ESC European Society of Cardiology

Cardiovascular Research (2017) **0**, 1–10
doi:10.1093/cvr/cvx151

Original article

Hierarchical statistical techniques are necessary to draw reliable conclusions from analysis of isolated cardiomyocyte studies

Markus B. Sikkel[1,2†], Darrel P. Francis[1†], James Howard[1], Fabiana Gordon[3], Christina Rowlands[1], Nicholas S. Peters[1,2], Alexander R. Lyon[1,4], Sian E. Harding[1]*, and Kenneth T. MacLeod[1]

[1]Myocardial Function Section, Fourth Floor, Imperial Centre for Translational and Experimental Medicine, National Heart and Lung Institute, Imperial College London, Hammersmith Campus, Du Cane Road, London W12 0NN, UK; [2]Department of Electrophysiology, Imperial College Healthcare NHS Trust, Hammersmith Hospital, London W12 0HS, UK; [3]Statistics Advisory Service, Imperial College London, London, UK; and [4]Department of Cardiology, Royal Brompton Hospital, London, UK

Received 17 March 2017; revised 27 June 2017; editorial decision 26 July 2017; accepted 29 August 2017

Time for primary review: 44 days

Aims It is generally accepted that post-MI heart failure (HF) changes a variety of aspects of sarcoplasmic reticular Ca^{2+} fluxes but for some aspects there is disagreement over whether there is an increase or decrease. The commonest statistical approach is to treat data collected from each cell as independent, even though they are really clustered with multiple likely similar cells from each heart. In this study, we test whether this statistical assumption of independence can lead the investigator to draw conclusions that would be considered erroneous if the analysis handled clustering with specific statistical techniques (hierarchical tests).

Methods and results Ca^{2+} transients were recorded in cells loaded with Fura-2AM and sparks were recorded in cells loaded with Fluo-4AM. Data were analysed twice, once with the common statistical approach (assumption of independence) and once with hierarchical statistical methodologies designed to allow for any clustering. The statistical tests found that there was significant hierarchical clustering. This caused the common statistical approach to underestimate the standard error and report artificially small *P* values. For example, this would have led to the erroneous conclusion that time to 50% peak transient amplitude was significantly prolonged in HF.

Spark analysis showed clustering, both within each cell and also within each rat, for morphological variables. This means that a three-level hierarchical model is sometimes required for such measures. Standard statistical methodologies, if used instead, erroneously suggest that spark amplitude is significantly greater in HF and spark duration is reduced in HF.

Conclusion Ca^{2+} fluxes in isolated cardiomyocytes show so much clustering that the common statistical approach that assumes independence of each data point will frequently give the false appearance of statistically significant changes. Hierarchical statistical methodologies need a little more effort, but are necessary for reliable conclusions. We present cost-free simple tools for performing these analyses.

Keywords Hierarchical statistics • Cardiomyocyte • Ca^{2+} transient • Ca^{2+} spark

1. Introduction

Changes in cellular Ca^{2+} handling are agreed to play a key role in the pathophysiology of heart failure (HF).[1] The standard experimental model, established for half a century is left anterior descending coronary artery ligation in the rat.[2] At the whole organ level, there is broad agreement between experimenters on the pattern of haemodynamic and echocardiographic variables.[3–8]

†These authors contributed equally to this work.

* Corresponding author. Tel: +44 207 594 3009; fax: +44 207 351 8145, E-mail: sian.harding@ic.ac.uk

At the cellular level, however, there is more controversy. Changes in Ca^{2+} transient amplitude and kinetics have been widely reported to contribute to organ dysfunction.[1,9] However, the 18 studies that report on changes in Ca^{2+} transient kinetics show considerable conflict (*Table 1*). On even smaller scales, Ca^{2+} spark measurements also show convincing changes, again in conflicting directions (see Supplementary material online, *Table S1*). This is not a peculiarity of the rat: mouse studies too show the same conflicts (see Supplementary material online, *Tables S2* and *S3*).

The accepted paradigm is that reduced Ca^{2+} transient amplitude contributes to reduced myocardial contractility.[10] However, the cellular data do not consistently fit this. For example, Ca^{2+} transient amplitudes in HF cells are reported as significantly increased in more studies than decreased (*Table 1*). Moreover, the two studies that reported changes in SR Ca^{2+} content indicated opposite directions of change (*Table 1*). In other variables, studies conflict between showing convincing changes and convincing absence of change (*Table 1*).

There is also no apparent time-dependent shift in variables following MI as would be expected for variables that change from a 'compensated' to 'de-compensated' cellular phenotype.

Authors have proposed explanations including a compensated phase following MI,[5] or differences in other elements of the myocyte (e.g. myofilaments) overriding the effects of enhanced Ca^{2+} transients. When one study found enhanced Ca^{2+} amplitudes simultaneously with reduced myocyte contractility, the proposed explanation was a reduction in myofilament sensitivity to Ca^{2+}.[11]

Rarely considered is the possibility that our statistical methods may have left us open to changes appearing falsely statistically significant. The studies in the field generally used standard statistical tests designed to be valid for independent data points, because these tests are widely available and straightforward to implement.

The challenge we face is that when we take *n* cells from each of *m* animals we do not truly have $n \times m$ independent data points.[12] We have *m* clusters, each containing *n* data points. The data points in each cluster (i.e. the cells from a single animal) will tend to be more similar to each other than they are to points in the other clusters (*Figure 1*). There are well-established statistical techniques for handling data that shows such clustering. Whether using cluster based analysis produces different results from our field's standard statistical tests, has never been tested.

When analysing such a set of $n \times m$ data points, investigators tend to take one of two approaches. The more conservative approach is to calculate a mean for each of the m animals and treat these *m* mean values as the only data points. This approach is not popular because it treats the sample size as only *m* and therefore, reduces the ability to detect changes.

The far more popular approach is to ignore the clustering and treat all $n \times m$ data points as though they are independent. The attraction of this approach is that when the standard statistical tests are performed, the large number of data points lead to the standard error being very small and therefore differences appear to be more statistically significant.[13] In fact the term 'pseudoreplication' has been coined for this.[14] Studies rarely specify which of these two approaches have been taken but the sample sizes reported generally appear consistent with the latter approach.

A third approach, rarely used, is to recognize the clustered (hierarchical) structure of the data. And use statistical methods designed for this situation.[15] Hierarchical statistical methods test for clustering and, if present, correct for this when performing significance testing.

Table 1 Summary of changes in electrically evoked Ca^{2+} transients and SR load assessment in studies of rats with post-MI HF. Arrows refer to whether variables (evoked Ca^{2+} transient amplitude and decay time, SR Ca^{2+} content and diastolic $[Ca^{2+}]_i$) were found to be significantly reduced (↓), the same (↔), or increased (↑) in myocytes isolated from post-MI animals compared with control animals in the post-MI rat HF model

Publication	Wks post-MI	Ca^{2+} transient amplitude			Transient decay time			SR Ca^{2+} content			Diastolic $[Ca^{2+}]_i$		
Cheung *et al.*[4] [a]	3				↔								
Huang *et al.*[6]	3					↑							
Zhang *et al.*[7]	3		↔		↔							↔	
Anand[8]	6		↔		↔								
Sande *et al.*[29]	6				↔								
Holt *et al.*[9]	6	↓				↑							↑
Soppa *et al.*[5]	6			↑		↑				↑			↑
Lee *et al.*[30]	7		↔			↑		↔				↔	
Maczewski and Mackiewicz[31]	8		↔			↑		↔					
Kaprielian *et al.*[32]	8			↑				↔				↔	
Loennechen[33]	8			↑		↑							↑
Yoshida *et al.*[34]	8		↔		↔							↔	
Cheng *et al.*[11]	8			↑		↑							
Saraiva *et al.*[35]	9	↓			↔								
Loennechen *et al.*[36]	13			↑		↑							↑
Lyon *et al.*[3]	16					↑							
Lyon *et al.*[37]	16					↑	↓						
Ait Mou *et al.*[38]	18	↓				↑							
Total		3	5	5	0	6	11	1	3	1	0	4	4

[a]At physiological Ca^{2+} (increased decay time at supraphysiological Ca^{2+} of 5 mM).

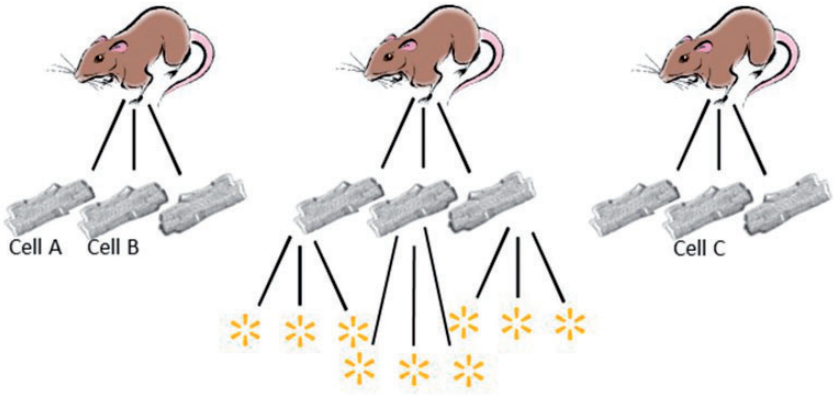

Figure 1 Hierarchical structure of data attained from studies of isolated cardiomyocytes. Multiple cardiomyocytes originate from each isolation. Differences in the animals from which the myocytes originate, as well as slight variations in quality of isolation or experimental conditions on any one day, may result in measurements taken from the myocyte from one rat being more closely related to each other than to measurements from a different isolations. That is, measurements in cell A are more likely to be similar to those in cell B vs. those in cell C in the diagram. A fundamental condition of common statistical tests (e.g. *t*-tests), that of independence of data points, is therefore contravened. An example of a further level of hierarchy is shown for the middle rat with multiple individual sparks recorded from each cell.

In this study we test the application of a hierarchical statistical approach to data from myocytes from the post-MI rat HF model. We examine whether (i) it makes a difference to the conclusions drawn, (ii) whether it is necessary, and (iii) whether it can be implemented conveniently by experimenters.

2. Methods

2.1 Rat HF model and myocyte imaging

The chronic rat HF model was produced surgically via LAD ligation[3] with an 8 week delay before echocardiography, culling and biometric measurement as described previously.[16] Briefly, rats were anesthetized with 2% isoflurane, intubated, and ventilated after preoperative buprenorphine (0.03 mg/kg SC) injection. Loss of righting reflex and pain responses confirmed dequate analgesia and anaesthesia before operating and tying off the LAD with 6-0 prolene. All animal surgical procedures and peri-operative management were carried out in accordance with the Guide for the Care and Use of Laboratory Animals published by the US National Institutes of Health (NIH Publication, 8th Edition, 2011) under assurance number A5634-01. Imperial College Ethical Review Committee authorized the project licence.

Echocardiography and biometric measurements showed a reduced ejection fraction compared with control animals (40 ± 2.2% vs. 83 ± 2.1%, $P < 0.0001$), as well as a significant increase in heart weight: body weight ratio (3.28 ± 0.1 vs. 2.79 ± 0.1, $P < 0.0001$) confirming established HF (see Supplementary material online, *Figure S1*).

Rats were culled by cervical dislocation and cells were isolated 8 weeks following ligation. Cells were loaded with Fura2-AM and imaged using ratiometric techniques as described.[17] Ca^{2+} transients were recorded at 1 Hz in control (76 cells from 10 isolations) and HF cells (79 cells from 10 isolations) using the ratiometric dye Fura2-AM. The ratio of the isosbestic point for Fura2 (360 nm) to the unbound form (380 nm) was used as a measure of $[Ca^{2+}]_i$. Several variables were assessed using automated transient analysis in IonWizard (Ionoptix Inc.) including diastolic ratio, peak systolic ratio, transient amplitude (peak ratio/diastolic ratio), time to 50% peak, time to 50% decay (TD50), and tau of decay.

In a separate set of experiments isolated cardiomyocytes were loaded with fluo4-AM as described previously[17] and spontaneous Ca^{2+} sparks were assessed using confocal microscopy. All experiments were performed at 37 °C in NT solution containing 2 mM Ca^{2+}. The control group consisted of myocytes isolated from age-matched controls. Data were viewed using Ionwizard and ImageJ (NIH) and analysed with a combination of custom-built macros and Sparkmaster. Amplitude and morphology of five Ca^{2+} transients per cell were averaged to give final result for that cell. Sparks were treated as separate data points.

2.2 Statistical analysis

Data were analysed using IBM SPSS Statistics in two ways. Firstly, using our field's common statistical approach of treating all data points as though they were independent. For each variable, an independent samples *t*-test was performed comparing all the cells from HF hearts against all the cells from control hearts.

The second analysis used hierarchical statistical techniques. These techniques are designed for situations where values may be clustered, with a possibility of similarity within each cluster.[13] We have designed cost-free methods which we have made available for the reader to use for similar analyses (see Discussion section).

2.2.1 Quantifying clustering

The amount of clustering can be quantified by the intraclass correlation coefficient (ICC). This varies between 0% for data showing no clustering, and 100% for intensely clustered data (i.e. each cluster contains multiple identical values, but different clusters have different values). This measure represents the proportion of total variability in an outcome measure that is attributable to the isolation of origin. Illustrative datasets with high, moderate, and low values are shown in *Figure 2*. The method of calculation is shown in Supplementary material online.

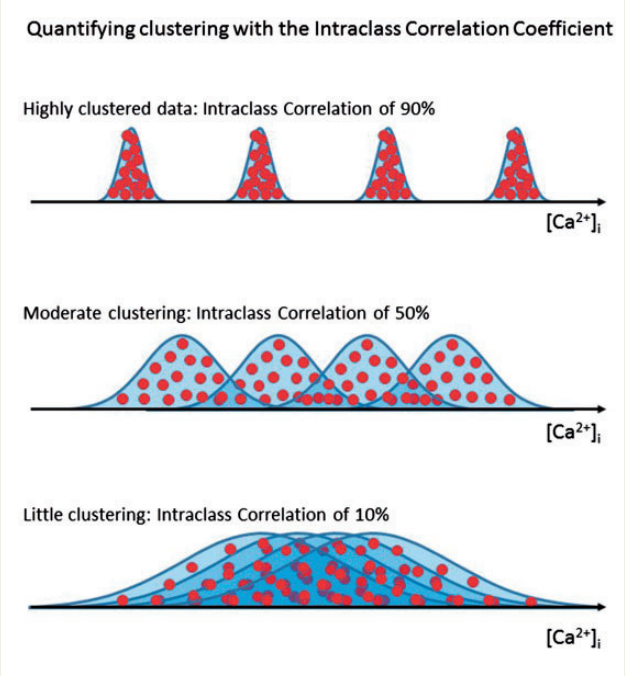

Figure 2 Intraclass correlation in cardiomyocyte studies. Examples of data with high, moderate, or low intraclass correlations (ICCs). Intracellular calcium concentration ($[Ca^{2+}]_i$) from individual cells (red dots) is clustered into histograms making up the data distributions in individual rat hearts. In the top panel, there is a high ICC with data clustered tightly in each heart. In the middle panel, there is an intermediate ICC and in the bottom panel there is a low ICC with a high proportion of the overall variability coming from each individual heart. A larger correction to common statistical techniques is required in data with a high ICC.

2.2.2 Correction of the analysis based on ICC

If the data show no clustering, then the hierarchical model works effectively identically to the commonly used statistical test, treating all the data points as independent. If, on the other hand, the data are tightly grouped within each cluster, with relatively large separation between clusters (as might happen when each day's isolation was internally homogeneous but differed from other days) then a relatively large correction would need to be applied. The hierarchical test quantifies the amount of clustering and applies the appropriate correction to the statistical significance test (*Figure 3*), further explanation of hierarchical testing is shown in Supplementary material online (see Supplementary methods and *Figure S2*).

2.2.3 Effective sample size

The hierarchical approach also provides a simple method of augmenting conventional sample size calculation to explain the fact that with m cells from each of n rats the true sample size is less than $n \times m$ but more than n. The correction required depends on the degree of clustering as quantified by the ICC. With low levels of clustering, where each cell's behaviour is independent of the animal it has come from, effective sample size approaches $n \times m$. With high degrees of clustering, where all cells from a single animal behave identically, effective sample size approaches n. The calculation of effective sample size is as follows:

For n animals, each with m cells

$$Effective\ \ Sample\ \ Size = \frac{n \times m}{1 + (m - 1) \times ICC}$$

For example, with 10 animals and 20 cells from each, with an ICC of 50%, effective sample size is neither 200 nor 10, and is calculated as follows:

$$Effective\ \ Sample\ \ Size = \frac{200}{1 + (19) \times 0.5} \approx 19$$

This has important implications for the interpretation of conventional power calculations in the context of hierarchical data. Power calculations are used to work out the appropriate sample size for experiments given a working knowledge of likely effect size and variance obtained from pilot data. In the context of a hierarchical data structure, the sample size given by such power calculations should be considered to be the *Effective Sample Size*. A rearrangement of the above equation would give the true sample size required:

$$n \times m = \text{Effective Sample Size} \times (1 + (m - 1) \times ICC)$$

2.2.4 Testing whether the hierarchical statistical model significantly improved the fit

We tested whether the hierarchical statistical model produced a significantly better fit to the data than the commonly used statistical approach, using the χ^2 test of the change in -2 Log Likelihood (χ^2-2LL). This is the recommended method for comparisons of this kind.[18–20]

2.2.5 Transforming non-normal distributions

Initial exploration of data showed that whilst variables pertaining to Ca^{2+} transients were symetrically distributed, spark morphological variables were skewed as has been shown previously.[21] The logarithmic transformations of morphological characteristics were therefore used to improve this lack of symmetry (see Supplementary material online, *Figure S3*). These transformed variables are referred to as LogAmp, LogFWHM, LogFDHM for the logarithm of spark amplitude, full width at half maximum (FWHM), full duration at half maximum (FDHM).

2.2.6 Running the two analysis approaches

Ca^{2+} transient and spark measures were the dependent variables in the statistical models. The only independent variable was the presence or absence of HF. Each model used was assessed for validity by ensuring predicted values closely corresponded to those observed. Residuals were assessed for normality and symmetry. Estimated marginal means were used to assess significance between control and HF groups.

3. Results

Ca^{2+} transient variables were measured from 76 cells from 10 control rats and 79 cells from 10 HF rats. Ca^{2+} spark variables were measured in 344 sparks from 17 cells from 7 control rats and 352 sparks from 22 cells from 5 HF rats.

3.1 Multiple cells from each of several rats: Ca2+ transient morphology

For all Ca^{2+} transient morphology variables, the hierarchical model produced a statistically significantly better fit than the commonly used analysis method (*Table 2*). The magnitude of the intraclass correlation (degree of clustering) varied between 12 and 47% (*Table 2*, ICC column).

For all the comparisons between HF and control, the commonly used analysis method reported a more statistically significant difference than the better-fitting hierarchical analysis. The mechanism of this was that the commonly used statistical method produced a smaller value for the standard error of the difference between HF and control (*Table 2*).

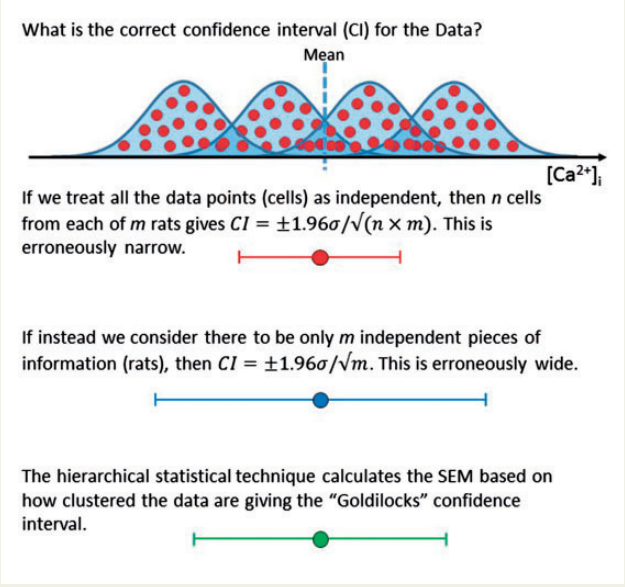

Figure 3 Correction for clustering produced by hierarchical tests. Description of correction to the confidence interval for the difference between control and HF for clustered data that is produced using hierarchical tests.

Figure 4 shows the 95% confidence interval calculated, by the two approaches, for the difference in time to peak between HF and control. The commonly used analysis method gives a different conclusions because it has an artificially small 95% confidence interval, causing the data to meet the criteria for statistical significance.

3.2 Multiple sparks from each of multiple cells from multiple rats

For sparks there is potential for an additional level of clustering, because the multiple sparks from each cell could be more similar to each other than the sparks from other cells of the same rat.

The hierarchical analysis was carried out twice (*Table 3*): once allowing only for clustering at the animal level and once allowing for clustering at both the cell and the animal levels. We tested whether the incorporation of the additional (cell) level improved the fit as defined by the χ^2-2LL. Where the fit was improved significantly, the additional level was incorporated into the analysis (methods shown in Supplementary appendix).

The *P*-values were substantially different between the commonly used method and the hierarchical analyses. Again, the hierarchical analyses fitted better for each variable. The cause of the difference in *P* values was that the hierarchical models calculated larger values for the standard error.

Even minor clustering made a substantial difference to the *P* value. For example, for logFWHM, even with only 8.5% clustering (as quantified by the ICC), the conventional approach which assumes no clustering reports a *P* value of 0.02, whereas the hierarchical model reports a *P* value of 0.78.

LogAmp showed significant clustering both within each cell and within rats (*Table 3*, *Figure 5A, B*). *Figure 5C* shows the 95% confidence interval calculated, by the two approaches, for the difference in LogAmp between HF and control. The artificially small confidence interval produced by the common test makes the difference between the groups appear highly significant, but when the hierarchical data structure is taken into account, there is no significant difference.

Looking at the best-fitting analysis method (*Table 3*, defined by χ^2-2LL criterion), it emerges that none of the spark variables is in fact significantly different between HF and control. LogAmp and LogFDHM appeared to be different between HF and control using the commonly used analysis approach, but this was because of an artefactually small standard error through failing to account for clustering.

Table 2 Analysis of Ca^{2+} transient morphology variables using standard and hierarchical statistical tests

	Clustering of data (ICC) (%)	Common test of HF vs. control		Hierarchical test of HF vs. control		Comparison of goodness of fit (common vs. hierarchical)
		Std error of difference	*P*-value	Std error of difference	*P*-value	
Diastolic ratio	27	0.0120	0.248	0.0203	0.623	<0.001***
Peak systolic ratio	23	0.0396	0.004**	0.0653	0.046*	0.002**
Transient amplitude (F/F_0)	21	0.0306	<0.001***	0.0493	0.010*	0.006**
Time to 50% peak (ms)	12	0.520	0.018*	0.716	0.109	0.021*
Time to 50% decay (ms)	44	4.86	0.444	98.2	0.400	<0.001***
Tau (ms)	47	8.55	0.535	17.7	0.424	<0.001***

Elements of the analysis of each variable are shown. The independent-samples *t*-test is shown as the common test used to compare cellular data. The clustering of data measured by calculating the intraclass correlation (ICC) is shown for each variable. The hierarchical technique is more appropriate with each variable as indicated by better goodness of fit (as measured by χ^2-2LL test). When using the more appropriate hierarchical test the standard error increases and the *P* values also increase making significant differences less likely. For the time to 50% peak this results in a change from a significant test to a non-significant test. Note that the change in standard error is greatest where the ICC is larger.
*$P < 0.05$.
**$P < 0.01$.
***$P < 0.001$. There were 76 cells from 10 control rats and 79 cells from 10 HF rats.

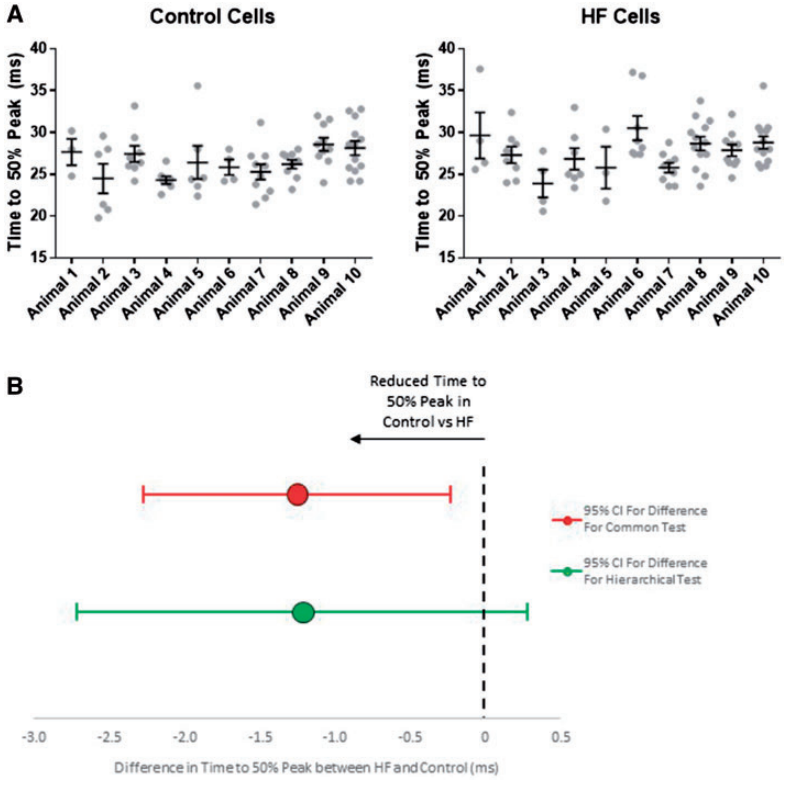

Figure 4 Clustering of time to 50% peak transient amplitude. (*A*) Clustering of time to peak is shown with the time to 50% peak for transient in each cell (grey dots) shown for each rat in HF and control. The mean and standard error bars are shown for each rat. By eye the data appear clustered and this is confirmed by the statistical testing in *Table 2*. (*B*) The mean difference and confidence interval for the difference are shown which indicates why the difference between HF and control becomes non-significant with the corrected confidence intervals of the hierarchical test (green bars) compared with the uncorrected confidence intervals of the common test (red bars). There were 76 cells from 10 control rats and 79 cells from 10 HF rats.

Table 3 Spark data analysed by hierarchical vs. standard statistical tests

	Clus-tering of data (ICC) (%)	Common method		Two-level hierarchy		Three-level hierarchy		Comparison of goodness of fit (common vs. hierarchical)
		Standard error	*P*-value	Standard error	*P*-value	Standard error	*P*-value	
Variables that have a single value per *cell*, and have clustering within *animal*.								
Spark freq (sp/100 μm/s)	24	0.537	0.886	0.664	0.540	N/A	N/A	0.048*
Variables that have a single value per *spark*, and show clustering within *cell* and within *animal*.								
LogAmp ($\Delta F/F_0$)	58	9.00×10^{-3}	<0.001***	0.0320	0.001**	0.0641	0.239	<0.001***
Variables that have a single value per *spark*, and show clustering within *cell*. There is minimal additional variability between rats such that analysis at cell level hierarchy is most appropriate. Here a comparison of goodness of fit between common and two-level test is significant ($P < 0.05$) but the same comparison between the two-level hierarchy and three-level hierarchy is not significant ($P > 0.05$).								
LogFDHM (ms)	8	0.0168	0.023*	0.0289	0.778	N/A	N/A	<0.001***
LogFWHM (μm)	7	0.0113	0.381	0.0185	0.782	N/A	N/A	<0.001***

Analysis using the common test vs. hierarchical methods is shown. Analysis using a rat-level hierarchy is most appropriate for spark frequency as only a single value is available for each cell. For variables describing spark morphology either an analysis that accounts for both cell-level and rat-level hierarchy is appropriate (as for spark amplitude) or where there is little additional variability per rat, and the goodness of fit is not further improved by a three-level hierarchy, analysis with a cell-level hierarchy is most appropriate (as for FDHM and FWHM). The hierarchical test out-performs the common test for each variable.

*$P < 0.05$.

**$P < 0.01$.

***$P < 0.001$. There were 344 sparks from 17 cells from 7 control rats and 352 sparks from 22 cells from 5 HF rats.

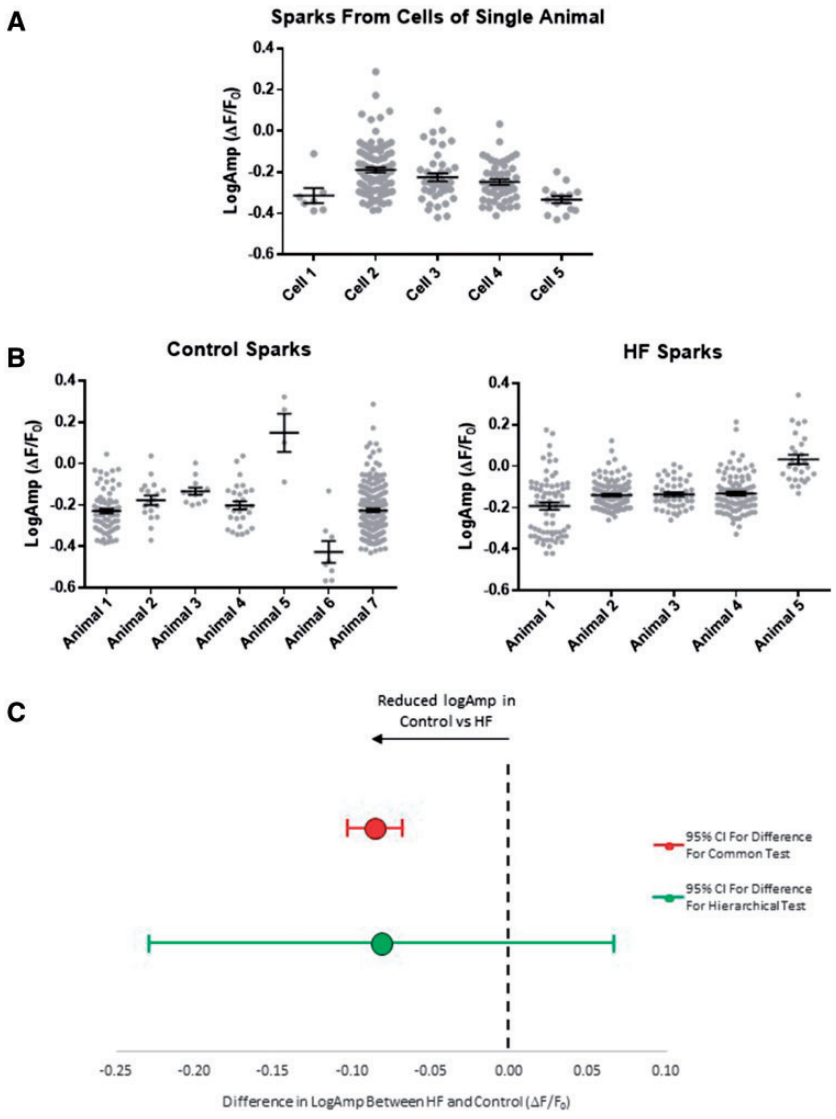

Figure 5 Multi-level clustering of spark logAmp. (*A*) Clustering of spark LogAmp is shown within each of five cells tested from a single rat heart. Each individual spark's logAmp is shown as a grey dot. The mean and standard error bars are shown for each cell. By eye the data appear clustered at this level and this is confirmed in *Table 3*. (*B*) Spark data may also be further clustered within individual rats. Here sparks are shown grouped by rat. As confirmed in *Table 3* there is also clustering at the level of the rat. (*C*) With two levels of clustering to correct for with a large ICC (58%) a large correction to confidence intervals is required. With the common test there is a highly significant difference comparing logAmp in HF and control (red bars). With the appropriate correction to confidence interval (green bars) it is clear there is no significant difference. There were 344 sparks from 17 cells from 7 control rats and 352 sparks from 22 cells from 5 HF rats.

4. Discussion

Different studies of Ca^{2+} fluxes in HF cells have given opposite results for aspects of Ca^{2+}handling in isolated cardiomyocytes. The usual interpretation of this is that the direction of the effect may vary with choice of HF model and species. Our present study, however, shown in *Table 1* and Supplementary material online, *Tables S1–3* that even within a single HF model, studies already show opposite direction of effects.

Our study suggests that that these contradictory directions of effect could be a manifestation of the subtle but important defect in our conventional statistical analysis, whereby we assume that all the data points are independent with no significant clustering. In reality for many biological variables, repeated measurements (e.g. of sparks) from the same cell are more similar than those of other cells, and measurements from multiple cells of the same animal are in turn more similar than measurements from different animals. The consequence is when we have multiple measurements we have less meaningful data than the simple number of data points would suggest. This problem, the ability of large sample numbers to deceive us when we do not recognize clustering, has been previously identified in the field of neuroscience.[12,14]

While this study focuses on the clustering of transient data from different myocytes isolated from the same rat, as well as the clustering of spark data within each cell, this should not be seen as a reason to limit the use of hierarchical statistics to these settings. We have found significant clustering in other aspects of isolated myocyte properties, such as cell size.[22] We have also found that these techniques are essential for

robust analysis of other models. For example in a recent study where multiple cardiac slices were taken from each dog, significant clustering of slice contractility data meant that data were most appropriately analysed using hierarchical techniques.[23] It would be safest, therefore, for the investigator to assume that any data with a hierarchical structure should be analysed in this way unless the hierarchical tests show non-significant clustering. We have made such testing freely available (below).

4.1 Hierarchical analysis exposes false positive findings

The transients of an individual cell are known to be almost identical and therefore, it is already conventional to recognize this clustering by treating the entirety of transient data from one cell as a single data point by averaging data from several transients.

Sparks, on the other hand, vary a great deal from one to the next within the same cell. This stochastic behavior results from the subcellular physiology of ryanodine receptors, which are present in variable configurations with variable opening probabilities and dynamics. The convention in our field has therefore been to not average spark data into a single value per cell, but consider each spark as a separate data point.

Our study indicates that the spark data show strong clustering at the cell level, which means that it is wrong to count the individual sparks for statistical purposes as though they were independent.

Strikingly, although the degree of clustering (ICC) varied from one variable to another, it was always significant for all of the variables. Moreover, even small values of the clustering measure (ICC) turn out to be surprisingly important. For example, spark logFDHM, which has an ICC of just 8.5%, changed its result from being clearly significant with the common analysis ($P = 0.023$), to being not at all significant ($P = 0.778$) with the hierarchical analysis. Many variables had far higher degrees of clustering (*Tables 2 and 3*), which means they are correspondingly even more susceptible to false positive results from the common analysis.

We were surprised to find that, after recognizing the clustering using hierarchical analysis, the only significant difference between HF and control is a higher Ca^{2+} transient amplitude. This is due to increased systolic Ca^{2+} without increase in diastolic Ca^{2+}. If we had used standard *t*-tests alone the conclusion of this study would have been that Ca^{2+} transients in HF exhibited both higher amplitude and a longer time to peak and that these changes might result from sparks (the elemental subunits of transients) which were more prolonged and of greater amplitude, giving a mechanistic explanation for our finding of higher transient amplitude which is not supported by the data.

4.2 Cruder alternatives to hierarchical statistics—aggregation and disaggregation

Because hierarchical statistics are sometimes considered difficult to implement, some workers have tried the two alternative approaches.

Aggregation is the approach of lumping multiple data points into one, typically by averaging. This is already done routinely for Ca^{2+} transients at the cell level but not the rat level. The disadvantage is that applying this approach universally is excessively conservative, and would require very large studies to test hypotheses that could be tested more frugally if hierarchical statistics were used. For example, the 696 sparks in our study would have been collapsed into just 12 averages, one for each rat, if we had indiscriminately aggregated.

Results of aggregated statistical analysis in this study are shown in Supplementary material online, *Tables S4*, *S5* and are similar to the outcomes of the hierarchical tests but with overestimation of *P*-values and standard errors.

Disaggregation is the opposite approach, splitting the data so that each data point is considered to be independent. This is the commonest approach used in our field but has the scientific disadvantage that the statistics conducted assuming this independence give a falsely small standard error, falsely small confidence interval and therefore falsely small *P* value for the difference between groups.

An example of the catastrophic effect of disaggregation is that in our data the hierarchical model shows only 2 of 10 comparisons between HF and control to be statistically significant; in contrast the commonly-used approach (disaggregation of data) shows five comparisons to be statistically significant: a 150% overstatement.

A further challenge is that in the cellular laboratory there can be 'good isolation days' and 'bad isolation days'. On good days large numbers of healthy cells are available for experimentation but on bad days smaller numbers of less healthy cells are available. If we aggregate, we give excessive weight to the likely unhealthier cells from the bad days. If we disaggregate, we give excessive weight to the likely healthier cells from the good days because they are more numerous.

The social sciences have long known of this problem and developed standardized approaches to avoid wasting resources pursuing false leads.[13,24,25] We describe a similar problem in studies of isolated cardiac myocytes but we also describe a solution and provide cost-free simple tools which can be used to produce robust statistical results for comparisons involving clustered data.

4.3 Hierarchical tests are not merely a method of *P*-value adjustment

There is a general focus on the *P*-value as the only important outcome of significance testing in the biological literature.[26] This approach leads to an overreliance on the apparent binary outcome of $P < 0.05$ vs. $P > 0.05$. In addition, the *P*-value gives the reader no indication of the magnitude of the effect and therefore its biological (rather than statistical) significance. An emphasis on point estimates and their precision (e.g. mean and 95% confidence interval) can prevent these issues.

If the hierarchical data structure is not considered, the 95% confidence intervals of an estimated mean are excessively narrow. Appropriate corrections can be made when performing hierarchical analysis. The magnitude of the required correction, as with other aspects of hierarchical testing, depends on the ICC. The estimated mean and 95% confidence intervals for each variable tested in this study, with and without hierarchical correction, are shown in Supplementary material online, *Tables S6* and *S7*.

4.4 A simple, cost-free, open source solution for statistical testing in cardiomyocyte studies

There are commercial software packages such as IBM SPSS Statistics, SAS (SAS Institute), STATA (StataCorp) which can do these hierarchical statistics very well. Readers can download the scripts which we show in Supplementary material online, Supplementary methods section. However, some readers may not have a particular commercial software package available or may be working with collaborators who do not.

We therefore present simple steps that any researcher can use without cost to perform these hierarchical statistics.

Step 1. Download the relevant files from Supplementary material online into a single working directory. These should include: *ratonly.Rmd,*

ratandcell.Rmd, *Hierarchical Transient analysis with Rat-Level Clustering.xlsx*, and *Hierarchical Spark analysis Cell & Rat-Level Clustering.xlsx*.

Step 2. Modify the relevant Microsoft Excel file. If the data have the potential for clustering only at one level, for example, at the rat level, use the layout in Supplementary material online, *Hierarchical Transient analysis with Rat-Level Clustering.xlsx*. For data such as spark values, which has potential for an additional (e.g. cellular) level of clustering, use the layout in *Hierarchical Spark analysis Cell & Rat-Level Clustering.xlsx*.

Cells should be numbered sequentially even if they come from different rats, that is, if cells 1–10 come from rat 1, then cells from rat 2 should be numbered from 11 upwards and so on.

Column titles can be changed to reflect what your relevant hierarchical structure is—for example, dogs and heart slices or cell line and culture number.

Step 3. Download and install R for Windows, Mac, or Linux from https://www.r-project.org/ (14 August 2017, date last accessed).

Step 4. Download and install RStudio from https://www.rstudio.com/products/rstudio/download/ (14 August 2017, date last accessed)—this is a more user-friendly interface for R. Choose *RStudio Desktop—Open Source License*.

Step 5. Run RStudio and install the two extra packages we will need ('lmerTest' for the statistical tests and 'readxl' for reading in the Excel spreadsheets). This can be done using the 'Tools → Install packages' option in RStudio, and typing 'lmerTest, readxl' (first letter of lmer is lower case 'L') in the 'Packages' field, or by typing the following commands in the 'Console' window (R Console) and pressing Enter after each command:

```
install.packages('lmerTest')
install.packages('readxl')
```

Step 6. Open the code corresponding to the Excel file you are working with in RStudio by clicking *File* menu and *Open File*. Open ratonly.Rmd if working with clustering at one level and ratandcell.Rmd if working with clustering at two levels.

Step 7. In the top right hand corner of the window in which the code has opened there is a *Run* command. Click this and then select *Run All*.

Step 8. Once the code has run, which may take several minutes, scroll down to the bottom of the same window where three tables (for ratonly.Rmd) or five tables (for ratandcell.Rmd). These tables show the following:

(1) The first table shows the outcomes of the common test (*P* value and standard error for the difference), the degree of clustering (ICC), the equivalent outcomes for the hierarchical test, and whether the hierarchical test is a better fit for the data under the column 'Superior fit'. If the *P* value in this column is < 0.05, there is sufficient clustering to make hierarchical techniques necessary. For ratandcell.Rmd, both cell-level hierarchical test (labelled group-level) and rat-and-cell level hierarchical test (labelled Parentgroup-group level) results are shown.

(2) The second table shows the point estimate, that is, the estimated group mean when taking into account the hierarchical data structure, as well as the standard error and the lower and upper confidence intervals for each variable for each condition (HF and control).

(3) The third table shows the pairwise comparisons. If there are only two conditions (e.g. HF and control) the *P*-values in this table are the same as the first table. This table becomes particularly relevant if there are >2 conditions since comparisons each pair of groups (A vs. B; B vs. C; A vs. C) may be desired. As with an ANOVA, these group-wise comparisons are only relevant if the overall test shown in the first table is significant.

(4) The fourth and fifth tables are only output by ratandcell.Rmd and are the equivalent of the second and third tables but for the rat-and-cell level hierarchical test.

The scripts we provide are sufficient to perform the analyses in this paper. Once readers are comfortable doing this, they can extend the R script to allow more levels of clustering or to include other variables such as gender, animal weight, or cell size.

Moreover, these scripts are equally suited outside the cardiomyocyte arena and even to clinical data.

For users who do not have R or RStudio installed, but who wish to view the source code and sample output, we have supplied two supplementary HTML (ratonly.nb.html and ratandcell.nb.html) files which can be opened in any browser and serve as manuals to the analyses.

4.5 Limitations

Even though the hierarchical test describes the data much better (as shown by the χ^2-2LL criterion), we should not assume that they are the ideal solution. It is merely a significant improvement for which we provide a straightforward method for implementation. Ideally researchers would engage with an expert statistician at an early phase of study preparation. Alternative models for performing hierarchical statistics are available including Generalized Estimating Equations (GEE), although we favour Mixed Models (with random coefficients) as this gives greater flexibility in the modelling process and is better at handling missing data.

Data transformations are frequently necessary, such as log transformation for the spark data which have a marked positive skew. In addition, even following valid transformation the meaning of the original data must be questioned. For example, some of the skewed distribution relating to variables describing spark morphology relates to the confocal line-scanning technique which is likely to miss the center of each spark, thus artefactually increasing the number of smaller sparks.[21,27,28]

5. Conclusions

The conventional approach used in our field to analyse cardiomyocyte data has a marked tendency to overstate the statistical significance of differences. This could be why some studies shows significant results in one direction and other studies show equally convincing significant results in the other direction.

We present simple steps that any researcher can implement to identify and allow for clustering, at the rat and cell level if need be, permitting them to use all their data points and yet obtain statistically valid results.

Supplementary material

Supplementary material is available at *Cardiovascular Research* online.

Acknowledgements

Thanks to Liam Couch and Filippo Perbellini for their help in testing the ease of use of the RStudio Code and to Peter O'Gara for his technical expertise in cell isolation.

Conflict of interest: none declared.

Funding

This work was supported by the Wellcome Trust (WT092852). M.B.S. is supported by a National Institute of Health Research Clinical Lectureship award (#2670). This work was also supported by the British Heart

Foundation, the ElectroCardioMaths Programme of the Imperial Centre for Cardiac Engineering and the National Institute for Health Research (NIHR Imperial Biomedical Research Centre).

References

1. Epstein F, Morgan J. Abnormal intracellular modulation of calcium as a major cause of cardiac contractile dysfunction. *N Engl J Med* 1991;**325**:625–632.
2. Johns TN, Olson BJ. Experimental myocardial infarction. I. A method of coronary occlusion in small animals. *Ann Surg* 1954;**140**:675–682.
3. Lyon AR, MacLeod KT, Zhang Y, Garcia E, Kanda GK, Lab MJ, Korchev YE, Harding SE, Gorelik J. Loss of T-tubules and other changes to surface topography in ventricular myocytes from failing human and rat heart. *Proc Natl Acad Sci U S A* 2009;**106**:6854–6859.
4. Cheung JY, Musch TI, Misawa H, Semanchick A, Elensky M, Yelamarty RV, Moore RL. Impaired cardiac function in rats with healed myocardial infarction: cellular vs. myocardial mechanisms. *Am J Physiol* 1994;**266**:C29–C36.
5. Soppa GKR, Lee J, Stagg MA, Felkin LE, Barton PJR, Siedlecka U, Youssef S, Yacoub MH, Terracciano CMN. Role and possible mechanisms of clenbuterol in enhancing reverse remodelling during mechanical unloading in murine heart failure. *Cardiovasc Res* 2008;**77**:695–706.
6. Huang B, Wang S, Qin D, Boutjdir M, El-Sherif N. Diminished basal phosphorylation level of phospholamban in the postinfarction remodeled rat ventricle: role of beta-adrenergic pathway, Gi protein, phosphodiesterase, and phosphatases. *Circ Res* 1999;**85**:848–855.
7. Zhang XQ, Moore RL, Tenhave T, Cheung JY. [Ca^{2+}]*i* transients in hypertensive and postinfarction myocytes. *Am J Physiol* 1995;**269**:C632–C640.
8. Anand IS. Ventricular remodeling without cellular contractile dysfunction. *J Card Fail* 2002;**8**:S401–S408.
9. Holt E, Tønnessen T, Lunde PK, Semb SO, Wasserstrom JA, Sejersted OM, Christensen G. Mechanisms of cardiomyocyte dysfunction in heart failure following myocardial infarction in rats. *J Mol Cell Cardiol* 1998;**30**:1581–1593.
10. Wehrens XHT, Lehnart SE, Marks AR. Intracellular calcium release and cardiac disease. *Annu Rev Physiol* 2005;**67**:69–98.
11. Cheng Y, Li W, Mcelfresh TA, Chen X, Berthiaume JM, Castel L, Yu X, Wagoner DRV, Chandler MP, Ta M, Jm B, Wagoner V, Mp C. Changes in myofilament proteins, but not Ca^{2+} regulation, are associated with a high-fat diet-induced improvement in contractile function in heart failure. *Am J Physiol Hear Circ Physiol* 2011;**301**:1438–1446.
12. Sikkel MB, Macleod KT, Gordon F. Letter by Sikkel et al. regarding article, 'late sodium current inhibition reverses electromechanical dysfunction in human hypertrophic cardiomyopathy'. *Circulation* 2013;**128**:e156.
13. Hox JJ, *Multilevel Analysis: Techniques and Applications*. 2nd ed. Hove, United Kingdom: Routeledge Academic; 2010.
14. Lazic SE. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci* 2010;**11**:5.
15. Coppini R, Ferrantini C, Yao L, Fan P, Lungo MD, Stillitano F, Sartiani L, Tosi B, Suffredini S, Tesi C, Yacoub M, Olivotto I, Belardinelli L, Poggesi C, Cerbai E, Mugelli A. Late sodium current inhibition reverses electromechanical dysfunction in human hypertrophic cardiomyopathy. *Circulation* 2013;**127**:575–584.
16. Sikkel MB, Kumar S, Maioli V, Rowlands C, Gordon F, Harding SE, Lyon AR, Macleod KT, Dunsby C. High speed sCMOS-based oblique plane microscopy applied to the study of calcium dynamics in cardiac myocytes. *J Biophoton* 2016;**9**:311–323.
17. Sikkel MB, Collins TP, Rowlands C, Shah M, O'gara P, Williams AJ, Harding SE, Lyon AR, MacLeod KT. Flecainide reduces Ca^{2+} spark and wave frequency via inhibition of the sarcolemmal sodium current. *Cardiovasc Res* 2013;**98**:286–296.
18. Field A, *Discovering Statistics Using IBM SPSS Statistics*. 3rd ed. London: Sage; 2009.
19. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Stat Med* 2002;**21**:3291–3315.
20. Thum Y. Hierarchical linear models for multivariate outcomes. *J Educ Behav Stat* 1997;**22**:77–108.
21. Cheng H, Song LS, Shirokova N, González A, Lakatta EG, Ríos E, Stern MD. Amplitude distribution of calcium sparks in confocal images: theory and studies with an automatic detection method. *Biophys J* 1999;**76**:606–617.
22. Sikkel MB. Arrhythmogenic sarcoplasmic reticulum calcium leak in isolated ventricular cardiomyocytes – changes in heart failure and mechanisms of pharmacological modulation. *Ph.D. Thesis*. Imperial College London, 2014. pp. 98–99.
23. Perbellini F, Watson SA, Scigliano M, Alayoubi S, Tkach S, Bardi I, Quaife N, Kane C, Dufton NP, Simon A, Sikkel MB, Faggian G, Randi AM, Gorelik J, Harding SE, Terracciano CM. Investigation of cardiac fibroblasts using myocardial slices. *Cardiovasc Res* 2017; doi: 10.1093/cvr/cvx152.
24. Park S, Lake ET. Multilevel modeling of a clustered continuous outcome. *Nurs Res* 2005;**54**:406–413.
25. Raudenbush SW, Bryk AS, *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA: Sage; 2002.
26. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. *Nat Meth* 2015;**12**:179–185.
27. Izu LT, Wier WG, Balke CW. Theoretical analysis of the Ca^{2+} spark amplitude distribution. *Biophys J* 1998;**75**:1144–1162.
28. Shkryl VM, Blatter LA, Ríos E. Properties of Ca^{2+} sparks revealed by four-dimensional confocal imaging of cardiac muscle. *J Gen Physiol* 2012;**139**:189–207.
29. Sande JB, Sjaastad I, Hoen IB, Bøkenes J, Tønnessen T, Holt E, Lunde PK, Christensen G. Reduced level of serine(16) phosphorylated phospholamban in the failing rat myocardium: a major contributor to reduced SERCA2 activity. *Cardiovasc Res* 2002;**53**:382–391.
30. Lee J, Stagg MA, Fukushima S, Soppa GKR, Siedlecka U, Youssef SJ, Suzuki K, Yacoub MH, Terracciano CMN. Adult progenitor cell transplantation influences contractile performance and calcium handling of recipient cardiomyocytes. *Am J Physiol Heart Circ Physiol* 2009;**296**:H927–H936.
31. Maczewski M, Mackiewicz U. Effect of metoprolol and ivabradine on left ventricular remodelling and Ca^{2+} handling in the post-infarction rat heart. *Cardiovasc Res* 2008;**79**:42–51.
32. Kaprielian R, Wickenden AD, Kassiri Z, Parker TG, Liu PP, Backx PH. Relationship between K^{+} channel down-regulation and [Ca^{2+}]*i* in rat ventricular myocytes following myocardial infarction. *J Physiol* 1999;**517**:229–245.
33. Loennechen JP, Wisløff U, Falck G, Ellingsen O. Effects of cariporide and losartan on hypertrophy, calcium transients, contractility, and gene expression in congestive heart failure. *Circulation* 2002;**105**:1380–1386.
34. Yoshida H, Tanonaka K, Miyamoto Y, Abe T, Takahashi M, Anand-Srivastava MB, Takeo S. Characterization of cardiac myocyte and tissue beta-adrenergic signal transduction in rats with heart failure. *Cardiovasc Res* 2001;**50**:34–45.
35. Saraiva RM, Chedid NGB, Quintero HCC, Díaz GLE, Masuda MO. Impaired beta-adrenergic response and decreased L-type calcium current of hypertrophied left ventricular myocytes in postinfarction heart failure. *Braz J Med Biol Res* 2003;**36**:635–648.
36. Loennechen JP, Wisløff U, Falck G, Ellingsen Ø. Cardiomyocyte contractility and calcium handling partially recover after early deterioration during post-infarction failure in rat. *Acta Physiol Scand* 2002;**176**:17–26.
37. Lyon AR, Bannister ML, Collins T, Pearce E, Sepehripour AH, Dubb SS, Garcia E, O'gara P, Liang L, Kohlbrenner E, Hajjar RJ, Peters NS, Poole-Wilson PA, Macleod KT, Harding SE. SERCA2a gene transfer decreases sarcoplasmic reticulum calcium leak and reduces ventricular arrhythmias in a model of chronic heart failure. *Circ Arrhythm Electrophysiol* 2011;**4**:362–372.
38. Ait Mou Y, Toth A, Cassan C, Czuriga D, de Tombe PP, Papp Z, Lacampagne A, Cazorla O. Beneficial effects of SR33805 in failing myocardium. *Cardiovasc Res* 2011;**91**:412–419.