(A plan to reveal) Extreme undisclosed analytical flexibility in HRV with automated p-mining software

Heathers JAJ, Holcombe A.

INTRODUCTION

- Heart rate variability (HRV) is the study of the beat-to-beat variability in the heart rate, which is a consequence of the immediate state of the autonomic nervous system. It is simple to collect, inexpensive, and is central to popular theories of brain-body integration (Thayer and Lane, 2000; Porges, 1992). It is one of the most widely used psychobiological dependent variables.

- There are a series of problems with HRV as it is typically employed. HRV measures which are used to reflect the function of the sympathetic nervous system may not reflect the underlying physiology (Goldstein et al. 2013; Moak et al. 2007; Heathers, 2013), multiply-employed methods of HRV analysis may be mathematically redundant (Burr, 2007; Heathers, 2013), and HRV may have a poor relationship to the desired underlying physiology of measurement either by necessity or without a correction for breathing.

- An additional, previously undocumented problem exists – *that flexibility within analytical methods allows any finding to be presented as significant*. The last broad statement on methodology typically cited was published in 1997, and its recommendations are frequently diluted. There is no well enforced formal analysis standard. As a consequence, individual analysis decisions are often made on an ad hoc basis. These decisions include but are not limited to methods of exclusion, correction, interpolation, and analysis method. These may be either chose by the experimenters or automated using commercial or custom software.

- With variations in the above, and selective reporting of the 'significant' outcomes of analysis, *our aim here is to demonstrate that any comparison between two groups may be manipulated into significance*, via software which simulates this process of selective analysis and reporting. We have automated the process of producing many different yet plausible analyses from the same data (see also Simmons et al., 2011).

- We expect this analysis flexibility will yield significant p-values from HRV datasets *entirely in the absence* of two additional forms of dishonest researcher behaviour – *data forgetting* (removing outliers or problematic values which ostensibly 'obscure' underlying effects) and *data peeking* (performing multiple analyses as sample collection proceeds, effectively multiplying the number of comparisons examined) – that previously have been shown to guarantee statistical significance (Simmons et al., 2011). This is due to the sheer volume of methods which may be employed, which are several orders of magnitude in excess of what might be expected in traditional social science experiments.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.277v1 | CC-BY 4.0 Open Access | received: 7 Mar 2014, published: 7 Mar 2014

SUMMARY FOR PRE-REGISTRATION, uploaded 7th March, 2014

- To investigate the above, we have developed custom MatLab software for 'p-mining' (aka p-hacking) the EKG signal used to create the HRV dependent variable (PMinerEKG; Heathers, 2014). This software takes raw beat-to-beat intervals, and chooses among the following methods from the literature randomly:

    - an error identification method (none, threshold, raw difference, percentage)
    - a correction method (mean, median, linear or cubic spline replacement)
    - a detrending method (none, polynomial, moving average, loess, rloess, etc.)
    - an 'interpolation' frequency to enable frequency analysis (1,2,4, or 7Hz)
    - a correction for the start of the recorded interval to ensure stability (yes/no)
    - popular time and frequency domain values of HRV (85 in total, counting corrected values as separate)

  This initial method provides for any given dataset more than 400,000 possible pathways for analysis, a number which would be trivial to expand. *All* decisions above can be at least superficially justified on the basis of precedents in the literature for each analysis choice.

METHOD

Three spurious comparisons will be conducted and are outlined below.

1. ***Within subject, at rest, over time.***
   Heart rate variability is frequently recorded in both 5 and 10 minute intervals. Results are similar between both, and are highly reproducible within individuals over both short and long time frames, and stable at rest. Is this p-mining method sufficiently powerful that it can demonstrate significance between the extremely similar records within the first 5 and second 5 minutes of the same participant at rest? In other words, can we demonstrate significant values to support the hypothesis that HRV either goes down or up over time in the absence of any intervention at all? N=50 records of two 5-minute consecutive recordings will be analysed within subjects.
2. ***Within subject, at rest, random times.***
   As above, but in the absence of a time variable. That is, with the two consecutive periods as defined in 1. scrambled within subjects.
3. ***Between subject, at rest.***
   From a publicly available database (Normal Sinus Rhythm Database; physionet.org), records of two equal groups of participants at rest are assigned at random. Thus, there is no real difference between the groups. As above, can statistically significant differences be generated between subjects from randomly-chosen data of high technical quality? Because even with this procedure, for the p=.05 criterion, 5% of random draws should yield a legitimate (not p-mined) statistical difference, this will be repeated with multiple random samples to estimate the type 1 error inflation rate.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.277v1 | CC-BY 4.0 Open Access | received: 7 Mar 2014, published: 7 Mar 2014
SUMMARY FOR PRE-REGISTRATION, uploaded 7th March, 2014

DISCUSSION

It should be possible to demonstrate 'extremely significant' - and entirely spurious - differences between the groups in all the scenarios above simply by executing enough iterations of PMinerEKG.

It is important to note that although this is an extreme scenario , the basic procedure of trying many different analyses may well be what many researchers do. The analysis reflects well the structure of outputs from HRV analysis software, which include not only multiple correction methods and options for analysis, but also multiple methods of frequency and time-frequency analysis (i.e. via Welch's periodogram, AR, Lomb-Scargle Periodogram, etc.), and so on. Many of these parameters can be changed wholesale and analysed by batch, at which point the more traditional 'researcher degrees of freedom', such as data forgetting, 'creative' outlier removal or might be employed.

In fact when analysing HRV the available degrees of freedom are *drastically in excess* of our demonstration. Here, we use every value provided, include no covariates, and perform a simple t-test. However, most psychobiological work has the possibility of using various forms of regression analysis, including (or excluding) covariates, creating between subject groups from psychometric measures (i.e. creating new groups via the score on a scale via median split, standard deviation split, scale cut-off and so on), including poorly understood and mathematically complex methods of quantifying HRV (detrended fluctuation analysis, various measures of entropy, symbolic dynamics, and so on), etc. Moreover, these decisions may be made entirely in good faith and alter a result entirely without the possibility of data forgetting or data peeking.

Possible solutions to reduce 'researcher degrees of freedom' vary by area, but HRV is well positioned to benefit from those which have been previously outlined.

Pre-registration

Biobehavioural models make testable predictions about approach/avoidance behaviour, emotional regulation, the application of fear/stress/shock and so on. In a pre-registered study, outlining both the analysis methods and the expected behaviour of specific HRV measures as dependent variables is recommended.

Of course, the data may behave in unexpected ways, rendering the planned analyses inappropriate. Exploratory (unplanned) analyses should continue to be encouraged to avoid excluding serendipitous discoveries. However, the p-values associated with these unplanned analyses should be taken with a grain of salt.

Data retention/uploading

Heart rate data, especially as RR intervals, is very compact – the entire dataset for a large behavioural experiment will typically have a file size in kilobytes. Thus, storage and uploading presents no technical difficulty. This data is also trivial to analyse, invariably using simple methods which are well understood. Consequently, it is easy to recreate analyses and confirm/question results accompanied by data. Finally, the aggregation of uploaded data is of great benefit for the testing of heartbeat detection algorithms, the identification of electrocardiographic phenomena en masse, the provision of population norms, and so on.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.277v1 | CC-BY 4.0 Open Access | received: 7 Mar 2014, published: 7 Mar 2014
SUMMARY FOR PRE-REGISTRATION, uploaded 7th March, 2014

SUMMARY FOR PRE-REGISTRATION, uploaded 7th March, 2014

Psychological research often involves using crude psychometric variables in abnormal populations to create complicated general models of human capacity – with various degrees of replicability. In such an environment, HRV provides a haven of preserved meaning – at its best, it has good construct validity, derived directly from well understood neurobiological principles, and is also straightforward to calculate and interpret. However, a robust discussion of its analytical flexibility is due, and should be seen as critical to preserve its ability to provide insight into human behaviour.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.277v1 | CC-BY 4.0 Open Access | received: 7 Mar 2014, published: 7 Mar 2014

SUMMARY FOR PRE-REGISTRATION, uploaded 7th March, 2014