Atrioventricular and interventricular delay optimization and response quantification in biventricular pacing: arrival of reliable clinical algorithms and research protocols, and how to distinguish them from unreliable counterparts

Berthold Stegemann1* and Darrel P. Francis2

1Bakken Research Center, Endepolsdomein 5, 6229 GW Maastricht, The Netherlands; and 2International Centre for Circulatory Health, National Heart and Lung Institute, Imperial College London, London, UK

Received 3 July 2012; accepted after revision 9 July 2012

This editorial refers to ‘Cardiac resynchronization therapy beyond nominal settings: who needs individual programming of the atrioventricular and interventricular delay?’ by M. D. Bogaard et al., on page 1746

In this issue of the journal, Bogaard et al.1 assess the achievable haemodynamic increment within individuals of optimizing atrioventricular (AV) and interventricular (VV) delays. Biventricular pacing, or cardiac resynchronization therapy (CRT), arguably the greatest advancement in the treatment of heart failure in the last decade, causes immediate increases in haemodynamics and then improvements in symptoms, exercise capacity, and reductions in hospitalization and mortality. The landmark endpoint studies included optimization of AV delay to give the most ideal appearance of transmitral Doppler.

Subsequently, many studies, not always randomized and controlled,2 have assessed alternative optimization techniques including echocardiographic aortic velocity–time integral (VTI), mitral VTI, left ventricular (LV) end-systolic volume and ejection fraction, transmitral Doppler (E-A) duration, E-A truncation; thoracic electrical impedance; cardiac output by rebreathing techniques; blood pressure, plethysmography and pulse contour analysis; intracardiac electrogram and surface electrocardiogram, QRS morphology, QRS axis; heart sounds and phonocardiography; and LV pressure and pressure derivatives (dp/dtmax) and pressure–volume analysis.

The mystery of abundant methods

Investigators mostly found what they expected: their proposed method was effective. If so many different methods of optimization appear beneficial, then either:

Hypothesis 1: all optimization methods agree regarding the optimum setting (which would explain why they all worked), or

Hypothesis 2: optimization methods do not agree well, but accidental bias may have caused some studies to overstate the optimization increment.

Hypothesis 1 is only viable if the optimization techniques are highly reproducible between one optimization session and another, since no two methods can agree better than they each agree with themselves. But there are little data from credible sources confirming between-method agreement, or even addressing blinded test–retest reproducibility from independently acquired datasets for individual optimization methods. Lack of convincing confirmation of between-method consistency, after many years, may be a warning sign: a ‘dog that didn’t bark’.3

Hypothesis 2 has many possible mechanisms. First, some studies used symptom status reported by patients even though they knew that they were undergoing ‘optimization’ to maximize their pacemaker function. Second, where response was defined echocardiographically, staff analyzing and – equally importantly – acquiring the images might not be blinded. Third, where the response was calculated from the optimization dataset itself, positive bias is guaranteed (Figure 1 and Supplementary File S1). Finally, some studies confuse between-patient variation with between-setting variation, reporting irrelevant statistics (Figure 2 and Supplementary File S2).

The opinions expressed in this article are not necessarily those of the Editors of Europace or of the European Society of Cardiology.

* Corresponding author. Tel: +49 228 9289 4161; fax: +49 228 9289 4824. Email: berthold.stegemann@t-online.de

Published on behalf of the European Society of Cardiology. All rights reserved. © The Author 2012. For permissions please email: journals.permissions@oup.com.
Noise alone consistently produces apparently very significant increments from optimization, if analysis is simply “pick-the-highest”

**Figure 1** How not to assess an optimization technique (part 1). Downloadable Supplementary File S1 shows why beat-to-beat variability generates an artefactual ‘benefit’ of optimization, if careful steps are not taken to exclude upward bias. In this simulation, all the atrioventricular delays have the same underlying value (which the reader can set in cell F9), but measurements vary randomly with a standard deviation (cell F7). The 12-patient study almost always shows a ‘highly significant’ increase from what is a worthless process, because the protocol incorporates upward bias. Bogaard et al. have avoided this, but most studies of physiological impact of optimization have not.

**Inputs**
- Enter the test-retest variability of the measurement (expressed as a standard deviation): 50
- Enter the underlying mean value: 1000

**Press F9 to re-run simulation**

<table>
<thead>
<tr>
<th>AV delay</th>
<th>Measurement</th>
<th>Apparent optimum</th>
<th>Apparent increment above AV120</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1023</td>
<td></td>
<td>+24</td>
</tr>
<tr>
<td>60</td>
<td>995</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>1000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>1034</td>
<td></td>
<td></td>
</tr>
<tr>
<td>140</td>
<td>1030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>1019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>180</td>
<td>928</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>998</td>
<td></td>
<td></td>
</tr>
<tr>
<td>220</td>
<td>1058</td>
<td>----</td>
<td>+24</td>
</tr>
<tr>
<td>240</td>
<td>1013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1127</td>
<td></td>
<td>+115</td>
</tr>
<tr>
<td>60</td>
<td>1101</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>974</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>996</td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>1012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>140</td>
<td>1066</td>
<td></td>
<td></td>
</tr>
<tr>
<td>160</td>
<td>1038</td>
<td></td>
<td></td>
</tr>
<tr>
<td>180</td>
<td>1057</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>998</td>
<td></td>
<td></td>
</tr>
<tr>
<td>220</td>
<td>944</td>
<td></td>
<td></td>
</tr>
<tr>
<td>240</td>
<td>1036</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary table of apparent increments in all 12 patients:

- Patient 1 +24
- Patient 2 +115
- Patient 3 +58
- Patient 4 +68
- Patient 5 +101
- Patient 6 +169
- Patient 7 0
- Patient 8 +130
- Patient 9 +32
- Patient 10 +170
- Patient 11 +87
- Patient 12 +102

Mean: +88
SD: 54
P: 0.000159

i.e. AV 120 happens to be highest
**Figure 2** How not to assess an optimization technique (part 2). Downloadable Supplementary File S2 shows that if the between-patient variation in the measurement (editable in cells J7 and J8) is much larger than the within-patient variation (J4), then the correlation across patients between any pair of measurements in each patient will be high (row 32). Therefore, a high correlation coefficient between a measurement at any predicted optimum and the actual highest observed value, tells the reader nothing about the effectiveness of the optimum-prediction technique (red dots). Published graphs typically do not show the non-selected measurements (grey dots) whose correlation may have been just as good, but usually highlight the irrelevant correlation coefficient and even-more irrelevant P value.
Well-designed, well-executed, well-reported trials of optimization methods – whose development was paradoxically obscure

Two large externally monitored randomized controlled trials of optimization have been carried out and reported.

The FREEDOM (Frequent Optimization Study Using the QuickOpt Method) trial (St Jude Medical) assessed whether echocardiographic optimization could be replaced by device-based AV/VV-delay optimization using intracardiac electrogram analysis by the QuickOpt algorithm. Rather than testing whether QuickOpt delivered the same proposed optimal AV as transmitral Doppler, the initial reports of QuickOpt were that it resulted in velocity–time integral that correlated (across patients) with the highest achievable velocity–time integral in those patients: this would have been the case even if the optima were completely uncorrelated. The FREEDOM results presented at Heart Rhythm Society in 2011 showed no statistically significant difference between echo and QuickOpt approaches.

The Smart AV trial assessed the efficacy of AV delay optimization by the electrogram-based Smart-AV algorithm, again designed to match the optimum identified by qualitative transmitral Doppler – itself of uncertain test–retest reproducibility. It found no significant reduction in LV systolic volume in patients randomized to the electrogram optimization, or randomized to the transmural Doppler optimization, in comparison with patients randomized to nominal settings.

A third study is in progress, defining the optimum as the point of inflection of the decline in amplitude of the first heart sound during AV lengthening (Son-R, Sorin) for reasons which are obscure. Its randomized partly blinded pilot study showed no statistically significant change in markers measured under blinded conditions, but a clear improvement in NYHA (New York Heart Association) class assessed unblinded.

Clinicians vs. guideline writers

Most clinicians defy guidelines and do not routinely optimize AV delay. Either clinicians, or guidelines, must be in error. Guideline writers might argue that the landmark survival trial CARE-HF (Cardiac RESynchronisation-Heart Failure) included AV optimization, and that clinicians must therefore comply. But not every constraint of a successful protocol contributes to survival benefit: it is (for example) immediately evident that the brand of pacemaker specified in CARE-HF does not. Separating contributory from non-contributory decision-making elements in a successful protocol can be difficult, but if a decision-making element is not an algorithm, i.e. does not reliably deliver the same decision when re-executed, it can be discarded as a worthless consumption of resources, or even potentially harmful.4,5

Guidelines could re-earn their scientific credibility by presenting experiments by which clinicians could quickly verify that the recommended AV optimization scheme is an algorithm (i.e. delivers a consistent optimum). For example, clinicians could be advised, in a few patients, to have two mutually blinded clinicians perform separate AV optimizations (i.e. distinct data, not re-reading identical images). Guideline writers should also state what level of contradiction they consider clinically acceptable, having checked that they themselves can achieve it. Clinicians should also play their part by openly challenging and rejecting guidelines if rudimentary checks do not appear to have been carried out.6

Reliable protocols for optimization

Bogaard et al. contribute experimental science into this landscape, by designing and implementing a study to calculate the achievable acute haemodynamic effect of optimization of AV and VV delay, through invasive assessment of changes in the maximum slope of LV systolic pressure rise (dp/dtmax) in a complex and demanding protocol. The authors took impressive efforts to prevent measurements becoming dominated by physiological variability. For calculation purposes, beat-to-beat variability – whether physiological (i.e. actually occurring in the body) or due to technical imperfection of measurement – has the same disruptive effect, and can be called ‘noise’ for short. Six-beat averages of haemodynamic measurements were used, predefined as the first six, because values vary between beats and the pattern of fluctuation in haemodynamics after a change in setting makes it wise to measure early.7 Four AV delays were tested to determine the optimum AV delay. Then seven different VV delays were similarly studied.

The authors carefully avoided the trap of simply picking the highest measurement as the optimum: for many commonly recommended protocol designs this systematically picks the wrong setting.6 Bogaard et al. fitted a parabola first to optimize AV delay, and then separately to optimize VV delay, to maximize the reliability of their estimates of patients’ increment from optimization.8 Many patients had submaximal effect from default device settings.

How not to assess an optimization

The authors’ greatest methodological innovation is the special care to quantify the impact of AV/VV optimization without accidental exaggeration. The large risk of inadvertent exaggeration is often not evident without example. A downloadable spreadsheet (Supplementary File S1), of which a snapshot is shown in Figure 1, is offered to illustrate this. When an optimum is selected by ‘pick-the-highest’, and the same dataset used to calculate the increment achieved by optimization then, for every patient, that increment is guaranteed to be positive (or zero). No patient’s increment can be negative. Therefore across patients, the average increment will always be statistically significantly positive, even if the pacemaker has no effect whatsoever.

Bogaard et al. have used curve fitting to greatly reduce this problem. Although the few percent increase in dp/dtmax they report may seem small, biventricular pacing itself gave only ~6% blood pressure increment in CARE-HF and COMPANION
(Comparison of Medical Therapy, Pacing and Defibrillation in Heart Failure). Their method permits the identification of optima between tested settings, separates signal from noise, and finds optima with quantifiable and greater precision than ‘pick-the-highest’.

Nor have Bogaard et al. fallen for the second commonest trap of confusing between-patient variability (whose magnitude is irrelevant to individual patients) with within-patient, between-setting differences (Supplementary File S2 and Figure 2).

Don’t ask an expert: do an experiment

Careful experimental design and elaborate analysis permits the authors to draw conclusions confident in the knowledge that they are not artefacts of noise, and that the findings will stand the test of time, which is critical for clinical application. Routine curve fitting also makes it easy to evaluate and report the confidence interval of the optimum, which is proportional to the noise standard deviation and inversely proportional to the curvature of the response curve and the number of repeated measures. Stating confidence intervals may be a quick way of eliminating unreliable optimization methods.

Bogaard et al. find that one-third of the variance of the haemodynamic response to optimization is determined by the intrinsic PR interval. This suggests that some element of fusion with native conduction is important for optimal function in many patients. That finding has a devastating further implication, because the sum of $R^2$ values of statistically independent predictors cannot exceed 1.0. Thus, if the $R^2$ of response with PR is indeed 0.34 in a large population, then no response predictor statistically independent of PR can have an $R^2$ in that population above 0.66. The ceiling on plausible prediction of echocardiographic response from current mechanical dyssynchrony predictors, recently found to be $\sim 0.3\text{--}0.4$, may therefore need to be lowered to $<0.66 \times 0.3\text{--}0.4$ to recognize the contribution of PR. A ceiling on $R^2$ of $\sim 0.2\text{--}0.3$ for ventricular dyssynchrony predictors accords with studies designed to resist bias, such as PROSPECT (Predictors of Response to CRT), but means that other claims, on the front line of the charge towards complex ventricular predictors of response, are far into untenable territory. It is fitting that Bogaard et al., rejecting expert recommendations for transmitral optimization, also reject the manufacturers’ 0 ms VV delay default. Confidence in the manufacturers’ default VV is unjustified, when their AV defaults are so discrepant, with paced AV defaults ranging between 130 and 200 ms.

Aside from these mechanistic and practical findings, there is one more enduring contribution in this study. A Bogaard approach, using AAI pacing as the reference, is finally a practical way to quantify ‘response’ reproducibly and — by doing enough replicates — to any desired degree of precision. This might even be conducted with a non-invasive marker. Narrow within-patient error bars for response might give much-needed plausibility to the long-running ill-fated search for mechanical dyssynchrony predictors. Ultimately, Bogaard response may be a variable one can bet one’s life on — which indeed is what we are asking our patients to do.

Supplementary material

Supplementary material is available at Europace online.

Conflict of interest: D.F. holds grants from the British Heart Foundation, European Research Council and the UK’s Medical Research Council. D.P.F. is employed by Imperial College London which has filed patents on haemodynamic methods for pacemaker optimization which emphasise reproducibility.

References

6. Francis DP. How to reliably deliver narrow individual-patient error bars for optimization of pacemaker AV or VV delay using a ‘pick-the-highest’ strategy with haemodynamic measurements. Int J Cardiol (in press) [Epub ahead of print 18 April 2012].