0.73, range 1.86 to 4.97), and the mean LVOT PW velocity was 1.23 m/s (SD 0.35, range 0.62 to 2.43) resulting in a mean DI of 2.67 (SD 0.71, range 1.14 to 5.04) and iDI of 2.58 (SD 0.65, range 1.58 to 5.58). All patients were in sinus rhythm.

Of these, 25 agreed to return for a second assessment.

4.4.1.2 Validation of the equivalence of inner profile and LVOT PW

There was no clinically or statistically significant bias between the LVOT peak velocity obtained conventionally from the PW trace, and that obtained by the iDI software from the inner profile of the trans-aortic CW trace (-0.018, 95% CI -0.050 to 0.015 m/s, p=0.28). Overall, there was good agreement between the LVOT velocity obtained from the PW and the inner profile of the CW trace (95% LoA -0.31 to 0.28 m/s) and a strong correlation (r=0.90, p<0.0001). In comparison, the limits of agreement between two individual beats of LVOT PW velocities were wider (95% LoA -0.31 to 0.35 m/s) with lower correlation (r=0.87, p<0.0001) and narrower for two inner profiles (95% LoA -0.21 to 0.19) with greater correlation (r=0.95, p<0.0001). Correlation and Bland-Altman plots are shown in Figure 4-21.

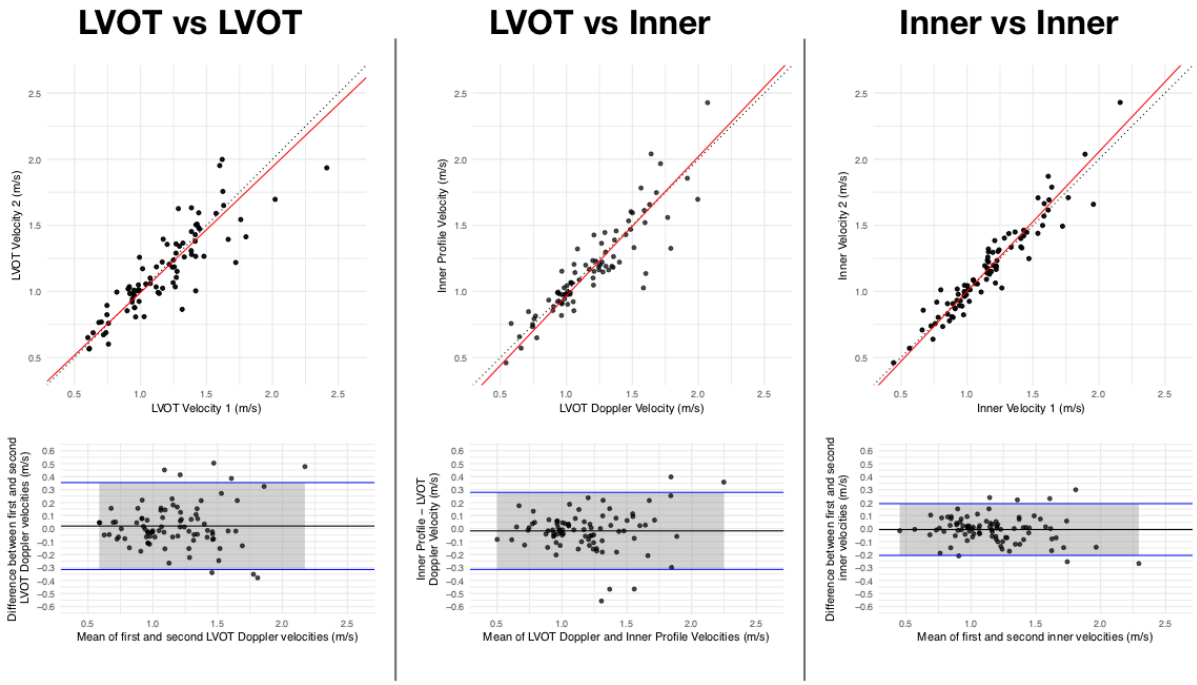


Figure 4-21 Comparison between PW of the LVOT and the inner profile

Comparison of the variability between replicate typical PW assessments of LVOT velocity (left panel), replicate assessments of LVOT velocity using the inner profile of CW (right panel), and between idealised PW and the inner profile (middle panel).

The LVOT PW data (left panel) show relatively poor reproducibility, which makes it impossible for any other variable to match LVOT PW closely.

The inner profile (right panel) shows better reproducibility.

The middle panel shows that the agreement between the two methods, is intermediate between the reproducibility of the two methods. This suggests that any apparent differences between the two measures can easily be easily explained by the inherent noise between each variable, and that the measures are not measuring different physical concepts.

4.4.1.3 Validation of the equivalence of the iDI and DI

There was clinically minimal, but statistically significant, bias between iDI and DI. iDI values were on average -0.08 units less than DI (95% CI -0.14 to -0.02, p=0.005). Overall, there was good agreement between iDI and DI (95% LoA -0.61 to 0.45) with a strong correlation (r=0.93, p<0.0001). In comparison, the limits of agreement between two DIs calculated from individual beats was larger (95% LoA -0.88 to 0.82) with a weaker correlation (r=0.83, p < 0.0001). The limits of agreement between two iDIs were narrower (95% LoA -0.41 to 0.35). The correlation and Bland-Altman plots are shown in Figure 4-22.

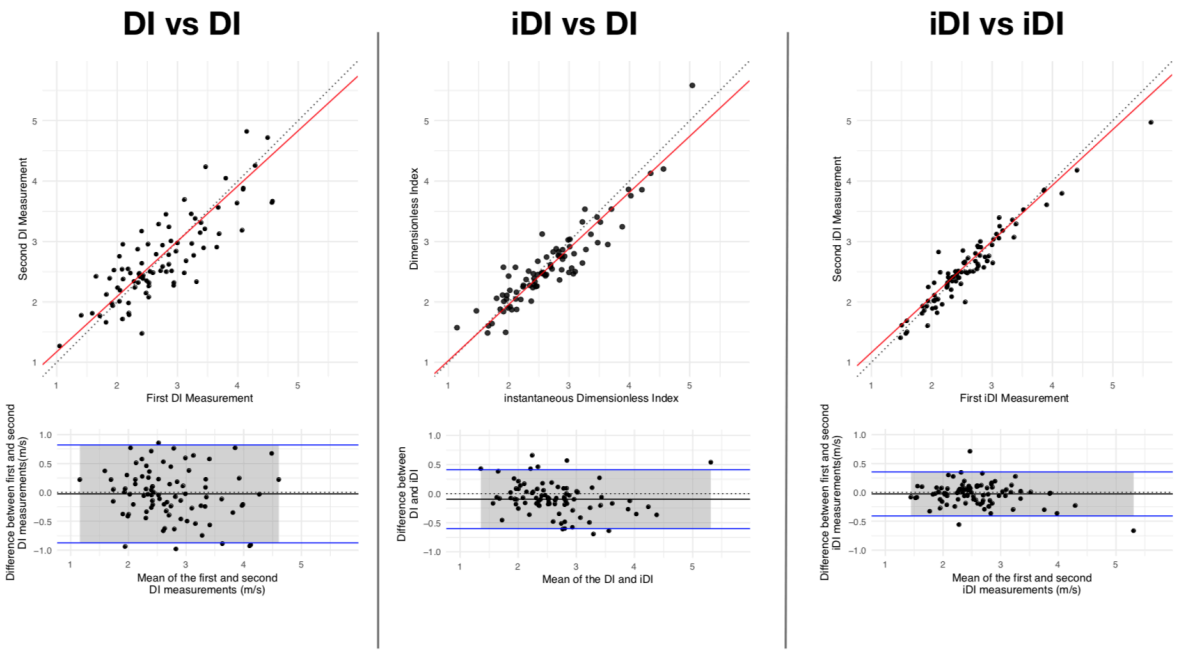


Figure 4-22 Comparison of the variability between replicate typical DI assessments (left panel), replicate iDI assessments (right panel), and between idealised DI and iDI (middle panel).

iDI versus *idealised* DI has limits of agreement (middle panel) that are intermediate in width between those of typical DI versus DI (left panel) and those of iDI versus iDI (right panel).

4.4.1.4 Comparative precision of instantaneous Dimensionless index and traditional Dimensionless Index: A worked example for an individual patient

The data of a single CW and PW acquisition of a single patient is shown in Figure 4-23. Each trace has been automatically thresholded, traced, and the peak velocity measured. The coefficient of variation of the beat-to-beat variability of the peak trans-aortic CW velocity over a single acquisition for this patient is 7.7%. For the LVOT PW in this single acquisition it is 4.6%. When each of the non-simultaneously measured CW and PW peak velocities are paired and the dimensionless index calculated the coefficient of variation of the dimensionless index is 8.2%.

The data from four continuous wave acquisitions from the same patient is shown in Figure 4-24. Each of the four continuous wave acquisitions of multiple-beats have been thresholded, traced, and merged to form the four ensembled beats shown. In each case the inner and outer profile has been extracted and their ratio taken to calculate the instantaneous Dimensionless Index. As the inner and outer profile are measured simultaneously, variability in these components cancels out when the ratio is calculated. The iDI thus has a much smaller coefficient of variation, which in this example is 1.7%, which is significantly less than that for the DI 8.2%.

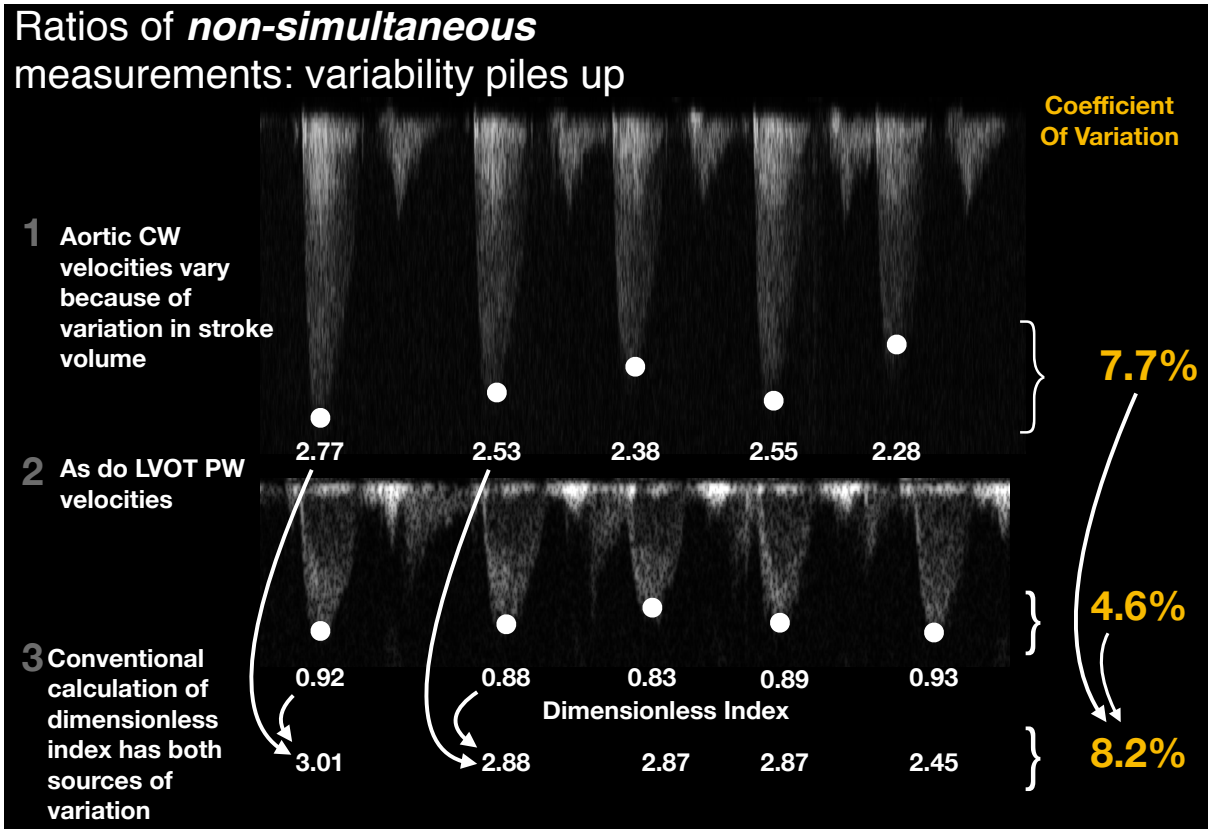


Figure 4-23 A single acquisition from an individual patient showing how variance compounds when calculating the traditional dimensionless index.

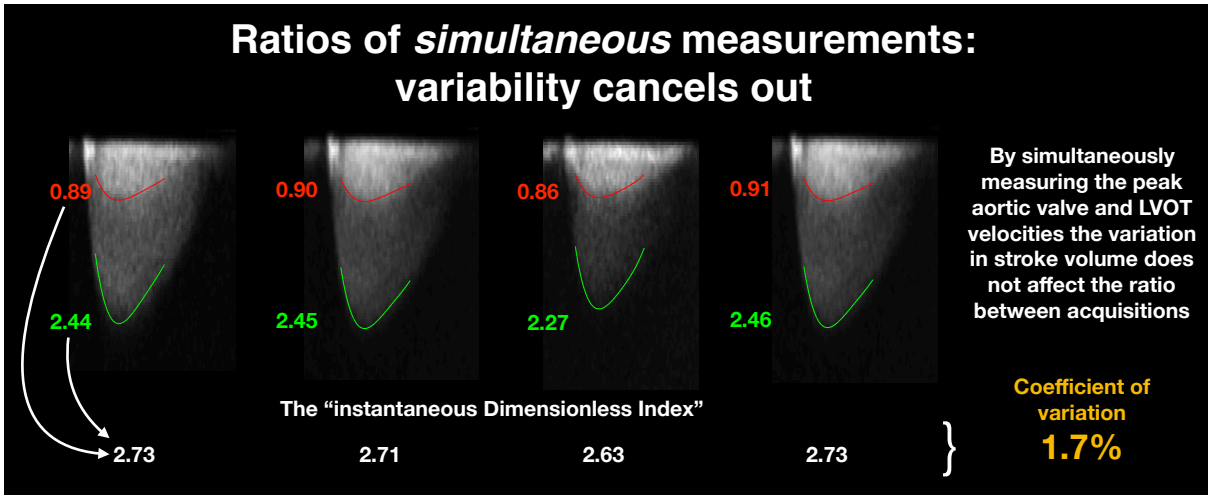


Figure 4-24 Simultaneous measurements cancel out variability with the instantaneous Dimensionless Index

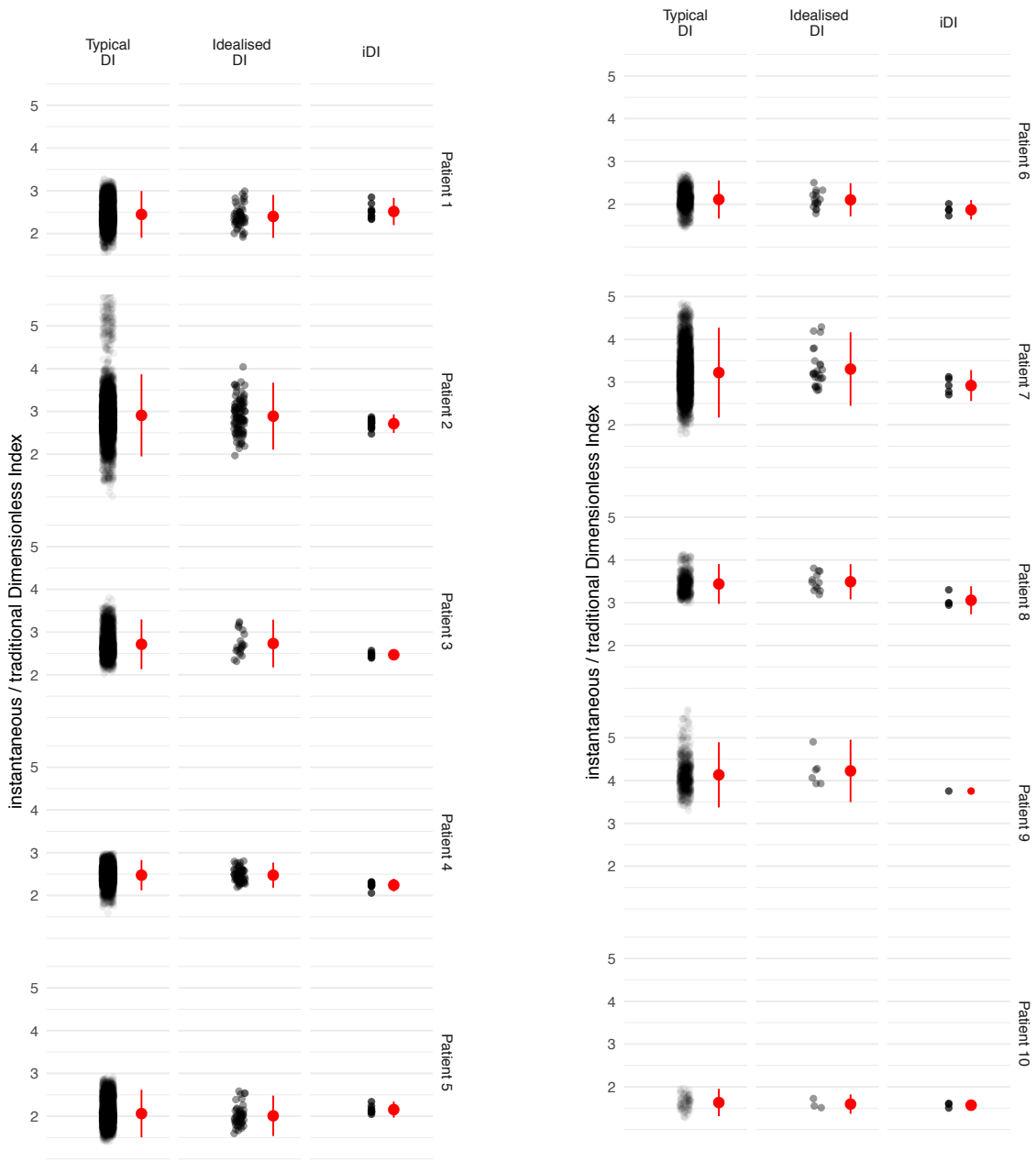


Figure 4-25 shows the variability of dimensionless index analysed in the three ways:

- Typical conventional DI (one beat of CW and one beat of PW).
- Idealised conventional DI (multi-beat average for CW and multi-beat average for PW).

- The instantaneous Dimensionless Index.

The Figure shows the results for the first 10 patients. The remaining 80 patients are shown in the Appendix.

Figure 4-25 shows that the coefficient of variation for iDI is consistently less than that of typical conventional DI or idealised conventional DI.

Across all 90 patients, the coefficient of variation for iDI was 5.8%. This was significantly less than the coefficient of variation for idealised DI (10.4%, $p<0.0001$), and the typical conventional DI (12.9%, $p<0.0001$).

The coefficient of variation for the peak continuous wave aortic velocity was 6.2%, which was not significantly different from that of the iDI (5.8%, $p=0.96$). The coefficient of variation for the peak pulsed wave LVOT velocity was 10.2%, which was significantly higher that the iDI ($p<0.0001$).

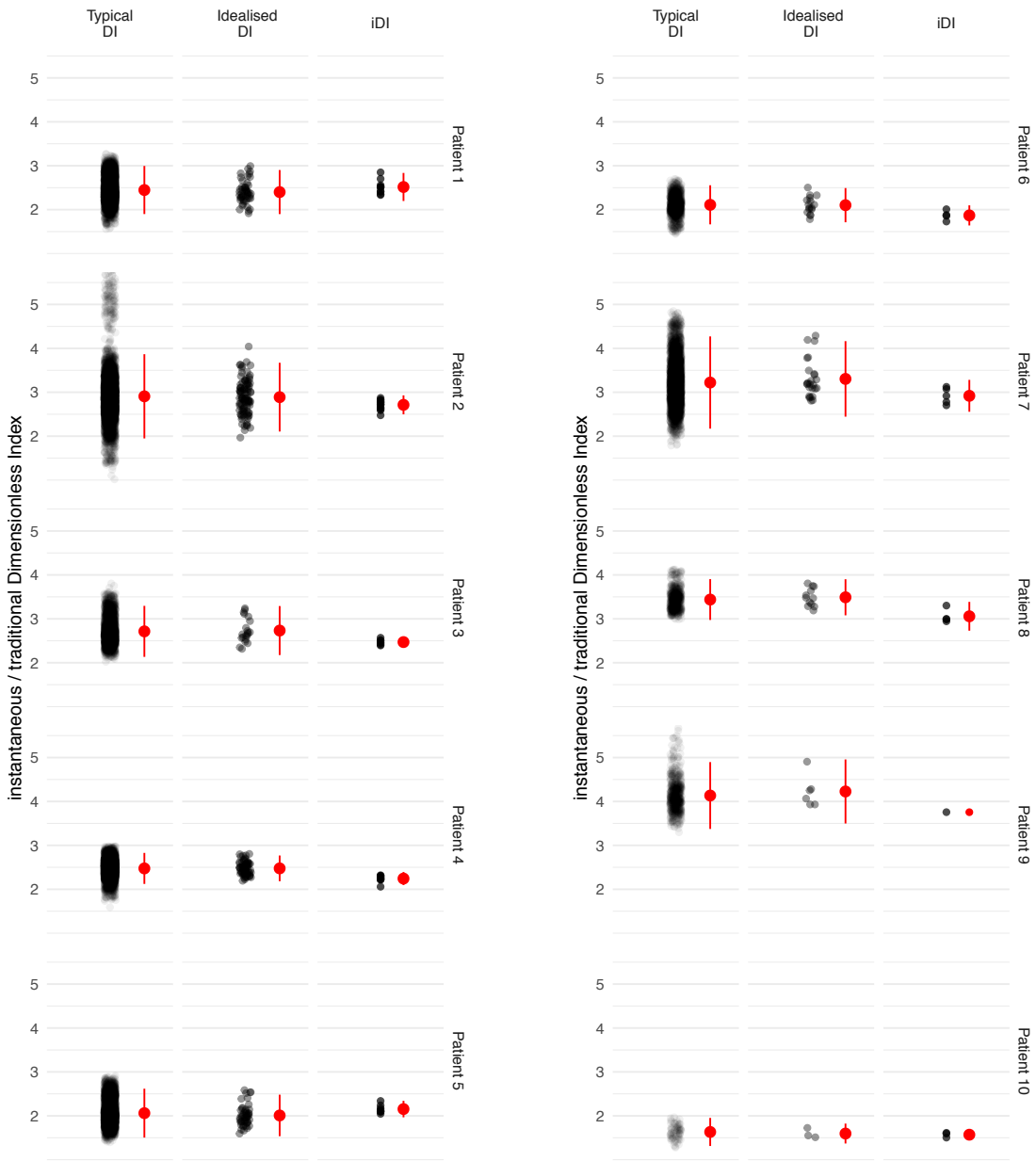


Figure 4-25 Individual values for the typical Dimensionless Index (DI), idealised DI, and the instantaneous Dimensionless Index (iDI) for the first 10 patients.

Black dots represent the individual values; red dots the mean and standard deviation.

4.4.1.5 *Clinical implications of improved precision*

The variability of a measurement can be reduced by making more measurements and taking a mean. Figure 4-26 shows the impact on the standard error of a measurement as more data is collected using. After 10 measurements the standard error of beat-by-beat DI is 3.9%, DI 3.5%, whilst iDI has a significantly lower standard error of 1.6% ($p<0.0001$, $p<0.0001$).

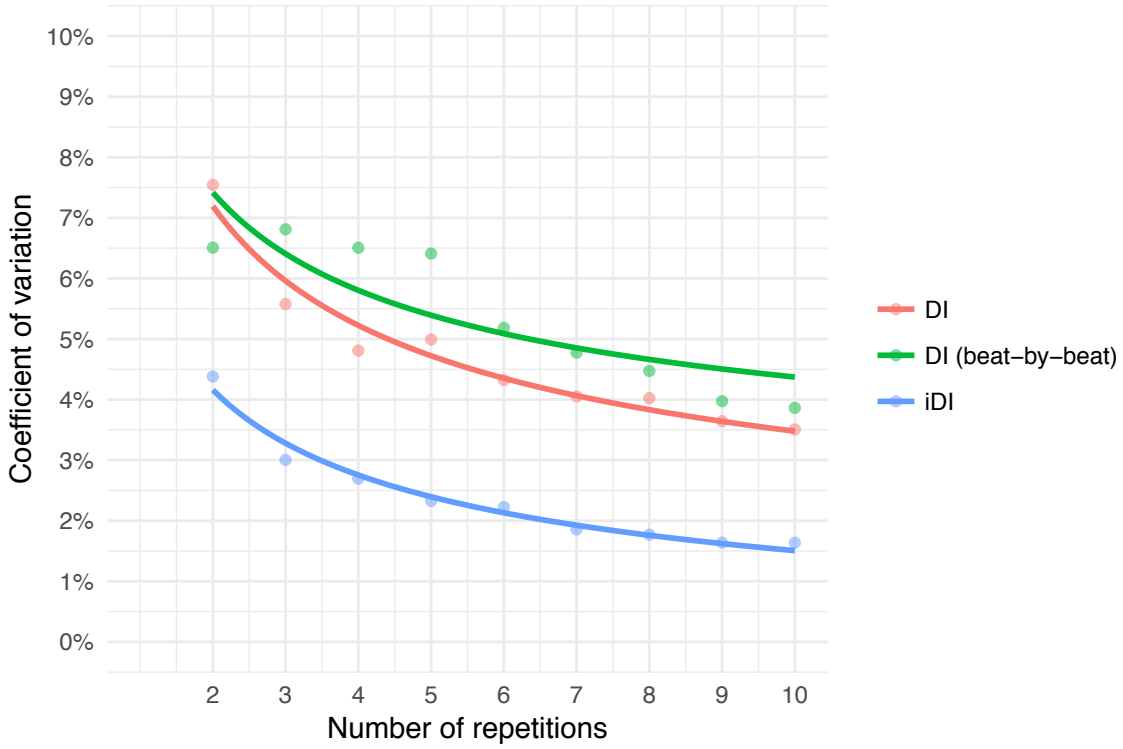


Figure 4-26 Interaction between the number of repeated acquisitions (x-axis) against the coefficient of variation (y-axis).

4.4.1.6 Visit-to-visit coefficient of variation

The mean interval between scans was 11 months (mean 332 days, sd 169). The overall visit-to-visit variability (i.e. the expected observed variability) for iDI was 9.1% for a single measurement at each visit. The typical DI and idealised DI were significantly larger than this (16.6%, 12.5%; p=0.003, p=0.007 respectively).

4.4.2 Validation of the instantaneous Dimensionless Index in Patient Group B

A second site collected data on 10 patients enrolled in a local study of aortic stenosis. As part of a local evaluation they collected continuous and pulsed-wave data from 10 patients with aortic stenosis at rest. iDI was feasible in all 10 patients (100%).

4.4.2.1 Patient Group B

Across the 10 patients the mean peak velocity was 3.37 m/s (SD 0.89, range 1.7 to 4.61) and the mean LVOT PW velocity was 1.01 m/s (SD 0.20, range 0.62 to 1.28). This resulted in a mean DI of 3.50 (SD 0.75, range 2.70 to 4.71) and iDI of 3.43 (SD 0.67, range 2.60 to 4.50). All patients had stenosis of their native valve.

4.4.2.2 Validation of the equivalence of inner profile and LVOT PW in Patient Group B

There was clinically minimal, statistically non-significant bias between the LVOT PW peak velocity and that of the inner profile of the trans-aortic CW trace (-0.038, 95% CI -0.098 to 0.023 m/s, p=0.19). Overall, there was good agreement between the LVOT velocity obtained from the PW and the inner profile of the CW trace (95% LoA -0.20 to 0.13 m/s) and a strong correlation (r=0.83, p<0.0009).

In comparison, the limits of agreement between two individual beats of LVOT PW velocities were wider (95% LoA -0.31 to 0.30 m/s) with lower correlation (r=0.70, p<0.02) and narrower for two inner profiles (95% LoA -0.21 to 0.20) with greater correlation (r=0.86, p<0.001).

Correlation and Bland-Altman plots for these agreements are shown in Figure 4-27.

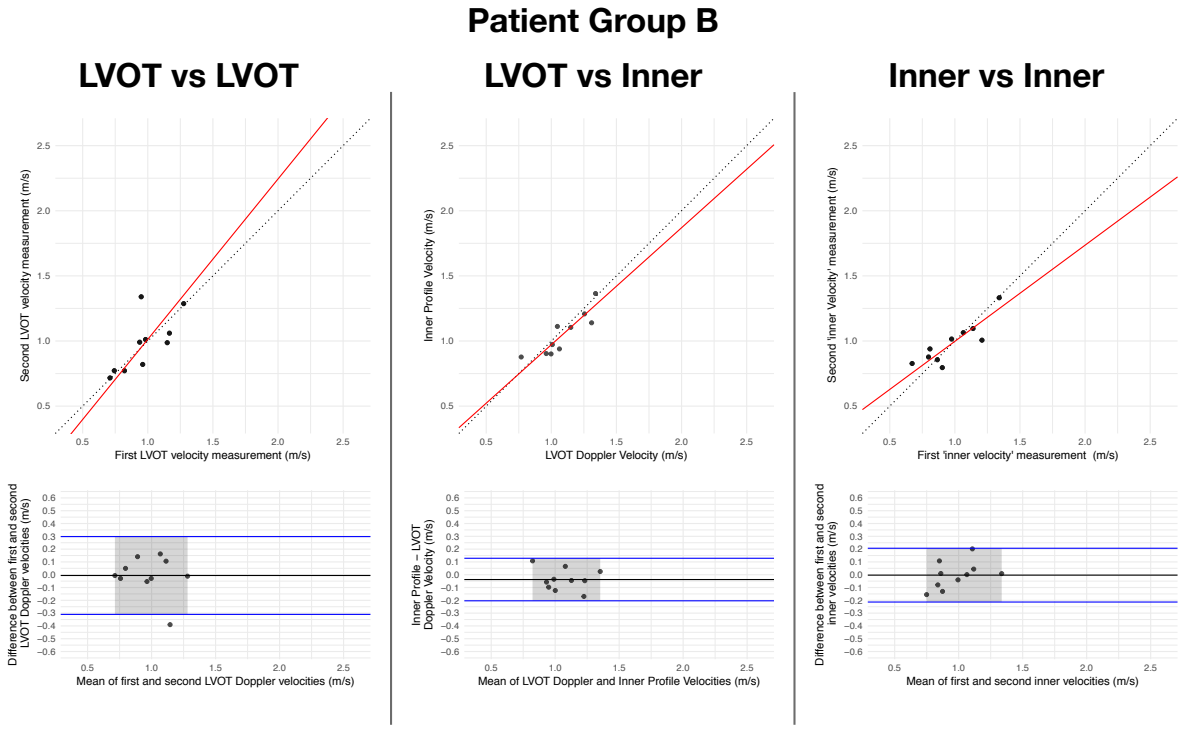


Figure 4-27 Agreement between the LVOT velocity obtained from PW and from the inner profile

Although this is a small cohort of patients, the relevant feature is that it is broadly similar to the equivalent graph obtained in my local 90 patients (Figure 4-27)

4.4.2.3 Validation of the equivalence of the DI and iDI in Patient Group B

Amongst Patient Group B there was clinically minimal, statistically non-significant bias between the iDI and DI, with iDI values being, on average -0.04 units less than DI (95% CI -0.10 to 0.02, p=0.194).

Overall, there was good agreement between the iDI and DI (95% LoA -0.20 to 0.13) with a strong correlation (r=0.87, p<0.0009). In comparison, the limits of agreement between two DIs calculated from individual beats was larger (95% LoA -0.88 to 0.82) with a weaker correlation (r=0.83, p < 0.0001) whilst the limits of agreement between two iDIs were narrower (-0.35 to 0.32). The correlation and Bland-Altman plots are shown in Figure 4-22.

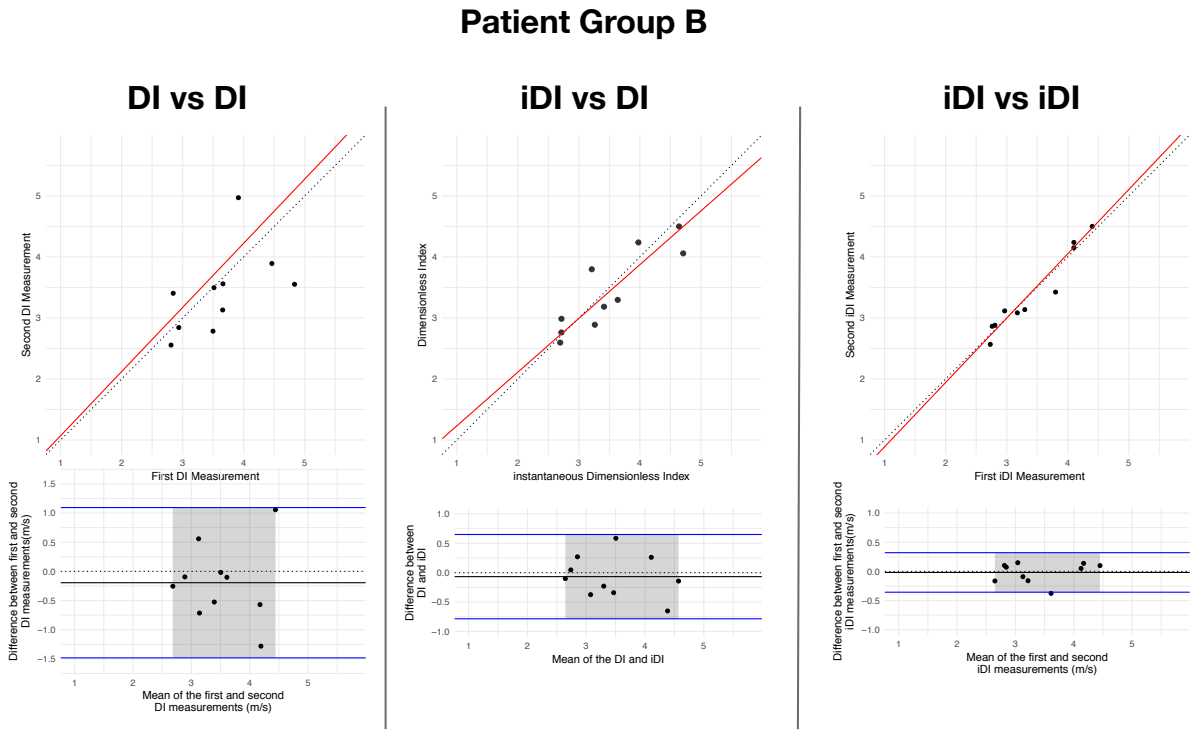


Figure 4-28 Agreement between the Dimensionless Index and the instantaneous Dimensionless Index

4.4.2.4 Comparative precision of instantaneous Dimensionless index and traditional Dimensionless Index at an independent site

Figure 4-29 shows the variability of the typical DI, the idealised DI, and the iDI for the 10 patients recruited at the independent site. In 8 of the 10 recruited patients, the coefficient of variation for iDI was less than both the typical, and idealised DI. Across all patients, the coefficient of variation for iDI was 3.8%. This was significantly less than the coefficient of variation for idealised lDI (9.2%, p=0.009), and the typical DI (17.0%, p=0.004).

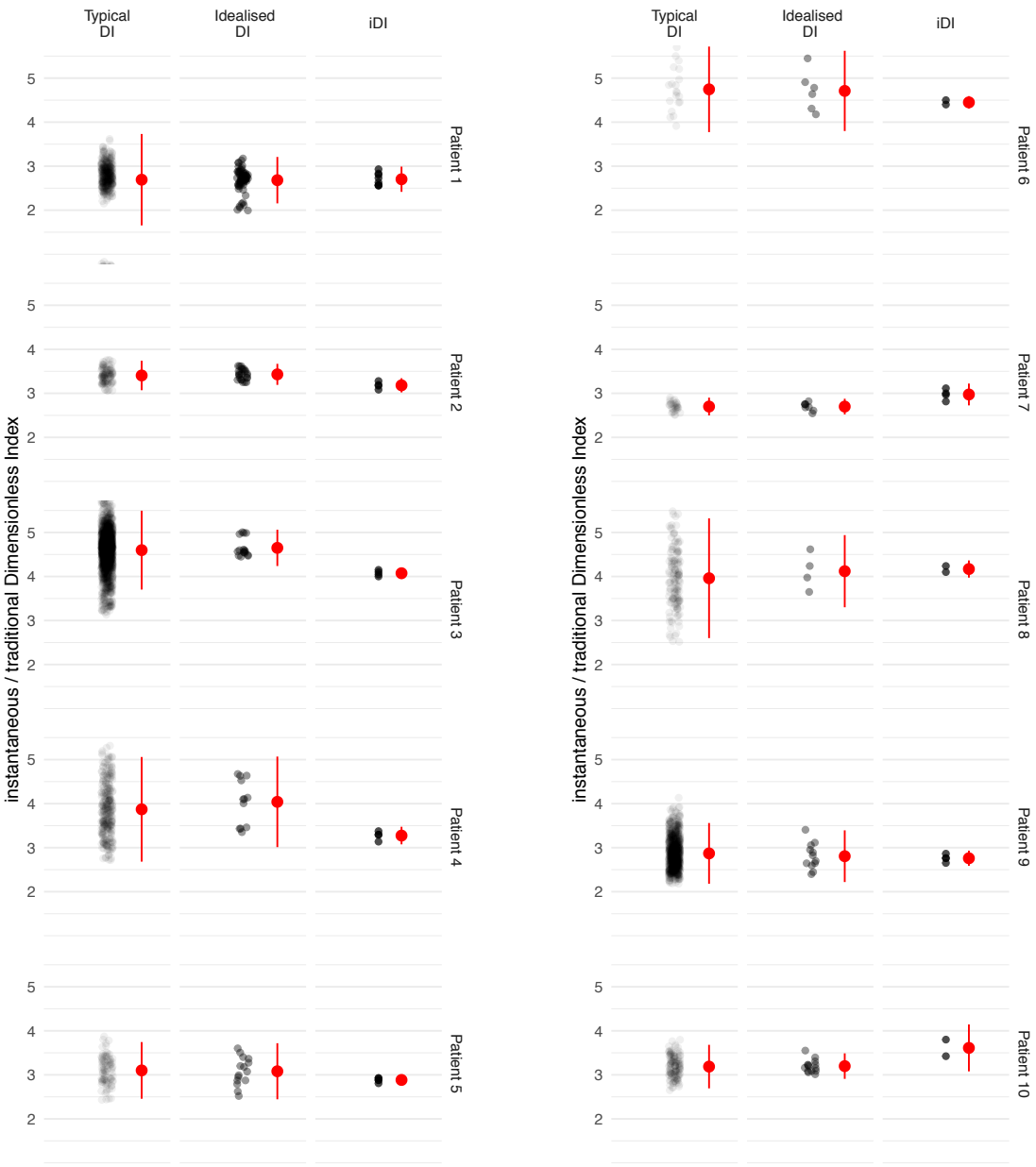


Figure 4-29 Individual values for the typical clinical dimensionless index, the idealised clinical dimensionless index, and the instantaneous Dimensionless Index

Black dots represent the individual values; red dots the mean and standard deviation.

4.5 Discussion

4.5.1 Main findings

In this Chapter, I have found:

- The instantaneous Dimensionless Index is feasible in the vast majority of patients with native or tissue valves and can be easily obtained by independent operators.

- The instantaneous Dimensionless Index can be used as a direct substitute for the traditional Dimensionless Index.

- The instantaneous Dimensionless Index provides the operator with far greater precision than the traditional Dimensionless Index with less than a quarter of the variance.

- That this improved precision translates to a reduction in visit-to-visit variability

It has long been remarked that the LVOT profile is visible within the trans-aortic continuous wave trace. Within this thesis I have demonstrated for the first time that, using enhancement through ensembling, its measurement is both highly feasible and truly reflects the LVOT peak velocity.

4.5.2 Paradoxical increase in precision

Usually, when one combines two variables, the error in the combination value is worse that the errors in each component. Statistically, this is because the errors are generally uncorrelated.

The instantaneous Dimensionless Index is unique as a compound echocardiographic variable, in that its value is less noisy than its two component parts: the coefficient of variation for the iDI was lower than either that of the peak continuous wave or pulsed wave velocity.

The reason for this surprising increase in precision is that in iDI, two noisy variables are being measured simultaneously, and the driving force for noise is minute fluctuations in stroke volume which in principle affects them in exact proportion.

4.5.3 Peak velocity or velocity time integral

When calculating the Dimensionless Index, the reader has the option of either calculating the ratio of either the peak velocities or the velocity time integral (VTI). Theoretically, one might predict that using the VTI would result in higher precision as it effectively averages the ratio over the whole profile. However, a colleague and I have previously extensively demonstrated (Sacchi et al. 2018) that when humans perform this task the variability is significantly lower when using peak velocities as compared to the VTI (CV 10.1% vs 18.0%, $P < 0.001$). We showed that significant errors arise from uncertainty in tracing at the start and end of the beat,

overwhelming any theoretical advantages of averaging the Dimensionless Index across the whole ejection period.

More fundamentally, aortic valve opening is not instantaneous. On a millisecond-by-millisecond basis, the ratio of velocities between LVOT and aortic valve may change progressively from near infinity as the valve just begins to open, to a nadir value, and then rise again.

We should not be using the velocity ratio in early or late systole because this will be artificially high.

Whilst I have described the instantaneous Dimensionless Index as using the peak velocity, in reality, because it detects the inner and outer peak using knowledge that the two profiles are scaled versions of each other, it effectively utilises the some of the advantages of the VTI.

Furthermore, as there I have demonstrated in previous work (Sacchi et al. 2018), there is significant uncertainty at the start and end of the trace. In part this is due to baseline artefacts interfering with the start and end of the trace. Furthermore, at the beginning and end of the ejection phase the valve is not fully open. Therefore, the ratio of points of the curve would not be expected to be the same as at the peak. During the development of the iDI algorithm, I found that utilising the information from 50ms before the peak to 100ms after provided a balance between maximising the information used and minimising the noise. This is shown in Figure 4-30.

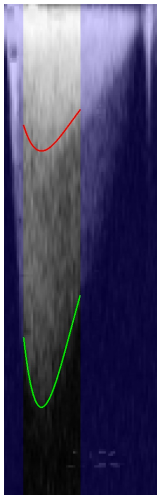


Figure 4-30 The portion of the traced profile used for the iDI calculation

The iDI is calculated using the data from 50ms before and 100ms after the peak (indicated using the green and red curve). This maximises the information used, whilst minimising the periods where there are baseline artefacts, and where the valve is not fully open.

4.5.4 Limitations

4.5.4.1 Dimensionless Index, aortic valve area, or peak velocity

The Dimensionless Index has been shown to be a key predictor of outcomes, with risk or death or valve replacement exponentially increasing as the ratio rises above 4:1. It is also added independent information to models with aortic valve area and peak aortic velocity, with no interaction with ejection, outflow tract diameter, or sex (Rusinaru et al. 2015). From data from the SEAS trial (Rossebø et al. 2008), the Dimensionless Index is also predictive of valve events in patients with low-flow, low-gradient with preserved EF aortic stenosis, even in addition to the aortic valve area (Jander et al. 2014).

Despite this, the Dimensionless Index does not directly feature either the European (Baumgartner et al. 2017) or US (Nishimura et al. 2017; Nishimura et al. 2014) guideline algorithm. However, whilst it is not explicitly featured in the diagnostic algorithm it underlies the calculation of the aortic valve area in the continuity equation.

$$AVA = \frac{\pi \times LVOT_{radius}{}^{2}}{DI}$$

Furthermore, for serial follow-up, changes in the LVOT dimensions are unlikely, and using a new recording of the "notoriously unreliable" measurement (Rusinaru et al. 2015) at each visit will only increase the variability of the AVA. Consequently, for assessing serial progression, whether in a clinical or research setting, the DI or iDI or will be at least as reproducible and sensitive as the AVA and likely to be better.

Finally, the European guidelines (Baumgartner et al. 2017) require the physician if diagnosing high gradient aortic stenosis to consider the possibility of it resulting from a high-flow. In this situation, the Dimensionless Index corrects for high flow and provides a "corrected" measure of the relative stenosis of the aortic valve.

4.5.4.2 Use of the Pedoff Tranducer

Use of the Pedoff Transducer, or "blind probe" has been recommended in assessing aortic stenosis for two reasons. Firstly, the transducer is associated with an improved signal-to-noise ratio compared to multi-element transducers. Secondly, due to its

small size, if the angulation of the jet does not align with direction of the Doppler interrogation in the default view, it allows alternative acoustic windows to be used.

The iDI algorithm was investigated on Doppler spectra obtained from the Pedoff transducer. From the apical window, similar dual-profile acquisitions can be easily obtained and simply processed. However, from alternate windows, the obtained traces are often brief and do not cut through the LVOT.

This might be regarded as a limitation of the iDI algorithm and associated protocols.

However, the iDI algorithm overcomes these two limitations.

Firstly, because the iDI is calculated on an ensembled beat, it greatly improves the signal to noise ratio, allowing for very precise measurement of the peak velocity. This is in contrast to the 3.3% coefficient of variation (Table 3.2) associated with human readers picking the peak of an already acquired Doppler trace.

Secondly, in a patient with a stable stroke volume and blood pressure, changes in the measured peak velocity arise from changes in probe position and angle. Failure to be perfectly aligned result in an underestimation of the peak velocity. Therefore, when using the peak velocity, calculating the mean of a series of acquisitions will, on average, lead to an underestimation of the true peak velocity. That is to say, as the distribution of obtained peak velocities are not symmetric, the mean is not an accurate representation of the true value.

However, this is not the case with the instantaneous Dimensionless Index. If the orientation of the LVOT and the aortic jet differ; with each acquisition an operator will typically either be better aligned with either the direction of the aortic jet or LVOT flow. They therefore, on each acquisition, they will have equal chance of overestimating the iDI or underestimating the iDI. If multiple acquisitions are made, the distribution of values would be expected to be symmetrical, centred around the true Dimensionless Index. Therefore, even if neither the "true" peak aortic velocity, and the "true" LVOT velocity have not been captured, the mean of the obtained iDIs would still be expected to be the true iDI.

4.5.4.3 Additional operator time

The iDI algorithm was designed to add minimal extra time. Inevitably, despite the development of both software that can be easily used post-hoc with exported DICOMs, or perform the analysis live on a laptop plugged into the video output port available on most echocardiographic machines, there is a small amount of added time. However, if the software was integrated "on-cart", no added time would be required for transferring images or utilising another device. Furthermore, as it eliminates the requirement for operators to trace and measure multiple beats, and provides improved precision with less captured data, it has the potential to both reduce the total time required and improve precision.

4.5.5 Implications for clinical practice

In routine clinical practice, patients and their clinicians will have far greater confidence in reported values, improving decision-making, and reducing the need for unnecessary follow-up. Our open-source method is vendor-independent, works using existing machines, adds no additional time, and could be immediately added to any workflow in laboratories worldwide.

4.5.6 Implications for research

With the prospect of disease modifying therapies for aortic stenosis, the role of imaging biomarkers for aortic stenosis is changing. A researcher assessing the clinical efficacy of therapies requires that the biomarker is both reproducible and responsive to changes, thereby maximising the signal-to-noise ratio and minimising the number of patients required within the trial.

The instantaneous Dimensionless Index provides a familiar measure for researchers that, through its equivalence to conventional DI has been shown to be predictive of outcomes. Furthermore, it provides this with almost a quarter of the variance of the traditional dimensionless index. Studies using the iDI as compared to the DI would therefore be able to recruit fewer patients, have greater precision, and better identify sub-groups that benefit from therapy. Furthermore, within a clinical trial, the addition of automated analysis helps ensure quality is maintained with less training and supervision, and further reducing costs.

4.5.7 Future directions

4.5.7.1 Use in pilot studies of aortic stenosis

Multiple pilot studies are underway investigating the use of existing pharmaceutical compounds in diminishing the progression of calcific aortic stenosis. As part of this work I have been in contact with these groups and they are collecting pilot data to see if they find it feasible, reliable, and suitable for use in their trials.

4.5.7.2 The challenge of aortic valve assessment during stress echo

Some patients with a severely stenosed aortic valve also have a low stroke volume. There are two principle challenges when stroke volume is low.

- Low stroke volume means all the velocities are accordingly reduced so that 4m/s is not reached.

- Low stroke volume means aortic valve leaflets may receive inadequate force to their fullest mechanically achievable orifice area.

In the former case, trans-valvular velocities (and the gradients calculated from them) will be lower than they would be with the same valve but a normally functioning ventricle. In contrast, in the latter case, the dimensionless index would be higher than it would be with the same valve and a normally functioning ventricle.

Increasing stroke volume with exercise or dobutamine is an opportunity to resolve both challenges.

As the stroke volume increases:

- All the velocities increase so that aortic valves that had too low volume flow rate to show a peak velocity above 4 m/s have an opportunity to show it (care must be taken to over-interpret increases in stroke volume above that require to counteract the baseline reduction in stroke volume).

- If the valve was opening inadequately due to inadequate ventricular force generation, this is now rectified, and the valve opens to the extent that its true inherent restriction permits.

However, performing echocardiography during exercise is difficult because the heart moves with ventilation, moving in and out of the imaging plane. Furthermore, during exercise placing the pulsed wave sample volume becomes even more difficult. Because the dimensionless index is calculated from dividing one value by another, growing noise on each value results in dramatically increased noise in the dimensionless index.

The instantaneous Dimensionless Index avoids these problems. By requiring only a continuous wave Doppler acquisition it eliminates the most unreliable part of the acquisition, which is placing the sample volume for the pulsed wave. It also allows the sonographer to focus solely on acquiring the continuous wave Doppler. Finally, by reducing the workload involved in analysing each beat, it makes it much easier for the sonographer to analyse multiple beats. All three advantages increase precision.

Figure 4-31 shows one such example. In the top panel, a patient with suspected severe aortic stenosis and reduced left ventricular function had a peak velocity was below guideline thresholds for severe aortic stenosis. However, the instantaneous Dimensionless Index was 4.0 suggesting severe aortic stenosis. Exercise was used to increase the flow. At peak exercise, flow increased by 44%, but the instantaneous Dimensionless Index remained 4.0 confirming the stenosis was severe.

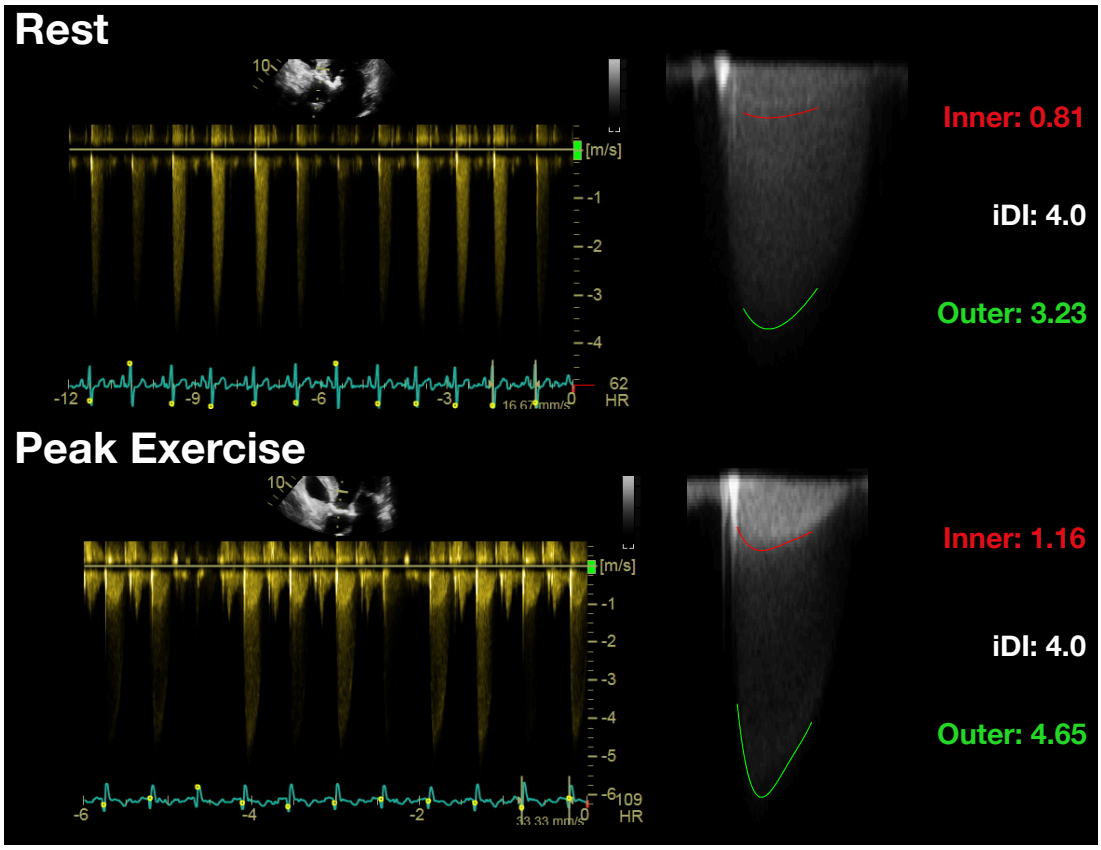


Figure 4-31 The use of the instantaneous Dimensionless Index on exercise

Using continuous wave Doppler, a screenful of beats (12 in this example) is captured using, in this case, a GE echocardiographic machine at a slow sweep speed. Within the DICOM file the Doppler spectrum is stored at a fast sweep speed. The DICOM file is processed and the instantaneous Dimensionless Index is calculated to be 4.0 meeting the criteria for severe aortic stenosis. However, the outer (peak) velocity is 3.23 m/s, which only meets the criteria for moderate stenosis. By increasing the flow through the valve using exercise the flow and the peak velocity rises by 44% to 4.65m/s. The inner (LVOT) velocity rises by the same proportion and the instantaneous Dimensionless Index remains 4.0 confirming that the valve remains non-compliant with higher flow-rates and the patient truly has low-flow low-gradient severe aortic stenosis.

4.6 Conclusion

The instantaneous Dimensionless Index is easily performed, direct substitute for the Dimensionless Index that doubles the precision of aortic stenosis measurements.

5 The development and pilot validation of “AQURO”, an automated method for the quantification of mitral regurgitation severity

5.1 Abstract

5.1.1 Introduction

In this Chapter I explain the rational behind, the further development of, and pilot clinical validation of AQURO (Automated Quantification of Regurgitant Orifice) for mitral regurgitation. AQURO provides automated quantification of the EROA. It is vendor-independent and uses novel automated assessment techniques.

Hypothesis: AQURO has equivalent performance to PISA in classifying the severity of mitral regurgitation but with higher precision.

5.1.2 Methods

I developed upon and improved the AQURO algorithm. AQURO improves the reproducibility of the assessment of the severity of mitral regurgitation by improving upon PISA through the automated interpretation of colour M-mode. It eliminates some of the limitations of PISA measurements such as the requirement of knowing the exact location of the orifice and the need to measure a small distance, and uses information from the whole cardiac cycle rather than at a single time-point. Furthermore, automation allows for the collection and analysis of a greater quantity of data from a patient, improving precision through averaging.

For the pilot validation study, I prospectively recruited patients with primary mitral valve regurgitation in sinus rhythm at my institution and its community clinics.

Multiple recordings of multiple-beats of colour Doppler M-mode and continuous wave Doppler through the regurgitant mitral valve were made in every patient. In

addition, videos suitable for calculating the EROA using PISA were captured. For each patient multiple assessments of the EROA using both AQURO and PISA were obtained. In addition, two independent echocardiographers reviewed the B-Mode and colour Doppler videos of all cases (without the AQURO and PISA measurements) and provided a visual estimate of the severity of mitral regurgitation. The relative precision and bias of AQURO and PISA derived EROAs were then compared.

5.1.3 Results

I recruited 41 patients with mitral regurgitation. 3 patients were excluded only a single PISA acquisition was usable, and 2 patients were excluded as the AQURO analysis failed. There was a significant relationship between the visual severity of the EROA derived by both AQURO ($p=0.003$, $rho=0.613$) and PISA ($p=0.0008$, $rho=0.617$). The coefficient of variation for a single measurement of AQURO and PISA derived were large (67% and 57% respectively). However, within the context of the spread of EROA values (from 7 mm^2 to 130 mm^2) the intra-class correlation coefficient is good (0.74 vs 0.8, difference -0.7, credible interval -0.21 to 0.06). As might be expected from the derivation of AQURO and PISA, AQURO derived EROA was on average 34% ($p=0.003$) larger than PISA derived EROA.

5.1.4 Conclusions

AQURO is an automated method of calculating the mitral regurgitant EROA from colour M-mode doppler images. It provides comparable precision and values to that

of PISA derived EROA when a single measurement is made, but allows for the simple acquisition of far greater amounts of data.

5.2 Introduction

Mitral regurgitation is a curious contrast to aortic stenosis. For aortic stenosis, the ESC guidelines describe over a dozen quantitative parameters that can be used to assess its severity. In mitral regurgitation, ESC guidelines recommend only one quantitative parameter. This is the Effective Orifice Regurgitant Area (EROA) (Baumgartner et al. 2017).

The regurgitant orifice cannot be directly visualised because it is moving, has an unknown shape and possibly variable shape, and is often small compared to the resolution of 2D ultrasound. Therefore, as with aortic stenosis, the regurgitation orifice area is assessed indirectly by effects on the flow patterns in its vicinity.

Originally described in 1988 (Bargiggia et al. 1991; Recusani et al. 1991; Bargiggia et al. 1988), the Proximal Isovelocity Surface Area (PISA) method is typically used to calculate the EROA. Similar to the calculation of the aortic valve area in aortic stenosis, it takes advantage of the continuity principle (mass conservation for a steady flow) to calculate the area at the level of the mitral valve during regurgitation. As mass is conserved, the flow at the mitral valve orifice is equal to a surface around the orifice (Figure 5-1).

Equation 1 The continuity equation

$$EROA \times V_{jet} = A_r \times V_r$$

V_{jet} : The peak velocity of the regurgitant jet (measured using continuous wave).

A_r: The surface area of the isospeed hemisphere.

V_r: The aliasing velocity (i.e. the speed of the isospeed hemisphere).

Equation 2, the surface of a hemisphere

$$A_r = 2\pi r^2$$

Substituting Equation 2 into Equation 1 and rearranging gives Equation 3

$$EROA = \frac{2\pi r^2 \times V_r}{V_{jet}}$$

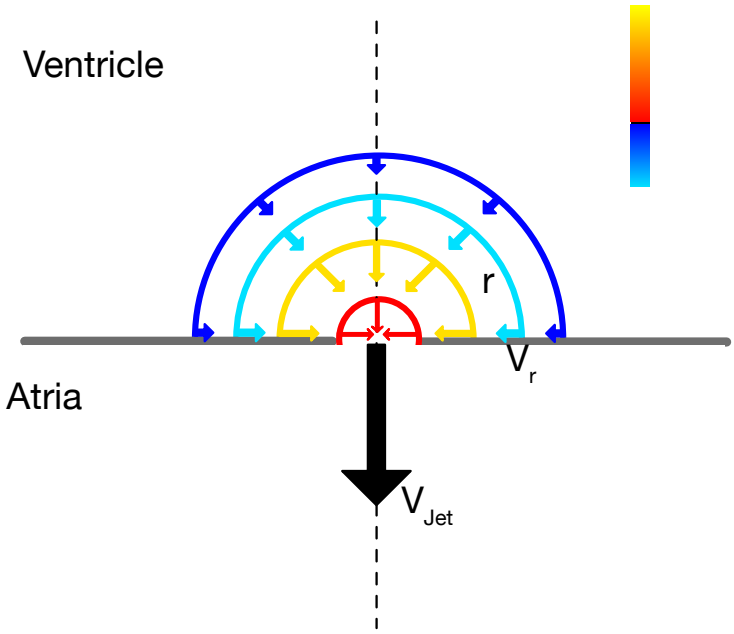


Figure 5-1 Using the PISA method to calculate the EROA of a regurgitant mitral valve

V_{Jet}: Velocity of the jet derived from a CW Doppler trace; V_r: Velocity at distance r (typically taken to be the aliasing velocity); r: distance of V_r from the orifice.

5.2.1 Difficulties encountered by clinicians in calculating the EROA using the PISA method

Despite the apparent simplicity of PISA, there is a great scope for error. This is true for both qualitative (Thomas et al. 2008) and quantitative estimates (Moraldo, Bergamini, et al. 2013).

Practical experience, reproducibility studies, and comparisons of guideline described behaviour with theoretical principals demonstrate there are multiple difficulties encountered by both trainees and experts alike in measuring the PISA distance.

PISA requires a small distance to be measured: Small errors, even of one or two pixels can cause significant changes in the EROA. As the EROA is proportional to the square of the PISA distance, a change in distance from 0.7 to 0.8cm leads to a 30% increase in the calculated EROA.

PISA requires the sonographer to guess *where* the regurgitant orifice is: Often this point is not identifiable with confidence to within the < 0.1cm required, because there is no recognisable object at that location. Rather it is an inference about where the regurgitation is emanating from.

PISA is labour intensive: Measuring the PISA distance requires significant time in acquiring a single zoomed frame and then measuring the PISA distance. This limits the number of beats worth of data that can be acquired. Often patients with significant mitral regurgitation are in atrial fibrillation, in which gathering more beats worth of data would be desired to minimise the impact of noise.

PISA does not maximise the use of information: The PISA distance is measured on one single frame, discarding the information available from the rest of systole.

PISA requires eyeballing the diameter: Measuring the PISA distance requires the operator to select the frame with the maximal PISA distance by eye.

PISA is very sensitive to the imaging plane: If the sonographer does not cut directly through the orifice the PISA distance will be under-estimated.

PISA is on colour Doppler does not appear hemispherical: Colour Doppler visualises ***non-hemispherical isovelocity*** contours, rather than the described ***hemispherical isospeed*** contours. Belief that the contours seen on colour Doppler are semi-circular rather an "urchinoid" shape leads to a systematic underestimation of the PISA distance. This is explained below in Section 5.2.2.

PISA measurement direction is unclear: If the orientation valve is not perpendicular to the ultrasound beam, the direction in which the PISA measurement should be made is not clearly specified, and is demonstrated incorrectly within guidelines

The shape of the isovelocity contours on colour Doppler has significant implications for the both where to measure the PISA distance (8), and in which direction (9).

5.2.2 What is the true PISA shape?

The visualised appearance of the convergence zone on the proximal side of an orifice on 2D colour Doppler is not a "hemisphere" as often described (Moraldo, Cecaro, et

al. 2013; Lancellotti et al. 2013), but is more of an "urchinoid" shape. The reason for this is whilst idealised isovelocity field consists hemispheres where the blood is travelling at the same ***speed***, colour flow Doppler can only measure the velocity component in the direction of the probe (Figure 5-2).

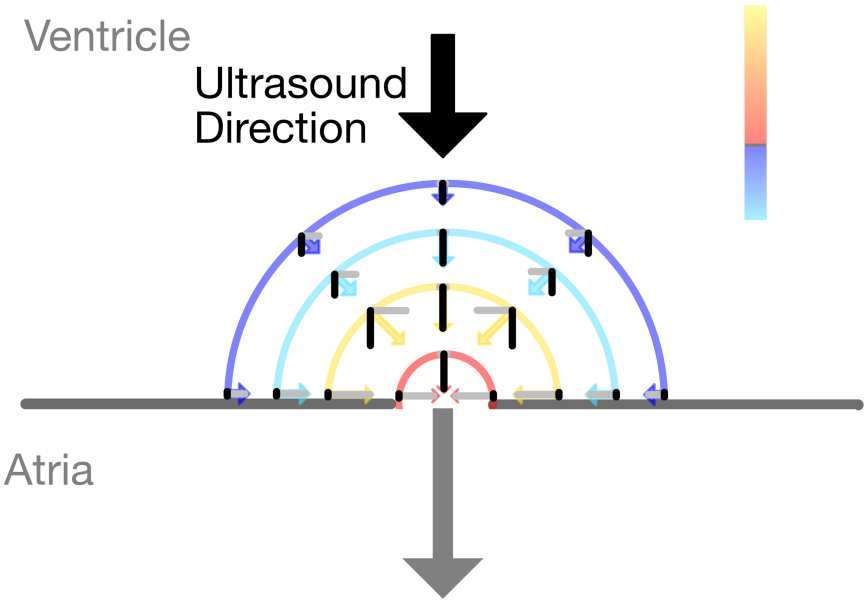


Figure 5-2 Depending on the direction of flow, the same speed will have a different velocity component in the direction "seen" by the ultrasound probe

The vertical velocity component in the direction of the probe is indicated in black, and the ignored horizontal component in grey. Along each coloured isospeed contour the vertical velocity

Therefore, the contours of the Doppler ***isovelocity*** field is different from that of the ***isospeed*** contours. Indeed, the Doppler ***isovelocity*** field forms an "urchinoid" shape (Figure 5-3).

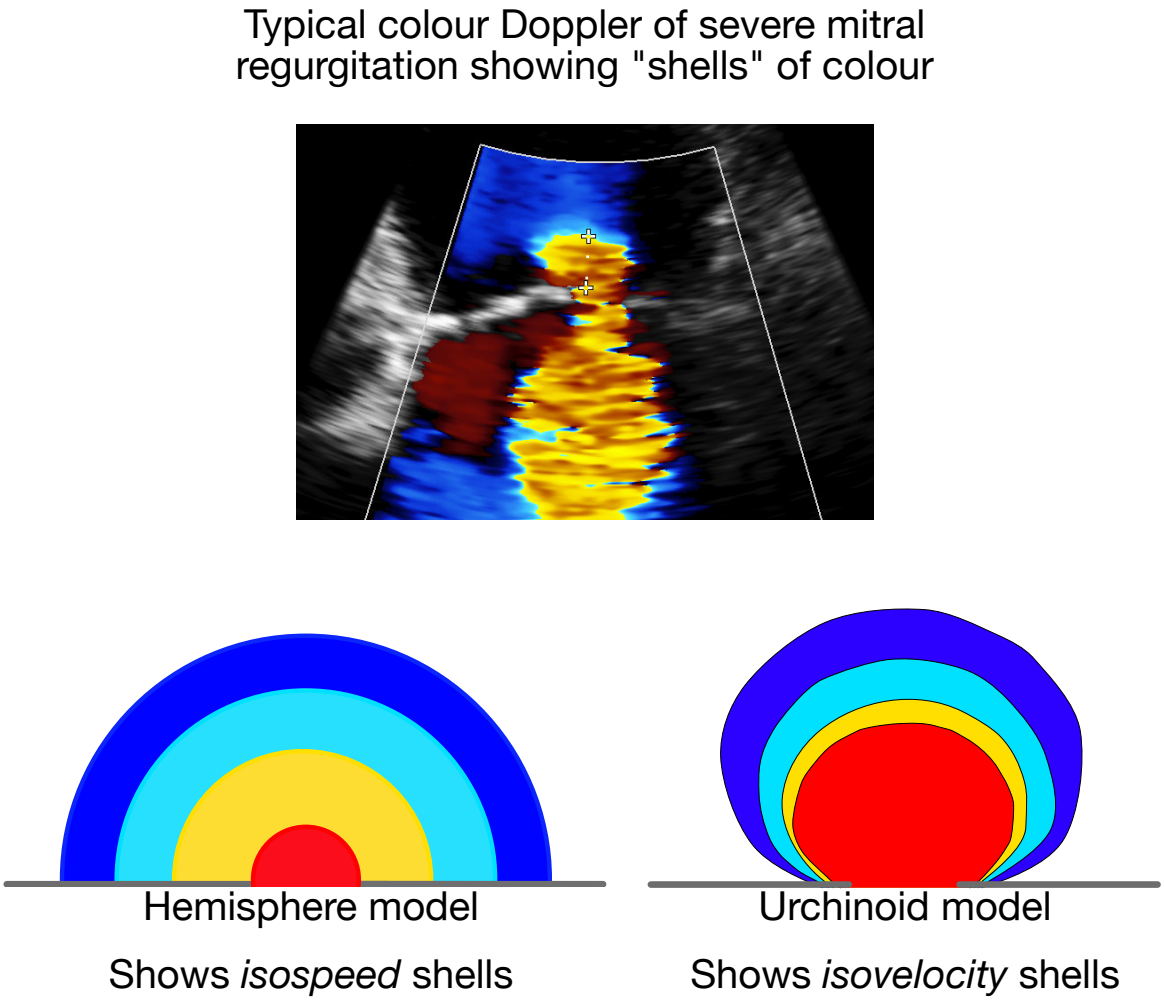


Figure 5-3 The colour Doppler PISA shape is more urchinoid in shape than hemispherical. Colour Doppler can only measure the velocity in the direction of the probe rather than the true speed. Therefore, the theoretical hemispherical isospeed contours become urchinoid isovelocity contours when visualised with Doppler.

As this relates to the fundamental principle that the echo Doppler only measures motion in the direction of the probe, it is true even in the relatively idealised experimental scenario of a small, round orifice.

No blame for this error of confusing the theoretical hemispheric ***isospeed*** contours with the Doppler urchinoid ***isovelocity*** contours can be laid at the authors of the original articles. In these articles (Recusani et al. 1991; Bargiggia et al. 1991), the authors very carefully refer to only the contours as being hemispherical within the theoretical section of the paper where they are referring to speeds (and not Doppler velocities). Even with their lower resolution colour Doppler, the visualised shells are not semi-circular, and furthermore, for the benefit of readers, the smooth traced shells have not been approximated to semi-circular shape. Unfortunately, they did not provide the readership with a memorable name for the Doppler profile shape, which may have led to people misapplying the term hemispherical.

I contributed to a systematic review of guidelines that found that out of 14 descriptions, 9 referred to the shape as a hemisphere, 1 as "spherical or non-spherical", one as "hemispheroid", 1 as "hemielliipsoid", and 3 did not give a description (Moraldo, Cecaro, et al. 2013). Furthermore, the current 2013 ESC Valve imaging guidelines (Lancellotti et al. 2013) specifically describe the "downward shift of zero baseline to obtain a hemispheric PISA".

5.2.2.1 Does the description of the shape matter?

Previous work I have assisted in a previous study of priming cardiologists with different types of instruction. The instructions:

- either described the shape as "hemispherical" and asked them to measure its radius,

- or as a “non-hemispherical urchinoid” and asked to measure the diameter.

This study found that trainees were particularly vulnerable to priming with the term “hemispherical”. When primed in this way, they tended to significantly underestimate the PISA distance (and hence EROA). In some cases, the underestimation was so dramatic that there was no overlap between trainee and expert results (

Figure 5-4).

Consequently, given that this is a substantial source of incorrect measurements in calculating the EROA, and any algorithm must take this into consideration.

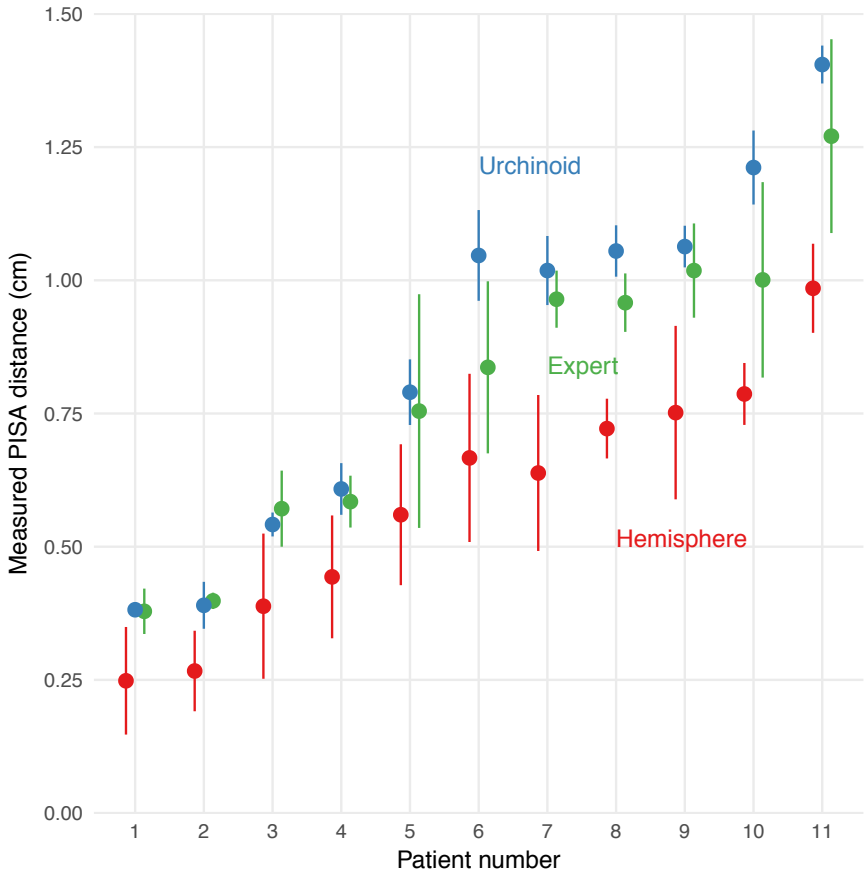


Figure 5-4 Relationship between description provided of PISA shape and measured PISA distance.

The description provided to the readers clearly influences the measured PISA distance. Adapted from (Moraldo, Cecaro, et al. 2013).

5.2.2.2 In which direction should the PISA distance be measured?

When the valve plane is not perpendicular to the ultrasound beam, the direction in which the PISA distance should be measured is not immediately obvious. The sonographer has a few options for measurement:

- Perpendicular to the orifice plane
- Directly up
- Directly towards the probe

The ***isovelocity*** contours in only coincide with the ***isospeed*** contours when in the direction of the ultrasound beam. Therefore, the PISA distance should be measured in the direction of the probe. Indeed, as the probe angle rotates around the valve the PISA urchinoid will appear to "look" towards the probe. This was described at the "Mona Lisa PISA" by my colleagues in their paper (Moraldo et al. 2012). Any other direction of measurement will lead to an underestimation of the PISA distance.

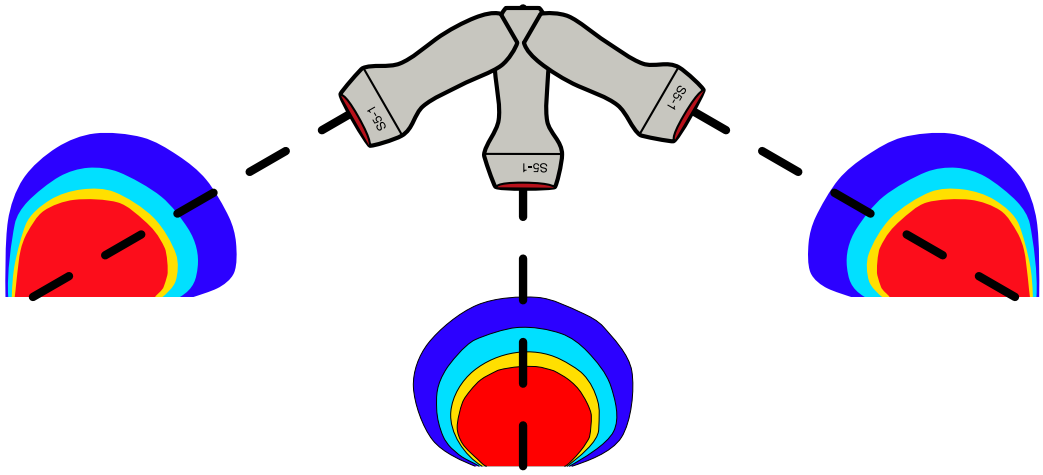


Figure 5-5 The Mona Lisa PISA effect

Because the velocity and the speed are of equal magnitude only when the flow is directed perpendicular to the probe the ***true*** velocity and hence the ***true*** PISA distance is seen when measured in the direction of the probe. As the angle at which ultrasound probe is directed towards the convergence zone, the apparent shape will change. Like the Mona Lisa, it will always "look" towards the viewer (in this case the ultrasound probe) (Moraldo, Cecaro, et al. 2013).

Unfortunately, guidelines do not specify, and indeed, in their example figures often get it wrong. For example, in Figure 6E of the 2013 ESC Guidelines (Lancellotti et al. 2013) the PISA distance is drawn at 45 degrees to the direction of the probe. Consequently, the drawn line goes from a slower velocity at the valve, to a faster velocity in the middle of the convergence zone, to a slower velocity at the edge of the convergence zone. When drawn correctly, the PISA distance should cross the velocity contours in a way that there is continuous acceleration towards the valve.

5.2.3 The AQURO method – resolving these difficulties

There have been attempts to improve upon the difficulties with PISA, by other authors (Giesler & Stauch 1992; T Shiota et al. 1994; Takahiro Shiota et al. 1994; Vandervoort et al. 1993). Our group built on this theoretical work to develop AQURO - Automated QUantification of Regurgitant Orifice area (Moraldo, Bergamini, et al. 2013).

The PISA formula can be reformulated to eliminate the need to specify the location of the regurgitant orifice. This is a crucial advantage because there is no algorithm available to automatically define the location of the orifice. This in turn may arise from this task being inherently vague.

The key step in the transformation is the realisation that the EROA can be calculated from:

- *either* the PISA radius (inherently requiring the orifice location to be known)
- *or* the velocity gradient over space inside the flow convergence zone (avoiding the need to know the orifice location).

5.2.4 The AQURO principle

The AQURO principle can most easily be understood by reference to a thought experiment.

5.2.4.1 The thought experiment of the draining pond.

Imagine you are standing in a wide shallow pond which is draining through a hole of unknown size. If you stay in the same place, the speed of water going past your legs is an index of the flow rate through the hole.

If you wanted to calculate the rate of flow out of the pond, you would need to know not only how fast water was going past your feet, but also where the hole was. This is because, closer to any given size of hole, water velocity is faster. Therefore, a given velocity against the feet may represent a small hole nearby, or a large hole far away.

However, suppose you were denied knowledge of the location of the hole. Instead, though, you have a friend a known distance away (in the direction of flow) who can also report the water speed. By chance, he reports that the water is going twice as fast across his feet as it is across yours, and in the same direction.

This allows you to know where the hole is (as shown below).

In essence, the volume flow rate through the outer circle must equal the volume flow rate through the inner circle. Since the velocity has doubled, the circumference of the circle must have halved. Therefore, the hole is located exactly as far away from your friend as you are (but on the other side).

This principle can be extended for any ratio of velocities between your feet and your friend's feet. The greater that ratio, the closer the hole is.

Although the example is given with velocities measured at two points, it can of course, reformulated for a series of velocities across a range of points in a straight line (as long as that straight line is going through the orifice).

When reformulated, the flow rate can be calculated from the velocity gradient over space.

However, this thought experiment refers to a two-dimensional velocity map namely in a shallow lake. In reality, fluid flows near a leaking mitral valve have a three-dimensional structure.

5.2.4.2 AQURO

The impact of the presence of the third-dimension is that it is no longer the circumference times the speed that is conserved across different radii, but now the hemispherical area times the speed. This has an important mathematical consequence that it is no longer a simple velocity gradient (dv/dx) but rather a transformed version. The derivation of the equation is shown below.

Equation 4, The continuity equation

$$EROA \times V_{jet} = A_r \times V_r$$

Rearranging gives Equation 5:

$$\frac{V_{jet}}{V_r} = \frac{A_r}{EROA}$$

Equation 6, the surface area of a hemisphere

$$A_r = 2\pi r^2$$

Substituting one into the other gives Equation 7:

$$\frac{V_{jet}}{V_r} = \frac{2\pi}{EROA} r^2$$

$$\sqrt{\frac{V_{jet}}{V_r}} = \sqrt{\frac{2\pi}{EROA}} \times r$$

This is of the form:

$$y = m \times x$$

With:

$$y = \sqrt{\frac{V_{jet}}{V_r}}$$

$$x = r$$

$$m = \sqrt{\frac{2\pi}{EROA}}$$

When calculating the EROA using PISA, a single distance from the valve (r) at the aliasing velocity (V_r) is measured. With AQURO, the velocities along each scan line (picking any arbitrary point as the reference distance) are extracted. A graph of $x = r$ against $y = \sqrt{\frac{V_{jet}}{V_r}}$ is then plotted. The gradient of this graph with a linear

relationship is: $\sqrt{\frac{2\pi}{EROA}}$. Thus, if the gradient of transformed velocities can be measured from an image, it can be used to calculate the EROA.

The AQURO equation, Equation 8:

$$EROA = \frac{2\pi}{gradient^2}$$

The reformulation of the PISA equation allows many of the previously described difficulties to be overcome. There is therefore no need to specify the location of the orifice, but it is now necessary to measure the rate of change of velocity with distance precisely.

Colour M-mode images are ideal for this purpose. The key is to place the interrogation line through the mitral valve (Figure 5-6). The AQURO calculation is outlined in Figure 5-7.

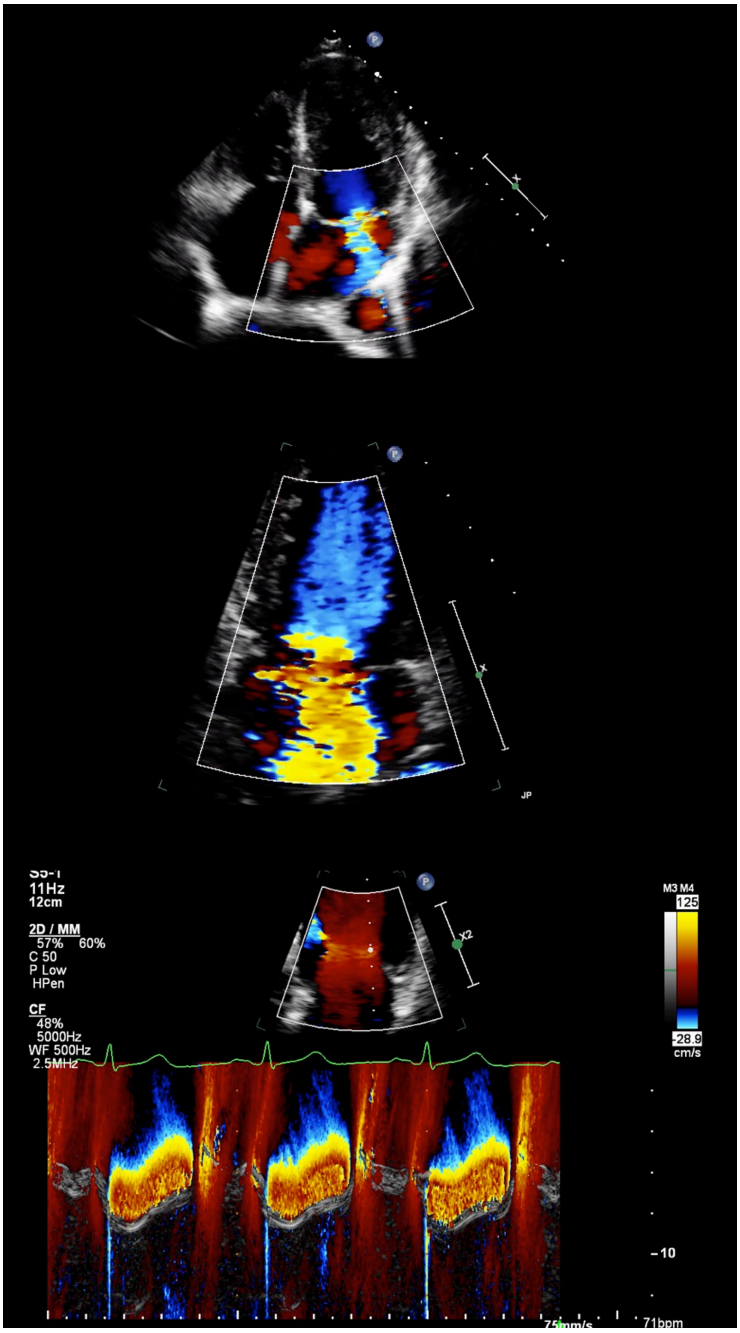


Figure 5-6 Acquired colour M-mode of severe mitral regurgitation for AQURO

Colour Doppler is places over the mitral valve (top panel), with the region magnified using the zoom function (middle panel). Colour M-mode is then acquired through the valve (bottom panel).

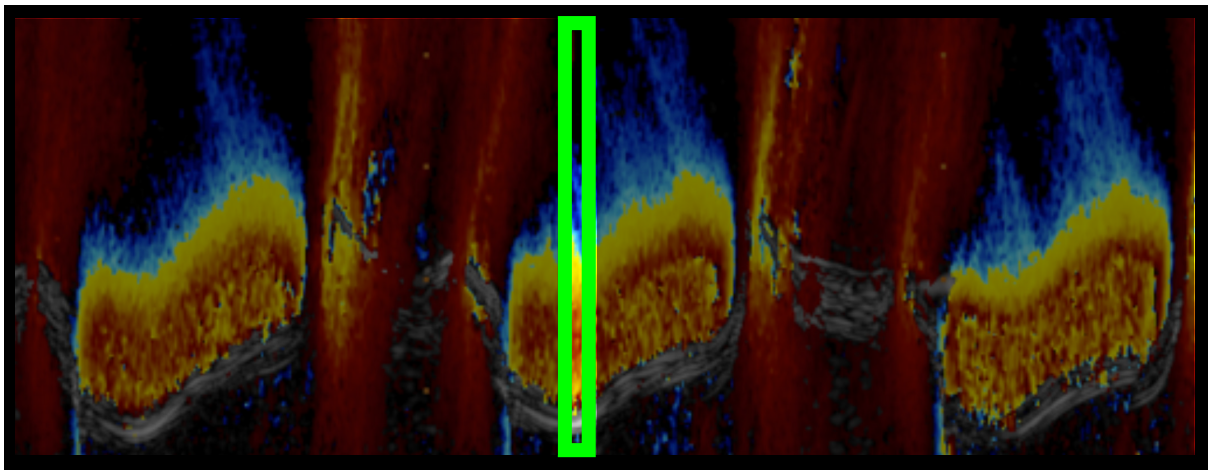


Figure 5-7 The AQURO method - Introduction

The AQURO method is based on looking not only at the velocity of blood but also the rate of its change with distance. If the velocity at a particular point is (say) 0.50 m/s, whether the regurgitant orifice is large or small depends on how quickly the velocity is changing with distance. If the velocity is changing quickly with distance, i.e. falling to near zero very quickly as one moves away, the regurgitant orifice must be nearby, and is (therefore) small. If the velocity is not changing much with distance, the regurgitant orifice must be far away, and (to still produce this velocity at such a distance) must be large. The tricky element of the calculation is correctly judging the rate of change of velocity with distance away from the orifice. Fortunately, we have many opportunities to do this. Each vertical scan line is, effectively, a graph of velocity against distance. One scan line is highlighted here (bright green).

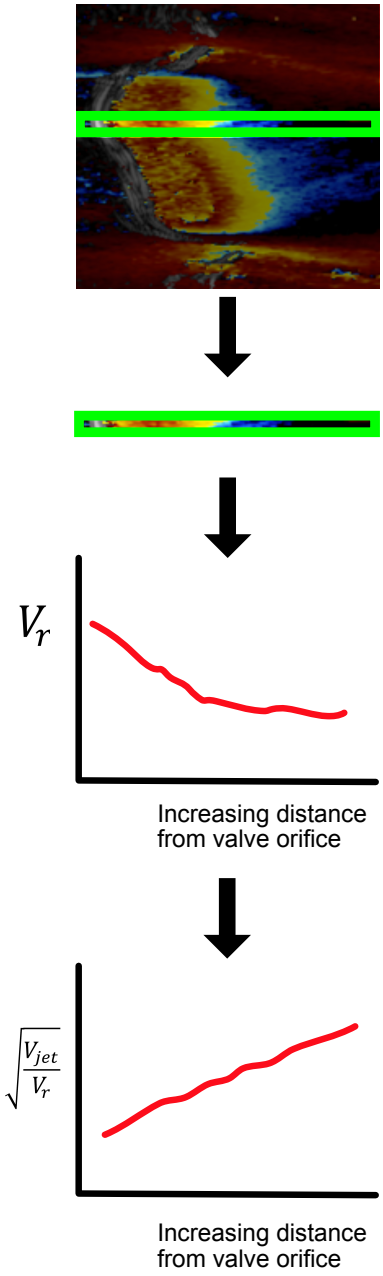
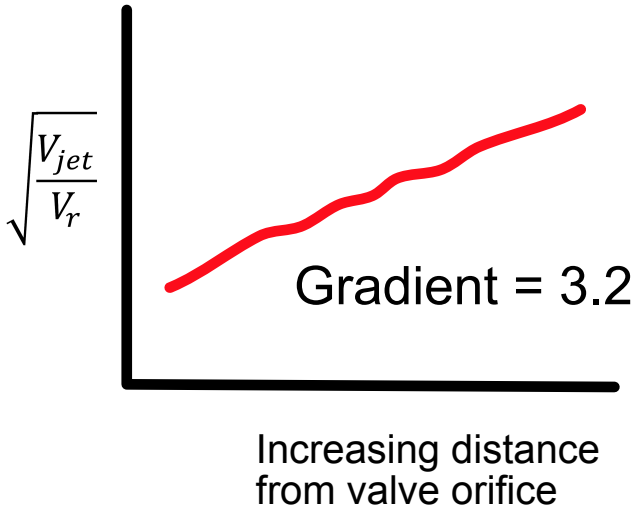


Figure 5-8 The AQURO method - Interpreting a single vertical scanline

This scan line of the colour image contains information about velocity changes in space, at a single instance in time. Somewhere near the left of the line (at an unknown position) is the orifice. The AQURO algorithm avoids having to know where it is: we only have to know that in the blue area, each step 1 mm to the right is 1 mm away from the valve.

Middle panel: From the colours along the single scan line, we can calculate the velocity profile against distance, at that single instant in time. Importantly, while we know absolute velocity at every point, we do not know absolute distance from the valve orifice: only that each 1mm to the right is 1mm further from the valve orifice. We see a velocity declining with distance, along a curved trajectory.

Bottom panel: We quantify that decline by applying a linearizing transform, "take reciprocal, and calculate the square root", which makes it into an upward going line. We can make it even easier by first transforming the velocities by expressing them as a ratio with orifice velocity, so that it becomes dimensionless and easier to handle. Therefore, we calculate $\sqrt{\frac{V_{jet}}{V_r}}$



$$EROA = \frac{2\pi}{gradient^2} = \frac{2\pi}{3.2^2} = 0.61 \text{ cm}^2$$

Figure 5-9 The AQURO method - Calculating the EROA from a single vertical scanline

The average gradient of the line of $\sqrt{\frac{V_{jet}}{V_r}}$ against distance from the valve orifice is calculated (in this example it is 3.2). The EROA is then calculated by "dividing 2 pi by the square of the gradient". In this example the EROA is 0.61cm^2.

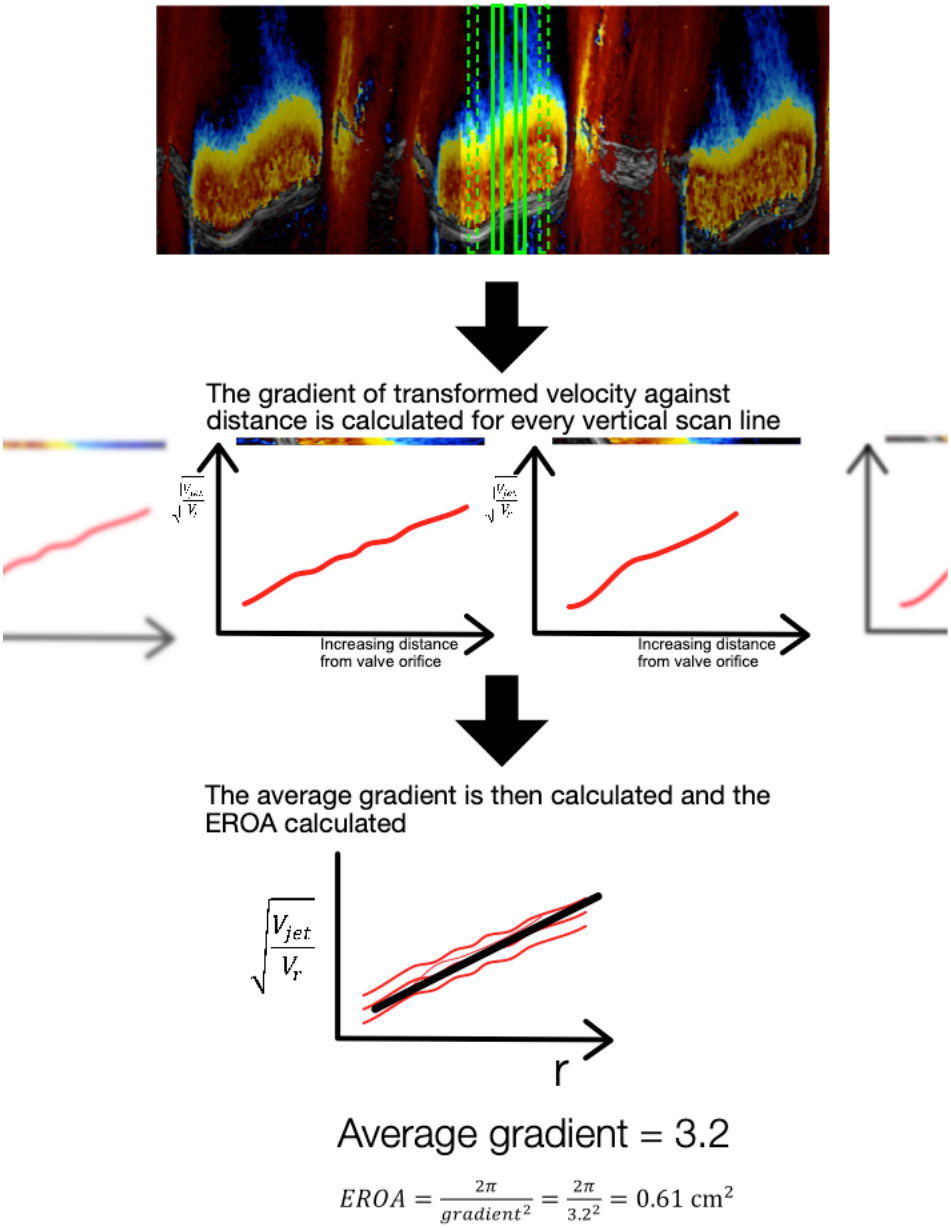


Figure 5-10 The AQURO method - Using data from multiple vertical scanlines to calculate the EROA

In reality the measurement of the gradient is noisy for many reasons. Most important is that the blood flow in this flow convergence zone is not smooth but somewhat wobbly between instants in time. Therefore, the measured orifice area from the method in Figure 5-9 varies, dependent on

the exact scan line we take for processing. We therefore take the opportunity of taking each valid scan line (each instant in time) and not just one picked arbitrarily as in Figure 5-8. The figure above shows two example scan lines, but we could have many (about 20) in that particular heartbeat.

To maximise statistical precision, we average the gradients to calculate a single consensus gradient, and use that to calculate the orifice area.

5.2.5 AQURO: Overcoming the difficulties of PISA

The design of AQURO allows it to overcome of the difficulties I described with PISA

PISA requires a small distance to be measured: With ***AQURO***, no single distance is measured, rather the change in velocity towards the valve automatically extracted from the colour M-mode is used.

PISA requires the sonographer to guess where the regurgitant orifice is: ***AQURO*** does not require knowledge of the exact location of the valve, only the change in velocity towards the valve.

PISA is labour intensive: ***AQURO*** is better defined than PISA as it does not rely on accurately identifying the valve location, therefore automation is easier, reducing sonographer time.

PISA does not maximise the use of information: ***AQURO*** can integrate the data from throughout systole and from multiple beats.

PISA requires eyeballing the diameter: As ***AQURO*** using information from throughout systole, the frame with the largest measurement does not been to be selected.

PISA is very sensitive to the imaging plane: ***AQURO*** provides a visual warning when the acquisition is off-axis. When in the correct imaging plane, blood accelerates in the direction of the valve. When the plane of acquisition does not go through the

valve, the blood will be seen to accelerate then decelerate before the line of acquisition goes through the atria (Figure).

PISA is on colour Doppler is not hemispherical and (8) PISA measurement direction is unclear: By using colour M-mode, the line of acquisition will always be in the direction of the probe. Therefore, with the ***AQURO*** method the sonographer will be making a measurement in the correct direction.

5.3 Methods

5.3.1 The AQURO algorithm

The AQURO algorithm, like the instantaneous Dimensionless Index, has multiple processing steps.

I began with the principles laid out in previous publications and refined many aspects of the algorithm in order to make it fully automated. The key steps of my method are as follows:

Data acquisition: Long, multi-beat runs of colour M-mode through the convergence zone and valve are acquired. In addition, continuous wave data of the mitral jet are also acquired for calculation of the peak velocity

Segmentation of the colour M-mode into individual beats: The colour Doppler strip is segmented into individual beats without reference to the ECG using a neural network.

Calculation of the velocity gradient towards the valve: For each vertical scan-line the velocity gradient towards the valve is calculated.

Quality control of the velocity gradient: The algorithm checks that the relationship linear with sufficient correlation to use the value.

Plotting of the velocity gradient throughout the cardiac cycle using multiple beats worth of data: The velocity gradient is plotted against time throughout the cardiac cycle. As the measurements may be noisy, data from multiple beats are used.

The peak velocity gradient is identified from the curve and converted into the EROA: As the EROA may change throughout the cardiac cycle, the peak EROA is identified on the time-gradient graph and transformed into an EROA.

Integration of AQURO into my easy to use echo quantification software suite.

My specific improvements to AQURO was the addition of the neural network (3) to the segmentation. This enabled (5), tracking the velocity gradient throughout the cardiac cycle, and ensuring that the velocity gradient that corresponded to the peak of the CW trace (6) was identified. I then integrated it into my echo quantification software suite (7).

5.3.1.1 Data acquisition

As with the instantaneous dimensionless index algorithm, multiple, multi-beat strips (8 beats or seconds) are recorded. Colour M-mode through the valve is acquired (Figure 5-11). Separately continuous wave Doppler is acquired of the regurgitant jet.

The benefit of utilising colour M-mode is that it allows the sonographer to see if they are off-axis. If, rather than cutting through the valve, they are instead cutting through a "segment" of the isochronal shells this is easily for them to recognise. Rather than the expected appearance of the blood continuously accelerates towards the valve, they will see instead the warning sign of the blood decelerating close to the valve.

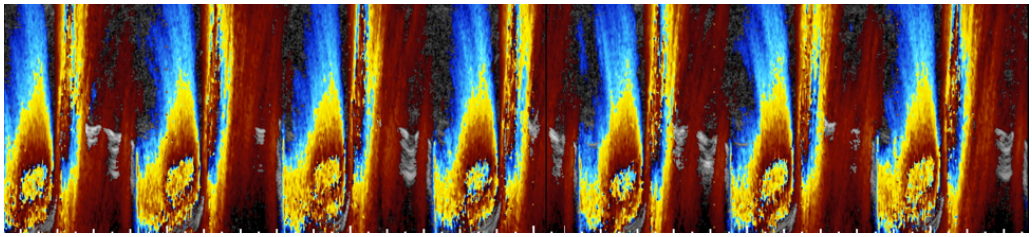


Figure 5-11 Colour M-mode through the convergence zone of a regurgitant mitral valve.

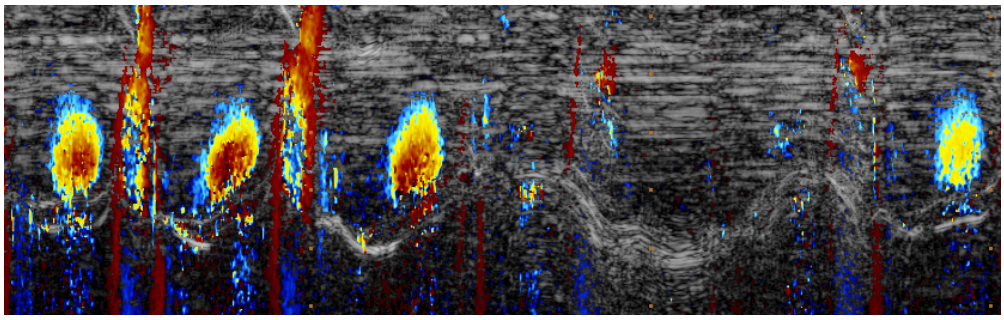


Figure 5-12 Colour M-mode through a convergence zone showing an off-axis beat.

All the beats in this figure are sub-optimal. The third beat might be acceptable because if shows a progressive acceleration with no deceleration. The first and second do show a few pixels of deceleration near the valve orifice.

Where the fourth Doppler beat have been expected there is no Doppler signal, this is because during this beat the beam missed the flow convergence zone completely.

In the final beat, there is almost as much deceleration at the bottom of the Doppler shape, almost as much as the top, indicated a very poor alignment of the beam with the intended axis.

5.3.1.2 ECG-free segmentation of the colour M-mode strip

Even in a stringent research setting, there is no guarantee that the ECG will either be attached or generating a useable trace. Furthermore, with the development of hand-held devices, ECG acquisition would be cumbersome, and is not present on current devices (V Scan, GE; Lumify, Philips).

Consequently, a process for segmenting the colour Mode strip into successive beats of systole and diastole was needed.

Designing an algorithm to segment the colour M-mode strip into individual periods of systole and diastole might, at first glance appear, easy. This is true for the case where the regurgitant jet is clear throughout the systole (Figure 5-13, left panel), however, in more difficult cases, where the convergence zone moves out of the imaging plane, it might be considered more difficult (Figure 5-13, right panel).

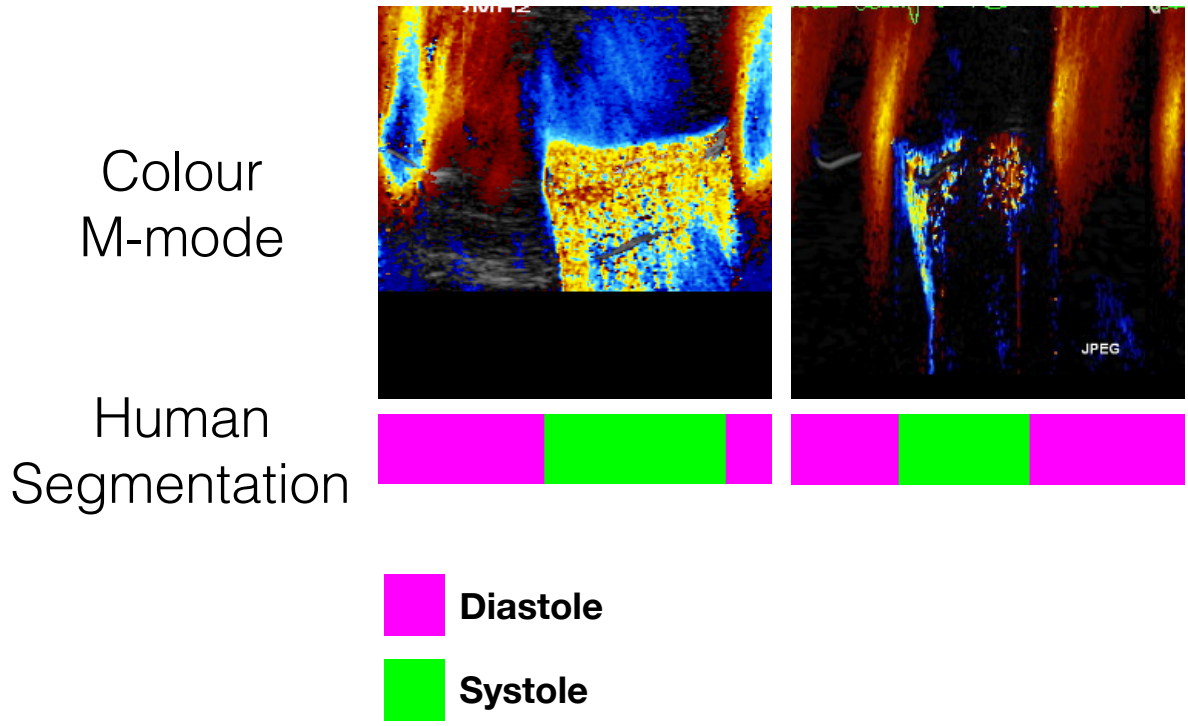


Figure 5-13 The segmentation of two excerpts of a colour M-mode Doppler strip into systolic and diastole

In the left panel, a simple case is shown. The right panel shows a much more difficult example.

5.3.1.2.1 Algorithms and deep learning

There are two types of approaches to evaluating biological signal data to improve its clinical utility. In the first, the physician researcher specifies in detail the steps that should be performed to the data and how these steps should influence the results. This is called algorithmic processing.

Meanwhile, the second approach, machine learning approach has begun to play a progressively bigger role. Machine learning is a branch of artificial intelligence that allows computers to "learn" form data to accomplish a task without being explicitly

programmed in that task. In the machine learning approach, the physician researcher provides a large dataset of raw clinical data and gold standard answers from clinical experts, and sets a process running on a complex algorithm that finds the best way to map the clinical data to the gold standard conclusions.

Traditional algorithms, with hand-crafted feature detectors, are laborious to make, often lack robustness, and when new data is presented that they fail on, are unable to learn from it. I therefore utilised "machine learning" to segment the images.

5.3.1.2.2 Artificial neural networks

There are multiple machine learning algorithms that have been developed over the years, of which artificial neural networks are one example. Whilst many of the other algorithms have been widely used in both research and commercial environments to good effect, much of the recent spectacular successes in complex tasks have come from artificial neural networks.

An artificial neural network takes in image data, represented as a matrix of numbers, and outputs another matrix of numbers that represents the result. In the case of simple image classification, the input may be a matrix of numbers 256 pixels in height, 256 pixels in width, and 3 channels (representing the red, green, and blue information) in depth. The output, if we were aiming to classify a picture as a cat or dog (or perhaps one of 98 other object), would be a matrix of 100 numbers with the first number representing the probability that it is a dog (on a scale of 0 to 1), the second if it were a cat, and so on. In our situation we would like the output to be a

matrix 256 numbers wide (to represent each of the vertical scan lines) and each number represent a probability that the scan line is systole (on a scale of 0 to 1).

The aim of an artificial neural network is to transform the input data to the output data. Each number in the input (256x256x3) would be fed to an input neuron in a single layer. The output layer neurons would be expected to output numbers that represented the result (256x1).

Every input layer neuron could be directly connected to every output layer neuron.

Each neuron in output layer (in this example we only have an input layer and output layer) applies a function to its inputs. This is part of the design of the neural network for a task. A typical function might be "take all the input values, ***weight*** each one of them by a value, ***add*** them together, and ***output*** a value". As part of training the neural network, the aim is to work out what the best ***weights*** are for each of the (256x256x3x256 ~ 50 million) connections. In fact, the neuron does a little more than I have just described and applies an "activation function" to the result to introduce some non-linearity to its output. Examples are a sigmoid function (so transforms a result from -∞ to +∞ to the range 0 to 1 giving a function with an S shape) or the "ReLU" (where negative numbers are changed to 0, and positive ones left unchanged) else the problem would be reduced to regression analysis.

A neural network with a single input and output layer with no hidden layers is called a perceptron (Rosenblatt 1958). There is only one level of connections.

More slightly more complex network could be imagined with a layer of hidden neurons (say 1,000) where every node in the input layer is connected to every neuron in the hidden layer (with each connection needing a weight to be learnt), and then every neuron in the hidden layer being connected to the output ways (and another set of weights needing to be learnt). It has been proven that, with a finite number of neurons, a network with only a single hidden layer can approximate any function (with some assumptions and restriction). This is referred to as the *universal approximation theorem* (Cybenko 1989; Hornik 1991). I.e. a network of this architecture should be sufficient for our needs (should we have the techniques to train it to work out what the weights should be). This architecture is called a multilayer perceptron (Rumelhart et al. 1986; Werbos 1994).

5.3.1.2.3 Deep learning

For our task, rather than use a single hidden layer, we will use multiple. It is this multi-layered approach that "deep" refers to, and well covered by reviews (LeCun et al. 2015). Whilst the addition of more layers makes training more difficult, it allows the network to learn more abstract representations of the data, and how features interact with each other.

5.3.1.2.4 Layers

Within the neural network I designed to classify scan lines, the connections between the layers are different. I have previously described "fully-connected" layers – by which I mean all the neurons in the previous layer are connected to all the neurons

in the next layer. This has multiple problems. Firstly, if each neuron in a layer is connected to the next, then the number of weights that would need to be learnt with perhaps 10s of layers would be insurmountable. Secondly, it fails to take into consideration the importance of local features. The information from surrounding pixels is far more useful than from distant pixels.

Therefore, rather than full-connected layers, the vast majority of image processing neural networks utilise ***convolutional*** layers (Lecun et al. 1998). A ***convolutional*** layer only looks at say the equivalent pixel in the layer before, and its immediate surrounding pixels (e.g. 3 for 3x3 pixels). Within each layer there will be many "channels", these can be thought of as each representing a filter (perhaps 16 or 32 in each layer). This means that for each layer perhaps only 2304 (9 input pixels * 16 input channels *16 output channels) weights might need to be learnt. This is a dramatic reduction as compared to the original 200 million (256*256*3*1000) that might be required for the first hidden layer if it were fully-connected. Sharing the weights in this way means that once the network learns a feature (perhaps a bit of red Doppler before a bit of blue Doppler) it will recognise this anywhere in the image.

However, as the aim of the network is to learn global abstract representations the ability to look at larger patches of pixels is required. Therefore, after a few convolutional layers, a layer that reduces the image dimension by pooling pixels (max-pool) are introduced (Nagi et al. 2011).

5.3.1.2.5 The AI (deep learning) boom

Despite the development of the perceptron in 1957, until recently there was little real-world use of neural networks, with other forms of machine learning, such as support vector machines, dominating. The recent boom in AI work, research use, and release of development frameworks has been sparked by many factors. Some of it relates to the development of better architectures (such as Resnets (He et al. 2016), Inception modules(Szegedy et al. 2015)), and methods (such as different activation functions, batch normalisation (Ioffe & Szegedy 2015)). However, the biggest advance has been in the availability of compute parallelised compute power provided by the widespread availability of relatively cheap GPUs (Raina et al. 2009). Originally designed for rendering 3D scenes for scientific and engineering purposes, the mass market of gamers drove a rapid increase in power at lower prices.

This is essential for deep learning as whilst using a neural network is relatively computationally easy, training it requires a large amount of computational resources. The reason for this is that the method by which these neural networks are trained.

5.3.1.2.6 Training a neural network

Training a network involves working out what value to set the weights to so that various inputs result in the correct output. This is an *optimisation* problem. Mathematically, some problems have an analytical solution. For example, consider the equation $output = weight \times input + 2$. It is easy to rearrange this to tell us what

weight is needed, given just one correct example of input and output. This is called an analytical solution. In this example it is: $weight = \frac{output-2}{input}$. For instance, if one knows that an input of 3 should produce an output of 8, then one knows that the required weight is (8-2)/3 = 6/3 = 2.

However, within a neural net, there is no simple analytical solution. An alternative to an analytical solution is an exhaustive search through all possible combinations. An analogy of this is trying to open a bicycle lock by trying all combinations of three digits. In a neural network there are many millions of weights (equivalent to dials on a bicycle lock), and therefore an exhaustive search is impossible.

Therefore, an *iterative optimisation* solution called *stochastic gradient descent* is used (Bottou et al. 2018). It relies on the neural network being *differentiable*. Test input is fed into the neural network and the output inspected. The difference between the actual output and the desired output is the error is calculated using a cost function. With training, the algorithm uses knowledge of the error derivative to work out in what direction and what way to nudge each weight to make it more likely to produce the desired output. The algorithm works backwards through the network, each layer at a time.

This process is repeated many, many times. The amount of training can vary, but typically, each image in the training set may be presented (as part of a small batch) perhaps 50 or more times (epochs).

5.3.1.2.7 The architecture of my neural network for segmenting M-mode colour Doppler

The neural network I designed is based on a series of convolutional layers. In between are max pooling to reduce the image dimensions and make the learnt features relatively invariant to translation and scale. It takes as input a 256 x 256 patch of colour M-mode Doppler, and outputs a strip of 256 x 2 pixels. This represents a classification for each of the 256 scanlines. For each scanline, there are two pixels, one representing plausibility of this scan line being in systole, and the other the plausibility of diastole. I used a Rectified Linear Unit (ReLU) activation function with batch normalisation after each convolution layer (aside from the last). The layer architecture is summarised in Table 5.1.

Table 5.1 Architecture of colour M-mode Doppler neural net

Layer	Output Size
Input	256 x 256 x 3 channels
Convolution (3 x 7 x 64 channels)	256 x 256 x 64 channels
Convolution (3 x 7 x 64 channels)	256 x 256 x 64 channels
Max pool (4 x 1)	64 x 256 x 64 channels
Convolution (3 x 7 x 128 channels)	64 x 256 x 128 channels
Convolution (3 x 7 x 128 channels)	64 x 256 x 128 channels
Max pool (4 x 1)	64 x 256 x 128 channels
Convolution (3 x 7 x 256 channels)	16 x 256 x 256 channels
Convolution (3 x 7 x 256 channels)	16 x 256 x 256 channels
Max pool (4 x 1)	4 x 256 x 256 channels
Convolution (3 x 7 x 512 channels)	4 x 256 x 512 channels
Convolution (3 x 7 x 512 channels)	4 x 256 x 512 channels
Max pool (4 x 1)	1 x 256 x 512 channels
Convolution (3 x 7 x 1024 channels)	1 x 256 x 1024 channels
Convolution (3 x 7 x 1024 channels)	1 x 256 x 1024 channels
Convolution (3 x 7 x 2 channels)	1 x 256 x 2 channels

5.3.1.2.7.1 Training Data

For training, I acquired 93 multi-beat strips of colour M-mode Doppler from 14 patients. These were labelled by hand to define systole and diastole.

The neural network was trained with 290 batches of data, each mini-batch containing one 256 x 256 image sample from each of 16 Doppler strips. The selection of the 256x256 image sample was random so that when a particular strip was used again, a different part was presented to a neural network. Because there were 290 batches of 16 image samples, i.e. 290*16 = 4640 image sample presentations, each of the 93 image strips was ultimately being presented 4640/93 ~ 50 times.

5.3.1.2.7.2 Validation

The validation data was obtained from a further 34 strips from 5 different patients who were not part of the training dataset. After training, on a vertical scanline-by-line basis a classification accuracy of 97% was achieved.

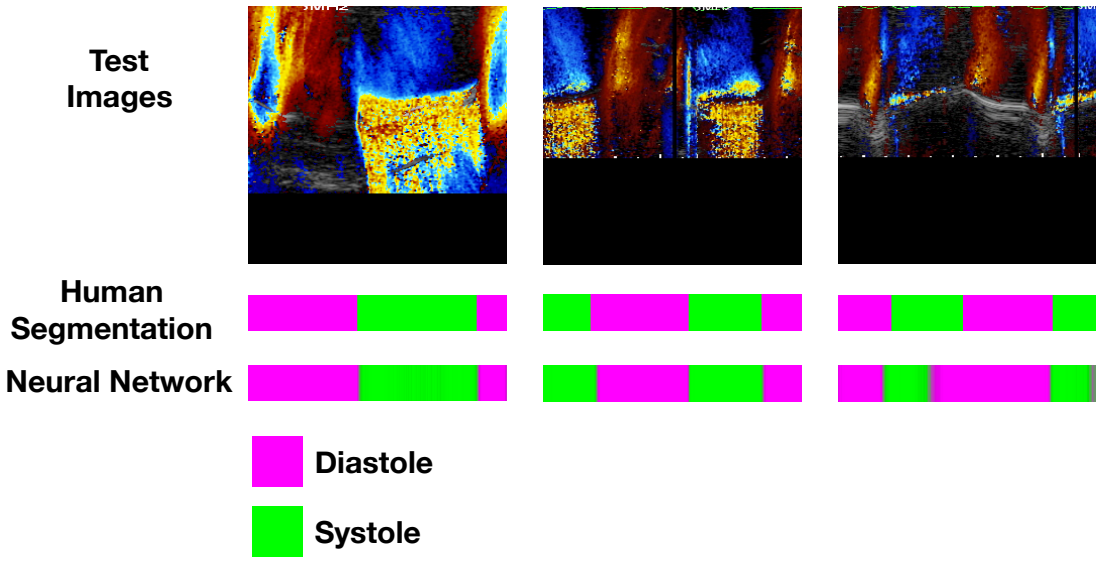


Figure 5-14 Results after training the neural network on the validation set of images.

Two almost perfect results are presented (left, centre), and one acceptable result from the test image dataset (right). Overall, on a vertical scan-line basis there was a 97% accuracy.

5.3.1.3 *Identifying the aliasing point*

The colour M-mode strip is segmented using the neural network developed in the previous section. The aliasing point is detected in each vertical scan line during systole is detected. Unlike with PISA, the aliasing point is *not* used for the calculation, but to localise the region of useful information on the colour M-mode strip. Some scan lines may appear to have multiple aliasing points due to noise. Ridge detection is used to find the continuous aliasing point throughout the vertical scan lines. This is shown in Figure 5-15

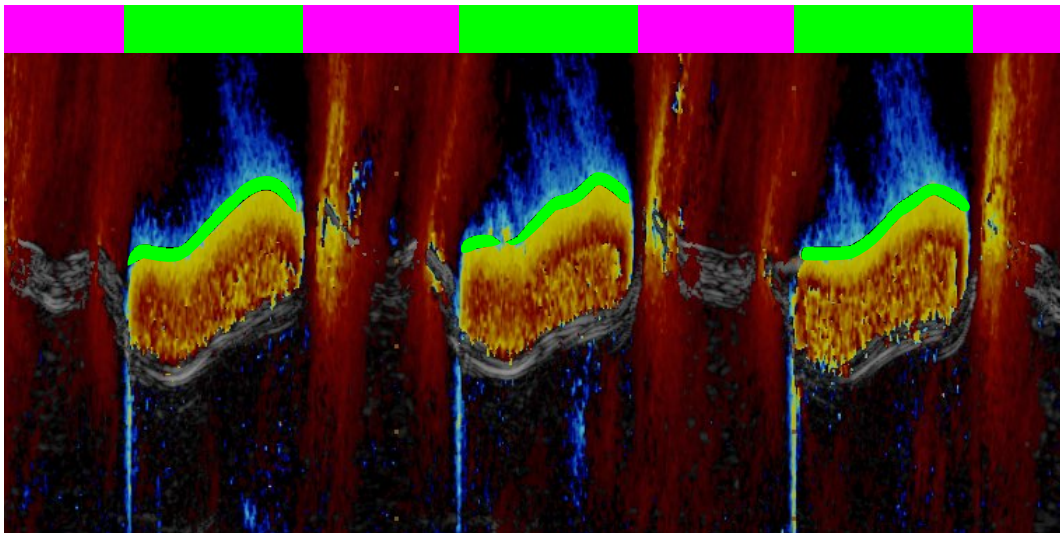


Figure 5-15 Segmentation of the colour M-mode strip into systole and diastole and identification of the aliasing point.

Systole (green) and diastole (purple) are highlighted above the strip, and the aliasing point highlighted in green.

5.3.1.4 *Calculation of the velocity gradient towards the valve*

Along each vertical scanline the velocity gradient (from a graph of $x = r, y = \sqrt{\frac{V_{jet}}{V_r}}$) is plotted. The Peak For each scan line, the region of the graph which is most linear (as assessed using the Pearson correlation coefficient) is found. If no part has a correlation of >=0.8, the vertical scan line is discarded. These linear segments are indicated on the strip in Figure 5-16

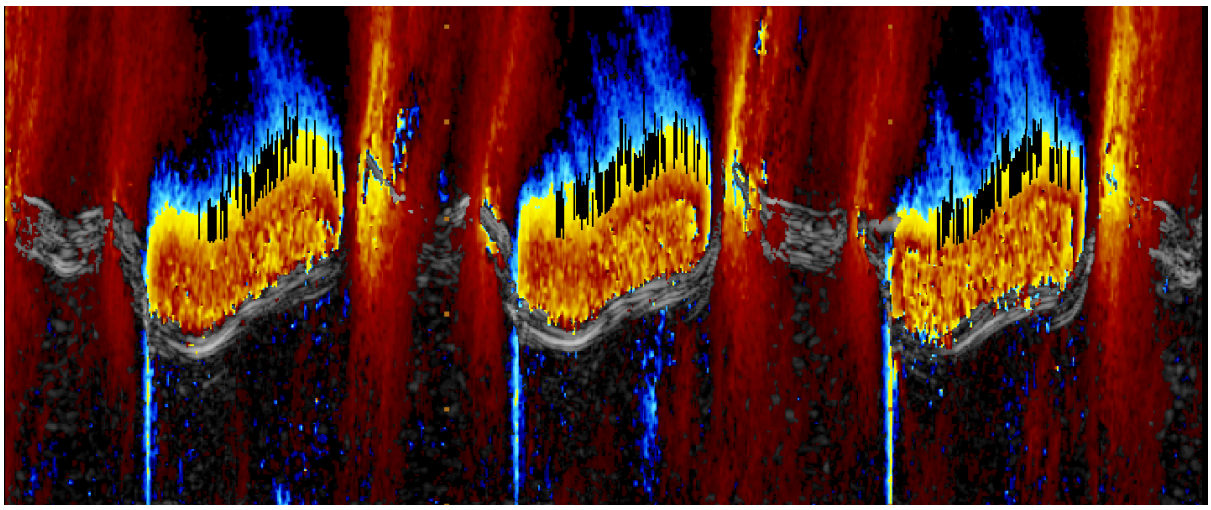


Figure 5-16 Segments of vertical scanlines showing most linear change in $\sqrt{\frac{V_{jet}}{V_r}}$ *with distance along the vertical scanline*

5.3.1.5 Calculation of the EROA

The beats are aligned at the onset of systole, and the gradients plotted (Figure 5-17). Robust regression (to cope with outliers) is used to fit a cubic spline with 5 knots through the data. The maximal velocity gradient (which corresponds to the maximal EROA) is then identified. As the velocity gradient is dependent on the peak velocity of the regurgitant jet (V_{jet}) only the maximal velocity gradient, coinciding with the peak mitral regurgitation velocity, represents the true EROA. Knowing that any change in the EROA occurs smoothly, by fitting a curve throughout systole it allows the algorithm to make use of all the available data.

Once the peak velocity gradient is identified, the EROA is calculated using the formula $EROA = \frac{2\pi}{gradient^2}$

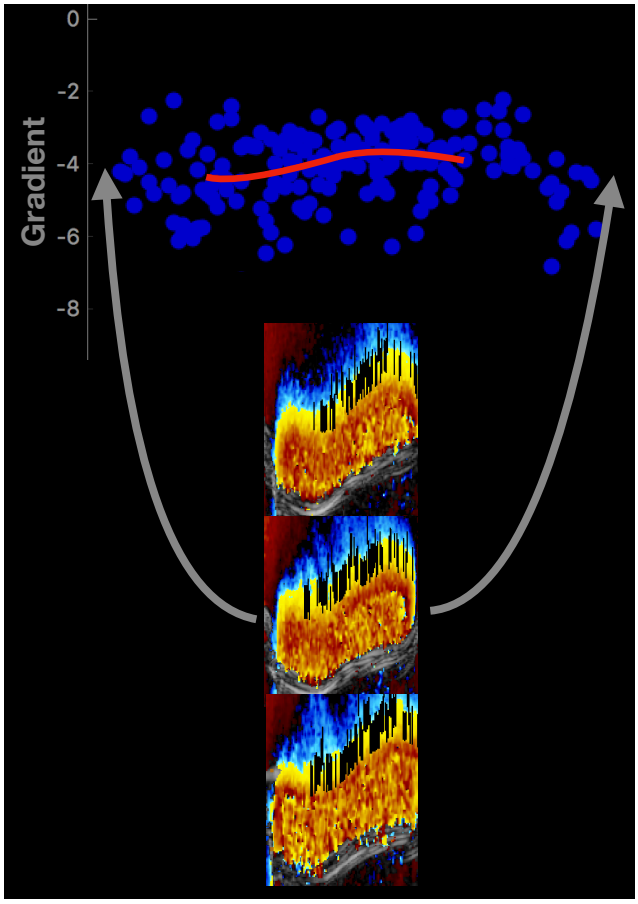


Figure 5-17 Plot of the velocity gradients over time

The segmented beats are aligned using the onset of systole. The gradients are plotted over time (x-axis). Using robust regression a cubic spline with 5 evenly spaced knots is fitted over the middle 70% of the data. The peak gradient is identified and used to calculate the EROA.

5.3.2 Study Population

For the pilot validation study, patients at a single institution (Imperial College NHS Trust, London) with mild, moderate, or severe primary mitral regurgitation in a native valve in sinus rhythm were recruited. Patients with bioprosthetic, mechanical, or repaired valves were excluded. All patients gave written, informed consent to

participate in this study which was approved by a local ethics committee (14/LO/1520). Aside from inability to give consent, there were no exclusion criteria.

5.3.3 Image acquisition

Echocardiographic data were acquired using a Philips IE33, Philips Affinity 70, or Philips Epic 7. Data were processed off-line.

A single echocardiographer performed a full study in line with British Society of Echocardiography guidelines (Wharton et al. 2015). Images were then acquired specifically for this study. At least 3 acquisitions suitable for AQURO analysis (multiple-beat colour M-mode images through the mitral valve) were performed, each time removing the probe from the patient's chest. Similarly, at least 3 acquisitions suitable for PISA analysis were obtained, again removing and replacing the probe between each acquisition. Finally, a continuous wave acquisition for the peak mitral regurgitant jet velocity was obtained as both the AQURO and PISA EROA analysis require this.

5.3.4 Image analysis

Images were analysed offline. AQURO EROA measurements were made using the software developed earlier in this Chapter. PISA EROA measurements were made using Horos (The Horos Project 2018).

Three sets of measurements were obtained in each patient.

1. Three echocardiographers independently reviewed the echocardiogram blinded to the PISA and AQURO data and provided a visual estimate of the severity of the mitral regurgitation. They selected on a scale of 1 to 5, representing mild, mild-to-moderate, moderate, moderate-to-severe, and severe.

2. A single AQURO EROA value was calculated for each of the multiple-beat colour M-mode acquisitions to provide at least 2 AQURO derived EROA measurements for each patient.

3. A single PISA derived EROA value was calculated for each of the colour Doppler acquisitions to provide at least 2 PISA EROA measurements for each patient. Measurements were performed by a single reader.

Both the AQURO and PISA EROA calculation require the maximum velocity of the regurgitant jet. For consistency, a single value was obtained in each patient using the automated software developed in Chapter 4 by taking the average of a multiple-beat CW acquisition. This value was then used in all AQURO and PISA calculations.

5.3.5 Feasibility and data exclusion

All patients with interpretable PISA images were included in the cohort. If the AQURO derived EROA was able to be calculated from at least one recording it was considered feasible in that patient. Only patients with both two AQURO and PISA derived EROAs were included in the assessment of agreement and reproducibility.

5.3.6 Data analysis

As previously shown (Chapter 2), amongst imaging biomarkers, the assessment of the severity of mitral regurgitation is particularly variable. As a biomarker cannot correlate better with another than it can with itself, I averaged multiple measurements (using a bootstrap confidence intervals without assuming normality) to allow comparisons between the different methods to better represent the underlying relationship.

The visual estimates of severity (on a 1 to 5 scale) from the three readers were averaged to get a representative visual estimate. Similarly, the AQURO and PISA measurements were averages across the acquisitions. Inspection of the data showed that the variability was highly correlated to the magnitude (i.e. at small EROAs the absolute variability was less than at larger EROAs), the geometric mean used. This was calculated by taking the exponential of the arithmetic mean of the logarithms of the EROA measurements.

Relationships between the AQURO and PISA derived EROAs and the visual assessments were assessed using an ordinal regression models (which requires fewer assumptions than the commonly used linear model estimation using ordinary least squares). A multivariate generalised linear mixed model with a Markov Chain Monte Carlo sampler was used to calculate the within-patient and between-patient variances and coefficient of variation. The intraclass correlation coefficient was

calculated using these obtained variances. Credible intervals and differences between these derived parameters were obtained via calculating the posterior mode and 95% highest posterior density interval of the modified samples.

Agreement between AQURO and PISA derived EROA was assessed using the methods of Bland-Altman (Bland & Altman 1986) on the logarithmic transformed EROAs. The exponential function was used to back-transform the results for interpretation.

Data was processed using Python (Python Software Foundation 2017), and statistical analysis using the statistical environment R (R Core Team 2016) with the "tidyverse" collection of packages (Wickham 2017) and "MCMCglmm" linear modelling package (Hadfield 2010).

5.4 Results

5.4.1 Patient population

41 patients were recruited for this initial pilot validation study. 3 patients were excluded as on review of the data, only one PISA EROA acquisition was useable. Of the remaining 38 patients, 2 patients were excluded as the AQURO analysis failed. In one, the beat-by-beat segmentation algorithm failed. In another, the colour M-mode strip also captured flow acceleration around a septal bulge and mistook it for the flow acceleration near the mitral valve. In the latter case, more diligent use of zoom would have prevented this failure. Therefore, AQURO EROA was feasible in 95% (36/38) of the patients. Overall 36 patients were included in the final dataset.

5.4.2 Relationship between the AQURO and PISA derived EROA and visual assessment

The relationships between the AQURO and PISA derived EROA and the visual assessment of mitral regurgitation severity are shown in Figure 5-18 and Figure 5-19 respectively. There was a significant relationship between visual severity and the EROA derived by both AQURO ($p=0.003$, $rho=0.613$) and PISA ($p=0.0008$, $rho=0.617$). Whilst all (4/4) patients with a consensus visual severity of severe regurgitation all had an EROA greater than 40mm^2 derived either from AQURO or PISA, some patients with mild-to-moderate, moderate, and moderate-to-severe also met this criterion (3/9, 2/8, 7/10; and 2/9, 3/8, 5/10 respectively). For both AQURO and PISA derived EROA, there was clear differentiation between visually mild and severe

disease, but substantial overlap within the mild-to-moderate, moderate, and moderate-to-severe categories.

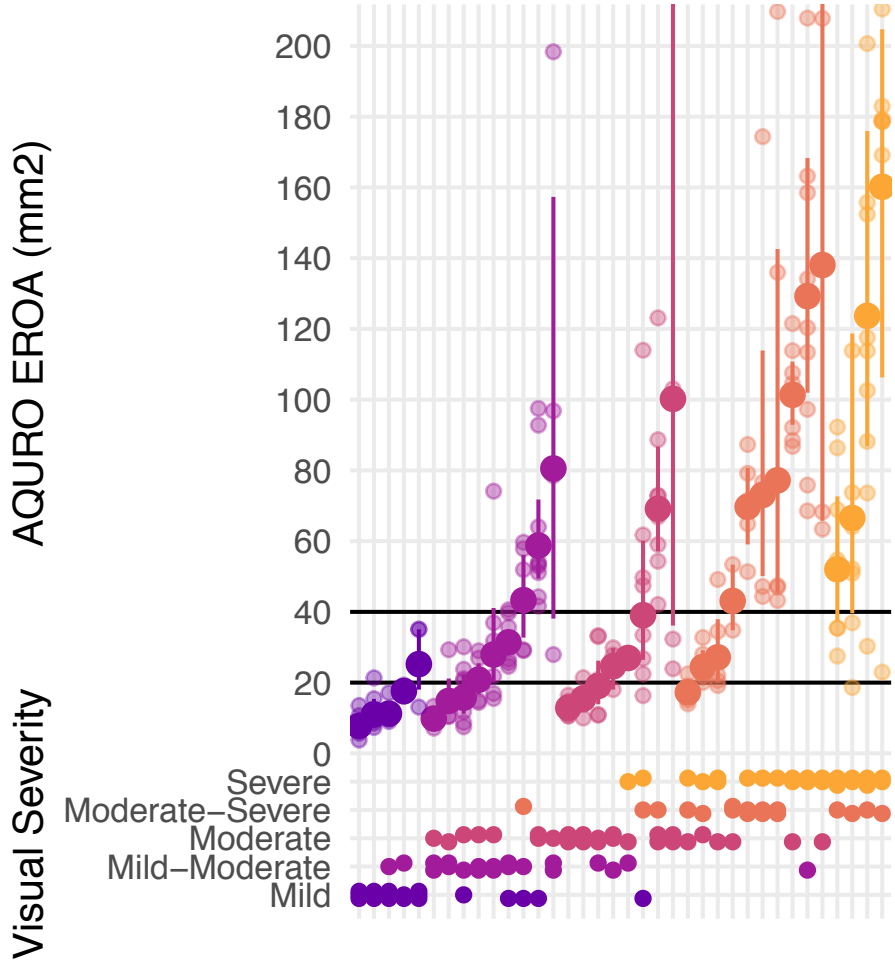


Figure 5-18 Relationship between AQURO derived EROA and the visual mitral regurgitation assessment

Each vertical line represents data from a single patient. Patients are first grouped by the visual severity (from left to right mild to severe) and then by the average AQURO derived EROA. The data at the top of the graph shows the individual AQURO derived EROAs for each patient, and their geometric mean and 95% confidence interval of the mean. The data at the bottom of the graph shows for each patient the visual assessment of severity from each of the three readers. The horizontal lines at 20 and 40 mm^2 reflect the current ESC boundaries between mild, moderate, and severe.

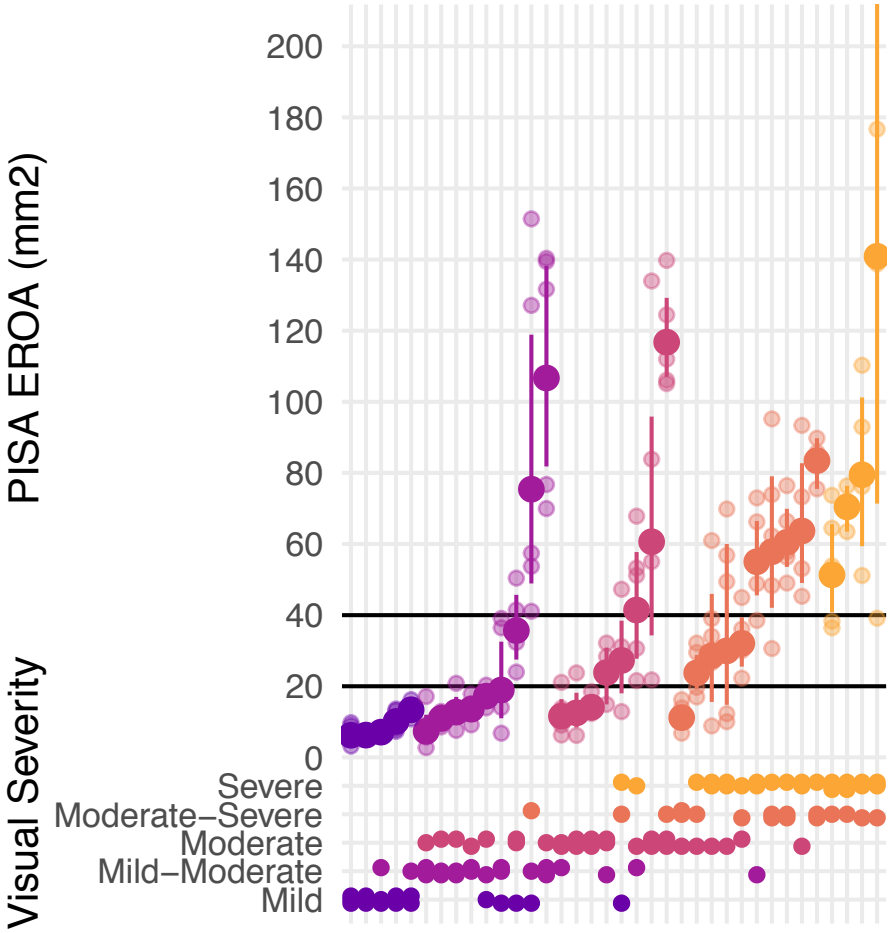


Figure 5-19 Relationship between PISA derived EROA and the visual mitral regurgitation assessment

Each vertical line represents data from a single patient. Patients are first grouped by the visual severity (from left to right mild to severe) and then by the average PISA derived EROA. The data at the top of the graph shows the individual PISA derived EROAs for each patient, and their geometric mean and 95% confidence interval of the mean. The data at the bottom of the graph shows for each patient the visual assessment of severity from each of the three readers. The horizontal lines at 20 and 40 mm^2 reflect the current ESC boundaries between mild, moderate, and severe.

5.4.3 Reproducibility of AQURO and PISA derived EROA

Using a mixed effects model, the immediate within-patient coefficient of variation was 67% for AQURO and 57% for PISA derived EROA. Whilst these numbers for reproducibility seem large, they must be compared to the large scale over which EROA values extend. Unlike, for example, peak aortic valve velocities may typically range from 1 to 4 (a 400% difference); calculated EROAs may extend from 5mm^2 to over 100mm^2 (a 2,000% difference). In this dataset, the mean EROA extended from 7mm^2 to 130mm^2.

The intraclass correlation coefficient (ICC) is the ratio of the measurement variance to the measured population variance and therefore puts the measurement variability in the context of the range of values obtained from the population. The ICC for the AQURO derived EROA was slightly smaller than that of the PISA (0.74 vs 0.80, difference -0.7, 95% credible interval -0.21 to 0.06).

5.4.4 Relationship and agreement between AQURO derived EROA and PISA derived EROA

Figure 5-20 shows the relationship between the average (geometric mean) of the AQURO and PISA derived EROA on a log-log scale. There was a significant correlation between AQURO and PISA EROA ($p<0.0001$, rho=0.814). However, AQURO values were, on average, 34% larger than the PISA values ($p=0.003$).

The Bland-Altman 95% limits of agreement were from -55% to 299%. Again, this large interval must be put in context of the large range over which EROAs extend.

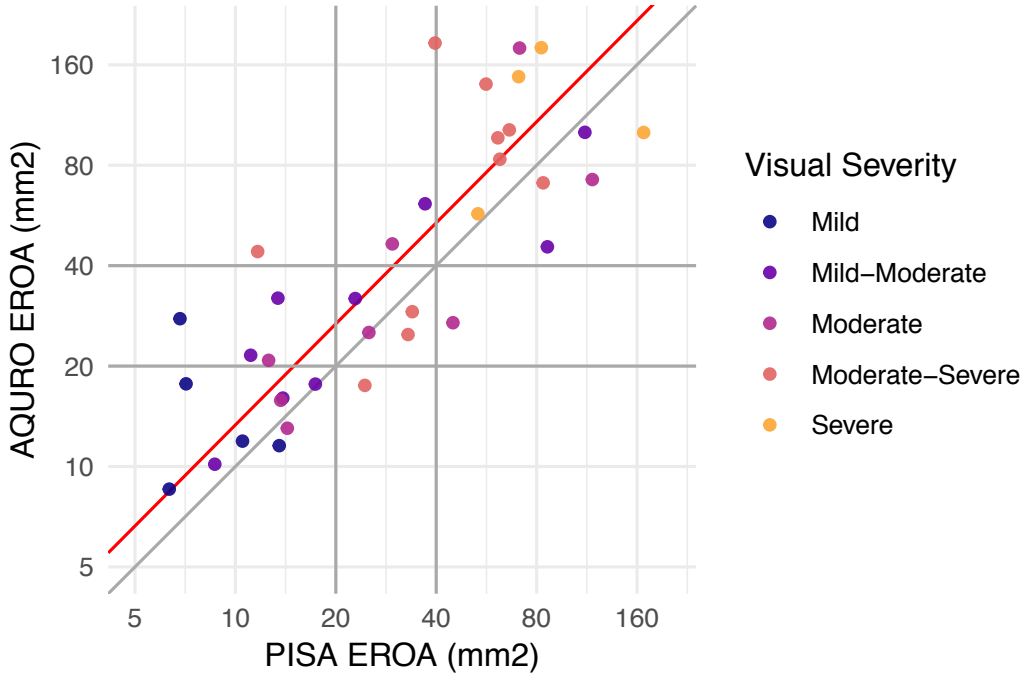


Figure 5-20 Relationship between AQURO and PISA derived EROA

For each patient, the geometric mean AQURO derived EROA and PISA derived EROA are plotted. The dark grey lines indicate the ESC guideline cut-offs for mild, moderate, and severe mitral regurgitation. The principal component regression line is red. A log-log scale is used.

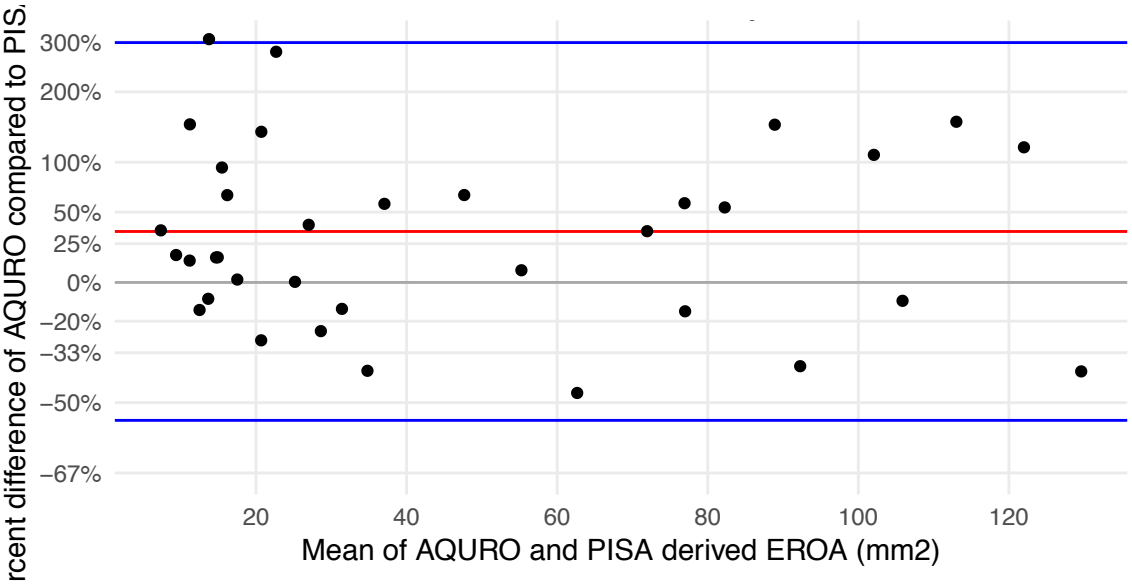


Figure 5-21 Bland-Altman plot of AQURO and PISA derived EROA

The geometric mean of the AQURO and PISA derived EROA are plotted on the x-axis, and the percent difference on the y-axis. The mean bias is indicated by the red line, and the 95% Limits of Agreement by the blue lines. Whilst numerically the Limits of Agreement are large, compared to the range of EROAs they are modest.

5.5 Discussion

5.5.1 Main findings

In this Chapter, I have found:

- AQURO derived EROA in mitral regurgitation is feasible in the majority of patients. It provides an automated method of calculating the EROA from multiple-beat acquisitions of colour M-mode.
- AQURO provides values 30% larger than PISA derived EROA.
- AQURO has a similar precision to PISA derived ERAO.

Within this thesis I have provided an open-source automated method of quantifying mitral regurgitation. Whilst it provides comparable measurements to PISA with a similar Intraclass Correlation Coefficient (ICC), it did not exceed the precision of PISA. This need not be a significant limitation as automation allows for the easy collection of much more data. With the collection of multiple acquisitions, the coefficient of variation will fall with the square-root of the number of acquisitions collected. For example, with 5 acquisitions (that would take no more than a minute), the coefficient of variation would be expected to be reduced from 67 to 26% and the ICC increase would be expected to increase from 0.74 to 0.93.

5.5.2 Implications of imprecision in EROA assessment

The large coefficient of variation of a single measurement (as typically performed in clinical practice) of both AQURO and PISA derived EROA might make a trained

metrologist have second thoughts about using the results determine a patient's surgical fate. In some ways, the biomarker is redeemed as the spread of EROA values is large. Therefore, the ICC is a respectable 0.74 when a single measurement is made (and with 5 repetitions is 0.94). However, all this means that you can be reasonably confident with "mild" and "severe" cases. Within EROAs within the narrow defined moderate range (20 to 40mm^2) significant diagnostic uncertainty remains.

It might be argued that this study is not representative. Perhaps there was a problem with the acquisition or interpretation technique. However, the findings are consistent with much cited validation study by Dujardin et al. (Dujardin et al. 1997) that defined the PISA derived EROA cut-points used for grading the severity of mitral regurgitation. In this study they compared the angiographic grade of mitral regurgitation to the PISA derived EROA. In Figure 1 in their paper, they produce a very similar plot to Figure 5-18 of this thesis. In their plot – at all angiographic grades of mitral regurgitation there were patients with a PISA derived EROA that were "mild", "moderate", and "severe". Indeed, there was significant overlap between the EROAs of angiographic grade "I" and "IV".

5.5.3 Calibration

AQURO derived EROAs were on average larger than PISA derived EROAs. This is to be expected from their respective formulations. An acquisition that does not cut the convergence zone through the centre of the orifice will have differential effects

on the AQURO and PISA. The convergence shell will appear smaller than it truly is. This will lead to a smaller PISA radius, and hence a small PISA estimate of the EROA. Conversely, because the PISA shell misses the point of maximal velocity, the acceleration through the convergence zone will appear to be slower than it truly is. This will lead to an overestimate of the AQURO derived EROA. Therefore, whilst PISA will tend to underestimate the EROA from mis-aligned images (assuming everything else is done correctly), AQURO will tend to overestimate. The reduction in acceleration apparent as the orifice is approached with AQURO can be detected both visually on the colour M-mode and an automated warning to the operator could be added.

5.5.4 Limitations

5.5.4.1 AQURO limitations

There are intrinsic limitations in using a 1-dimensional measurement from a 2-dimensional image to calculate the properties of a 3-dimensional jet and orifice. AQURO and PISA share this limitation. They rely on the assumption that the convergence zone is hemispherical, and the orifice circular. If the regurgitant zone becomes elongated from a small circle to a wider area along the line of coaptation both AQURO and PISA will underestimate the true area.

AQURO and PISA require two measurements that would ideally be undertaken simultaneously. With the instantaneous Dimensionless Index, I was able to derive methods to automatically capture these two parameters simultaneously which

contributed to a substantial reduction in variability. However, for mitral regurgitation, due to the large difference in velocities between the convergent zone and the jet this was not possible. If modifications could be made to the echocardiographic machine, interleaving the pulsed-wave and continuous wave scan lines is a possibility. Atrial fibrillation would exacerbate this lack of simultaneous acquisition.

5.5.4.2 Study limitations

This was a small pilot study. Future validation work would requires it to be conducted at an independent institution.

To maximise the ability to assess the correspondence between the two techniques, biological noise was minimised by selecting patients in sinus rhythm. Many patients with mitral regurgitation have atrial fibrillation; in these patients the coefficient of variation of both AQURO and PISA derived EROA would be expected to be larger than found in this study.

Similarly, to minimise other sources of variation, all patients had primary mitral regurgitation. Again, future validation work should include patients with secondary mitral regurgitation.

Like other studies, this study is limited by the lack of a "gold-standard" for the assessment of mitral regurgitation. When there are large discrepancies between either the ordinal classification and the calculated EROA, or between AQURO and

PISA, we are limited in our ability in understanding which is "correct" and which is "incorrect", and hence modify the algorithm. Cardiac MRI is one possibility, but this measures flow rates and regurgitant fractions. Whilst some guidelines advocate multiplying the calculated EROA by the VTI of the regurgitant jet to derive the flow – this makes a further assumption that the regurgitant orifice opens and closes instantly and risks introducing further errors to the calculation.

5.5.5 Implications for clinical practice

The coefficient of variation of echocardiographic EROA assessment is large. Clinicians must be aware of this when serially assessing patients. One way to overcome measurement variability is to average multiple measurements. Automation, using methods such as AQURO enables this, allowing the coefficient of variation to roughly fall with the square-root of the number of measurements made. If AQURO is not used, multiple measurements are still required to reduce the variability to an acceptable level.

Clinicians should be reminded that, as will all biomarkers, diagnosis at the extremes is easy as there was clear separation between mild and severe cases. It is those that are in the middle of the biomarker range that are the most challenging.

5.5.6 Implications for research

Researchers using EROA as an inclusion criterion to clinical studies must be aware that because of the variability of EROA that the moderate range (20 to 40mm^2) of EROAs is compatible with a visual severity that ranges from a little more than mild

to severe mitral regurgitation. For example, two trials of transcatheter mitral valve repair (MitraClip) in patients with mitral regurgitation and heart failure have been recently published. Despite comparable end-points (including death, and hospitalisation), there was a significant discrepancy between the results with one trial (COAPT, Stone et al. 2018) was strongly positive, and the other (MITRA-FR Obadia et al. 2018) neutral. They differed significantly in their inclusion criteria. COAPT used an EROA cut-off of 40mm^2 and MITRA-FR used 20mm^2 as indicating severe mitral regurgitation. As discussed, based both on the work in the paper, and Figure 1 of the original PISA validation study (Dujardin et al. 1997), using a cut-off as low as 20mm^2 is liable to include patients that on visual estimation would be considered little more severe than mild mitral regurgitation.

Studies investigation changes in the severity of mitral regurgitation as a secondary result of another intervention (such as aortic valve intervention or CRT) should ensure multiple replicate EROA assessments are made due to the significant variability.

5.6 Conclusion

AQURO provides an automated calculation of the mitral regurgitant EROA from colour M-mode Doppler. It provides results comparable to PISA with a similar single measurement variability. Through automation it will allow the capture of much more data, allowing precising to be increased.

6 Synthesis

My thesis addresses several dimensions of variability of echocardiographic assessment of aortic stenosis and mitral regurgitation:

Correct recognition of the existence of the problem (Chapter 2)

Developing a practical approach to measuring it, and a clear way to represent it to clinicians (Chapter 3)

Creating a freely-downloadable Open-Source package of software, the instantaneous Dimensionless Index, to solve or reduce several of the major sources of variably in the assessment of aortic stenosis. This was then tested in my hands, and in the hands of an independent colleague. I found it to reduce the variance more than four-fold (Chapter 4).

Developed and improved AQURO, an automated method of calculating the EROA in mitral regurgitation from colour M-mode Doppler images. I performed a pilot validation study finding it to provide, in an automated method, comparable results and precision as PISA derived EROA(Chapter 5).

6.1 Recognising that the literature on reproducibility is almost worthless

If there is a useful summary of this, it is that the existing reproducibility literature within echocardiography can only be used like the proverbial drunk uses a lamp-post – for support rather than illumination. The most charitable interpretation of the

large discrepancy between test-retest reproducibility of the dedicated studies against the placebo-arm of RCTs is that the interval between assessments was on average much shorter than within RCTs (median 1 day vs 6 months). Even if this interpretation is correct, it means that reproducibility studies are failing to study the largest source of variation (Chapter 2).

6.2 Targeting the true sources of variability

Understanding reproducibility and the sources of noise is hampered by the difficulties of conducting and analysing large scale studies within the two-operator paradigm. In Chapter 3 I develop and use a new framework to investigate the sources of noise in the quantification of aortic stenosis. This allowed me to develop a targeted approach in the following chapter. I found that the subtle, inevitable variations in probe placement are the greatest source of noise in the quantification of aortic stenosis severity. This is in contrast to the traditionally, and as it turns out, unfairly blamed sonographer and reader.

6.3 Designing and validating a new biomarker – the instantaneous Dimensionless Index – with the aim to improve reproducibility

Chapter 4 sets out the design and rational for the instantaneous Dimensionless Index. By simultaneously estimating the peak trans-aortic and LVOT velocity, it eliminates the beat-by-beat variability and most of the probe placement variability.

Furthermore, through automation, it completely eliminates completely reader and tracing variability. The iDI adds less than 1 minute of additional acquisition time,

and analysis can be performed instantly using the open-source software I have provided.

I find that the iDI doubles the precision of aortic stenosis quantification. A colleague at a different institution provided data confirming this finding.

6.4 Improving and validating AQURO derived EROA for mitral regurgitation

I updated the AQURO algorithm (Chapter 5) for the quantitative assessment of mitral regurgitation to improve its robustness and performed a pilot validation study at my own institution. AQURO provides comparable results and precision to PISA derived EROA with the benefit of automated analysis.

6.5 Future plans

The iDI is currently being evaluated by independent institutions for consideration of use within clinical trials. Whilst I did analyse independently collected data from 10 patients for this thesis, further data from 'different-hands' is essential to demonstrate its validity. Furthermore, as I found in Chapter 3, reproducibility data collected as a by-product of clinical trials often finds different coefficient of variations. Similarly, my updated version of AQURO for the quantitative assessment of mitral regurgitation needs clinically revalidating.

7 References

Altman, D.G., 1990. *Practical statistics for medical research*, Boca Raton, Fla: Chapman & Hall/CRC.

Altman, D.G. & Bland, J.M., 2005. Standard deviations and standard errors. *BMJ (Clinical Research Ed.)*, 331(7521), p.903.

Bargiggia, G.S. et al., 1991. A new method for quantitation of mitral regurgitation based on color flow Doppler imaging of flow convergence proximal to regurgitant orifice. *Circulation*, 84(4), pp.1481–1489.

Bargiggia, G.S. et al., 1988. Quantitative assessmentof mitral regurgitation by color Doppler analysis of flowconvergence region: Usefulness of continuity equation. *Proceedings of the Sixth International Congress on Cardiac Doppler*, p.140.

Baumgartner, H. et al., 2017. 2017 ESC/EACTS Guidelines for the management of valvular heart disease. *European Heart Journal*, 38(36), pp.2739–2791.

Baumgartner, H. et al., 2009. Echocardiographic assessment of valve stenosis: EAE/ASE recommendations for clinical practice. *European Journal of Echocardiography*, 10(1), pp.1–25.

BCIS Audit, 2016. Audit Results 1991 to present - BCIS - BCIS. Available at: https://www.bcis.org.uk/resources/audit-results/ [Accessed July 25, 2018].

Bland, J.M. & Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 1(8476), pp.307–310.

Bland, M., 2004. How do I analyse observer variation studies? *How do I analyse observer variation studies?* Available at: https://www-users.york.ac.uk/~mb55/meas/obshead.htm [Accessed December 28, 2018].

Bland, M., 2006. How should I calculate a within-subject coefficient of variation? Available at: https://www-users.york.ac.uk/~mb55/meas/cv.htm [Accessed July 29, 2018].

Bland, M., 2005. What is the origin of the formula for repeatablity? *What is the origin of the formula for repeatablity?* Available at: https://www-users.york.ac.uk/~mb55/meas/repeat.htm [Accessed December 28, 2018].

Bohbot, Y. et al., 2017. Impact of Mean Transaortic Pressure Gradient on Long-Term Outcome in Patients With Severe Aortic Stenosis and Preserved Left Ventricular Ejection Fraction. *Journal of the American Heart Association*, 6(6).

Bottou, L., Curtis, F.E. & Nocedal, J., 2018. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2), pp.223–311.

Brignole, M. et al., 2013. 2013 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy: the Task Force on cardiac pacing and resynchronization therapy of the European Society of Cardiology (ESC). Developed in collaboration with the European Heart Rhythm Association (EHRA). *European Heart Journal*, 34(29), pp.2281–2329.

Brubakk, A.O., Angelsen, B.A. & Hatle, L., 1977. Diagnosis of valvular heart disease using transcutaneous Doppler ultrasound. *Cardiovascular Research*, 11(5), pp.461–469.

Currie, P.J. et al., 1985. Continuous-wave Doppler echocardiographic assessment of severity of calcific aortic stenosis: a simultaneous Doppler-catheter correlative study in 100 adult patients. *Circulation*, 71(6), pp.1162–1169.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4), pp.303–314.

Douglas, P.S. et al., 2009. Echocardiographic imaging in clinical trials: American Society of Echocardiography Standards for echocardiography core laboratories: endorsed by the American College of Cardiology Foundation. *Journal of the American Society of Echocardiography*, 22(7), pp.755–765.

Du, P., Kibbe, W.A. & Lin, S.M., 2006. Improved peak detection in mass spectrum by

incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, 22(17), pp.2059–2065.
Dujardin, K.S. et al., 1997. Grading of mitral regurgitation by quantitative Doppler echocardiography: calibration by left ventricular angiography in routine clinical practice. *Circulation*, 96(10), pp.3409–3415.
Dweck, M.R. et al., 2013. Aortic stenosis, atherosclerosis, and skeletal bone: is there a common link with calcification and inflammation? *European Heart Journal*, 34(21), pp.1567–1574.
Enriquez-Sarano, M. et al., 2005. Quantitative determinants of the outcome of asymptomatic mitral regurgitation. *The New England Journal of Medicine*, 352(9), pp.875–883.
Evangelista, A. et al., 2008. European Association of Echocardiography recommendations for standardization of performance, digital storage and reporting of echocardiographic studies. *European Journal of Echocardiography*, 9(4), pp.438–448.
Finegold, J.A. et al., 2013. Choosing between velocity-time-integral ratio and peak velocity ratio for calculation of the dimensionless index (or aortic valve area) in serial follow-up of aortic stenosis. *International Journal of Cardiology*, 167(4), pp.1524–1531.
Galderisi, M. et al., 2011. Recommendations of the European Association of Echocardiography: how to use echo-Doppler in clinical trials: different modalities for different purposes. *European Journal of Echocardiography*, 12(5), pp.339–353.
Galderisi, M. et al., 2017. Standardization of adult transthoracic echocardiography reporting in agreement with recent chamber quantification, diastolic function, and heart valve disease recommendations: an expert consensus document of the European Association of Cardiovascular Imaging. *European heart journal cardiovascular Imaging*, 18(12), pp.1301–1310.
Giesler, M.O. & Stauch, M., 1992. Color Doppler determination of regurgitant flow: from proximal isovelocity surface areas to proximal velocity profiles. An in vitro study. *Echocardiography*, 9(1), pp.51–62.
Gosse, P. et al., 2004. Serial echocardiographic assessment of left ventricular mass: how blinded should readers be? *Journal of Hypertension*, 22(9), pp.1813–1818.
Grayburn, P.A. et al., 2014. Defining "severe" secondary mitral regurgitation: emphasizing an integrated approach. *Journal of the American College of Cardiology*, 64(25), pp.2792–2801.
Grayburn, P.A., Stoler, R.C. & Mack, M.J., 2017. The 2017 ACC/AHA updated valve guidelines regarding mitral regurgitation: the guidelines get it right. *Structural Heart*, 1(1–2), pp.31–33.
Hadfield, J.D., 2010. MCMC Methods for Multi-Response Generalized Linear Mixed Models: TheMCMCglmm *R* Package. *Journal of statistical software*, 33(2).
Hatle, L., Angelsen, B.A. & Tromsdal, A., 1980. Non-invasive assessment of aortic stenosis by Doppler ultrasound. *British heart journal*, 43(3), pp.284–292.
He, K. et al., 2016. Deep residual learning for image recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 770–778.
Holen, J. et al., 1976. Determination of pressure gradient in mitral stenosis with a non-invasive ultrasound Doppler technique. *Acta medica Scandinavica*, 199(6), pp.455–460.
Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2), pp.251–257.
International Standards Organisation, 2017. Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation. *Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation*. Available at:

https://www.iso.org/obp/ui/#iso:std:iso:21748:ed-2:v1:en [Accessed July 29, 2018].
Ioffe, S. & Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv*.
Jander, N. et al., 2014. Velocity ratio predicts outcomes in patients with low gradient severe aortic stenosis and preserved EF. *Heart*, 100(24), pp.1946–1953.
Joint Committee for Guides in Metrology, 2008. Evaulation of measurement data - Guide to the expression of uncertainty in measurement. *Joint Committee for Guides in Metrology*.
Kullback, S. & Leibler, R.A., 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), pp.79–86.
Lancellotti, P., Tribouilloy, C., et al., 2010. European Association of Echocardiography recommendations for the assessment of valvular regurgitation. Part 1: aortic and pulmonary regurgitation (native valve disease). *European Journal of Echocardiography*, 11(3), pp.223–244.
Lancellotti, P., Moura, L., et al., 2010. European Association of Echocardiography recommendations for the assessment of valvular regurgitation. Part 2: mitral and tricuspid regurgitation (native valve disease). *European Journal of Echocardiography*, 11(4), pp.307–332.
Lancellotti, P. et al., 2013. Recommendations for the echocardiographic assessment of native valvular regurgitation: an executive summary from the European Association of Cardiovascular Imaging. *European heart journal cardiovascular Imaging*, 14(7), pp.611–644.
Lancellotti, P. et al., 2016. Recommendations for the imaging assessment of prosthetic heart valves: a report from the European Association of Cardiovascular Imaging endorsed by the Chinese Society of Echocardiography, the Inter-American Society of Echocardiography, and the Brazilian Department of Cardiovascular Imaging. *European heart journal cardiovascular Imaging*, 17(6), pp.589–590.
Lang, R.M. et al., 2015. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *European heart journal cardiovascular Imaging*, 16(3), pp.233–270.
Lecun, Y. et al., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), pp.2278–2324.
LeCun, Y., Bengio, Y. & Hinton, G., 2015. Deep learning. *Nature*, 521(7553), pp.436–444.
Leon, M.B. et al., 2010. Transcatheter aortic-valve implantation for aortic stenosis in patients who cannot undergo surgery. *The New England Journal of Medicine*, 363(17), pp.1597–1607.
Marwick, T.H., Zoghbi, W.A. & Narula, J., 2014. Redrawing the Borders. *JACC: Cardiovascular Imaging*, 7(3), pp.333–335.
Maus, T.M., Nhieu, S. & Herway, S.T. eds., 2016. *Essential Echocardiography: Transesophageal Echocardiography for Non-cardiac Anesthesiologists* illustrated., Springer.
Medvedofsky, D. et al., 2017. Reproducibility and experience dependence of echocardiographic indices of left ventricular function: Side-by-side comparison of global longitudinal strain and ejection fraction. *Echocardiography*, 34(3), pp.365–370.
Minners, J. et al., 2010. Inconsistent grading of aortic valve stenosis by current guidelines: haemodynamic studies in patients with apparently normal left ventricular function. *Heart*, 96(18), pp.1463–1468.
Mitchell, M., Muftakhidinov, B. & Winchen, T., 2018. *Engauge Digitizer*, Engauge Digitizer Software.
Monin, J.-L. et al., 2009. Risk score for predicting outcome in patients with asymptomatic

aortic stenosis. *Circulation*, 120(1), pp.69–75.
Moraldo, M., Bergamini, C., et al., 2013. A novel fully automated method for mitral regurgitant orifice area quantification. *International Journal of Cardiology*, 166(3), pp.688–695.
Moraldo, M. et al., 2012. Avoiding bias in measuring "hemisphere radius" in echocardiographic mitral regurgitation quantification: Mona Lisa PISA. *International Journal of Cardiology*, 155(2), pp.318–320.
Moraldo, M., Cecaro, F., et al., 2013. Evidence-based recommendations for PISA measurements in mitral regurgitation: systematic review, clinical and in-vitro study. *International Journal of Cardiology*, 168(2), pp.1220–1228.
Nagata, Y. et al., 2015. Intervendor variability of two-dimensional strain using vendor-specific and vendor-independent software. *Journal of the American Society of Echocardiography*, 28(6), pp.630–641.
Nagi, J. et al., 2011. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA). IEEE, pp. 342–347.
Nagueh, S.F. et al., 2016. Recommendations for the Evaluation of Left Ventricular Diastolic Function by Echocardiography: An Update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *European heart journal cardiovascular Imaging*, 17(12), pp.1321–1360.
Nishimura, R.A. et al., 2014. 2014 AHA/ACC Guideline for the Management of Patients With Valvular Heart Disease: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation*, 129(23), pp.e521-643.
Nishimura, R.A. et al., 2017. 2017 AHA/ACC focused update of the 2014 AHA/ACC guideline for the management of patients with valvular heart disease: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Circulation*, 135(25), pp.e1159–e1195.
Nkomo, V.T. et al., 2006. Burden of valvular heart diseases: a population-based study. *The Lancet*, 368(9540), pp.1005–1011.
Obadia, J.-F. et al., 2018. Percutaneous repair or medical treatment for secondary mitral regurgitation. *The New England Journal of Medicine*, 379(24), pp.2297–2306.
Otsu, N., 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1), pp.62–66.
Otto, C.M., 2018. Aortic stenosis: treat the patient not the numbers. *Heart*, 104(3), pp.190–191.
Otto, C.M. et al., 1997. Prospective study of asymptomatic valvular aortic stenosis. Clinical, echocardiographic, and exercise predictors of outcome. *Circulation*, 95(9), pp.2262–2270.
Pawade, T. et al., 2018. Computed tomography aortic valve calcium scoring in patients with aortic stenosis. *Circulation. Cardiovascular Imaging*, 11(3), p.e007146.
Pawade, T.A. et al., 2016. Optimization and Reproducibility of Aortic Valve 18F-Fluoride Positron Emission Tomography in Patients With Aortic Stenosis. *Circulation. Cardiovascular Imaging*, 9(10).
Pham-Gia, T. & Hung, T.L., 2001. The mean and median absolute deviations. *Mathematical and computer modelling*, 34(7–8), pp.921–936.
Python Software Foundation, 2017. *Python Language Reference, version 3.6*, Python Software Foundation. Available at: https://www.python.org [Accessed June 11, 2018].
R Core Team, 2016. *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Raina, R., Madhavan, A. & Ng, A.Y., 2009. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. the 26th Annual International Conference. New York, New York, USA: ACM Press, pp. 1–8.

Ramirez, M.L. & Wong, M., 1985. Reproducibility of stand-alone continuous-wave Doppler recordings of aortic flow velocity across bioprosthetic valves. *The American Journal of Cardiology*, 55(9), pp.1197–1199.

Recusani, F. et al., 1991. A new method for quantification of regurgitant flow rate using color Doppler flow imaging of the flow convergence region proximal to a discrete orifice. An in vitro study. *Circulation*, 83(2), pp.594–604.

Rosenblatt, F., 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65(5), pp.386–408.

Rossebø, A.B. et al., 2008. Intensive lipid lowering with simvastatin and ezetimibe in aortic stenosis. *The New England Journal of Medicine*, 359(13), pp.1343–1356.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J., 1986. Learning representations by back-propagating errors. *Nature*, 323(6088), pp.533–536.

Rusinaru, D. et al., 2015. Relation of Dimensionless Index to Long-Term Outcome in Aortic Stenosis With Preserved LVEF. *JACC. Cardiovascular Imaging*, 8(7), pp.766–775.

Sacchi, S. et al., 2018. Doppler assessment of aortic stenosis: a 25-operator study demonstrating why reading the peak velocity is superior to velocity time integral. *European heart journal cardiovascular Imaging*.

Savitzky, A. & Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8), pp.1627–1639.

Schindelin, J. et al., 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7), pp.676–682.

Schroeder, E.B. et al., 2004. Repeatability of heart rate variability measures. *Journal of Electrocardiology*, 37(3), pp.163–172.

SCTS, 2016. SCTS Blue Book Online - Cardiac Surgery Audit Data. Available at: http://bluebook.scts.org/ [Accessed July 25, 2018].

Seabra, J. & Sanches, J., 2008. Modeling log-compressed ultrasound images for radio frequency signal recovery. *Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2008, pp.426–429.

Senn, S., 2011. Francis Galton and regression to the mean. *Significance*, 8(3), pp.124–126.

Shiota, Takahiro et al., 1994. Estimation of regurgitant flow volume based on centerline velocity/distance profiles using digital color M-Q Doppler: Application to orifices of different shapes. *Journal of the American College of Cardiology*, 24(2), pp.440–445.

Shiota, T et al., 1994. Evaluation of mitral regurgitation using a digitally determined color Doppler flow convergence "centerline" acceleration method. Studies in an animal model with quantified mitral regurgitation. *Circulation*, 89(6), pp.2879–2887.

Siostrzonek, P. et al., 1988. Day to day reproducibility of Doppler sonographic measurements in patients with valvular aortic stenosis. *Clinical Cardiology*, 11(11), pp.748–750.

Slabaugh, G. et al., 2009. Statistical region-based segmentation of ultrasound images. *Ultrasound in Medicine & Biology*, 35(5), pp.781–795.

Spearman, C., 1904. The Proof and Measurement of Association between Two Things. *The American journal of psychology*, 15(1), p.72.

Stone, G.W. et al., 2018. Transcatheter Mitral-Valve Repair in Patients with Heart Failure. *The New England Journal of Medicine*, 379(24), pp.2307–2318.

Szegedy, C. et al., 2015. Going deeper with convolutions. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1–9.

Taylor, B.N. & Kuyat, C.E., 1994. Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. *NIST Technical Note 1297*.
The Horos Project, 2018. *Horos*, The Horos Project.
The Inkscape Team, 2018. *Inkscape*, inkscape.org.
Thiago, L. et al., 2016. Statins for aortic valve stenosis. *Cochrane Database of Systematic Reviews*, 9, p.CD009571.
Thomas, N. et al., 2008. Intraobserver variability in grading severity of repeated identical cases of mitral regurgitation. *American Heart Journal*, 156(6), pp.1089–1094.
Tribouilloy, C.M. et al., 1999. Impact of preoperative symptoms on survival after surgical correction of organic mitral regurgitation: rationale for optimizing surgical indications. *Circulation*, 99(3), pp.400–405.
Tzoulaki, I., Siontis, K.C.M. & Ioannidis, J.P.A., 2011. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. *BMJ (Clinical Research Ed.)*, 343, p.d6829.
Vandervoort, P.M. et al., 1993. Automated flow rate calculations based on digital analysis of flow convergence proximal to regurgitant orifices. *Journal of the American College of Cardiology*, 22(2), pp.535–541.
Werbos, P.J. (Paul J., 1994. *The roots of backpropagation: From ordered derivatives to neural networks and political forecasting*, New York: Wiley.
Wharton, G. et al., 2015. A minimum dataset for a standard adult transthoracic echocardiogram: a guideline protocol from the British Society of Echocardiography. *Echo research and practice*, 2(1), pp.G9–G24.
Wickham, H., 2009. *ggplot2: Elegant Graphics for Data Analysis* illustrated., Springer Science & Business Media.
Wickham, H., 2017. tidyverse: Easily Install and Load the "Tidyverse."
Zaman, J.A.B. et al., 2014. Focal impulse and rotor mapping (FIRM): conceptualizing and treating atrial fibrillation. *Journal of atrial fibrillation*, 7(2), p.1103.
Zamorano, J.L. et al., 2016. 2016 ESC Position Paper on cancer treatments and cardiovascular toxicity developed under the auspices of the ESC Committee for Practice Guidelines: The Task Force for cancer treatments and cardiovascular toxicity of the European Society of Cardiology (ESC). *European Heart Journal*, 37(36), pp.2768–2801.
Zoghbi, W., 2003. Recommendations for evaluation of the severity of native valvular regurgitation with two-dimensional and doppler echocardiography. *Journal of the American Society of Echocardiography*, 16(7), pp.777–802.
Zoghbi, W.A. et al., 2017. Recommendations for noninvasive evaluation of native valvular regurgitation. *Journal of the American Society of Echocardiography*.

`