SUPPLEMENTARY MATERIAL

Allometric Indexing of LV Mass in Echocardiography: Differences between Males and Females

Anders SAHLÉN, Muh Tyng TEO, Nadira HAMID, Weiting HUANG, Abigail DEL ROSARIO AGPAOA, Ruan WEN, See Hooi EWE, Khung Keong YEO, Zee Pin DING

Supplement Index

Supplementary Methods

Subjects

Healthy male and females without known cardiovascular disease, were recruited from a previously characterized cohort of subjects investigated within the SingHEART study, which in turn was a substudy of the so-called SingHealth Biobank program.[1] Original method for recruitment was predominantly advertisement including posters and local newspapers.

The SingHEART program characterized all subjects using (i) written questionnaire, (ii) fasting blood tests, (iii) electrocardiogram (ECG), (iv) ambulatory blood pressure monitoring (ABPM), (v) continuous ECG monitoring, (vi) activity and sleep tracker, (vii) CT calcium score, (viii) cardiovascular magnetic resonance (CMR), (ix) lipidomics, and (x) genomics analyses.

Exclusion criteria were as follows: (i) any cardiovascular risk factors including hypertension (either based on clinical history, or abnormal ABPM defined as mean 24-hour blood pressure ≥ 130 / 80 mmHg or daytime blood pressure ≥ 135 / 85 mmHg or night time blood pressure ≥ 120 / 70 mmHg), impaired fasting plasma glucose (≥ 6.1 mM; ≥ 110 mg/dL) or diabetes mellitus (fasting plasma glucose ≥ 7.0 mM; ≥ 126 mg/dL), hypercholesterolemia (LDL-C > 4.9 mM; ≥ 190 mg/dL), or coronary calcification (Agatston score ≥ 100 units); (ii) significant arrhythmia including atrial fibrillation and bradycardia; (iii) any respiratory disease including COPD; (iv) any renal disorder including chronic kidney failure; (v) any hepatic disease including known or suspected fatty liver disease and/or raised levels of hepatic transaminases; (vi) any autoimmune or connective tissue disorder; (viii) history of cerebrovascular disease (based on clinical or radiological evidence), (ix) present or historical cancer, or (x) any other acute or chronic medical condition deemed by the investigators to potentially affect results of echocardiograms e.g. by altering the hemodynamic state of the patient or producing adverse cardiac structural change.

Echocardiography

Equipment and Staff

Acquisition of standard two-dimensional echocardiograms used in the present report was performed by 2 sonographers using Philips Epiq 7 (Philips, Andover, MS, USA) and measurements were performed at a dedicated workstation using commercially available software (GE EchoPAC, GE Healthcare, Horten Norway; Agfa HealthCare IMPAX Cardiovascular Suite, Greenville, SC, USA).

Variability Testing

Echocardiographic core laboratory data on inter-observer variability is obtained through an annual validation exercise administered by core lab office. Measurements are made in a blinded fashion by core lab sonographers and are calculated as the deviation from mean across observers, ranging from 3.6% to 9.8% (lowest for internal LV diameter, highest for posterior wall thickness).

Statistical Analysis

Overview of Bayesian modeling

Parameters at different levels in a hierarchical Bayesian model are expressed as coexisting in a joint parameter space. The presence of high-level parameters acts to pull low-level parameters closer together than they would be if there were not a higher-level distribution: so-called shrinkage. A prior distribution is supplied on the top-level parameters, and an entire posterior distribution is inferred across the joint parameter space (Supplementary Figure 1). The posterior distribution demonstrates which parameter values are credible and their uncertainty, given the data.

Sharing of information leads to pooling as the estimation of parameters occurs simultaneously across the levels of the model: means and their variability are estimated together for all observations.

Bayesian parameterization of multi-level models benefit in the same way as conventional mixed models from estimates being informed by data from all observations. However, multiple interesting differences between these 2 also exist: firstly, the entire parameter space of a Bayesian model is estimated in the posterior draw with uncertainty about parameters expressed with a 95% credible interval, in contrast to the random effect term of a mixed model which has no defined standard error (and therefore no 95% confidence interval).[2] Moreover, while shrinkage helps to achieve a conservative estimate for both, the use of a meaningful prior in a Bayesian model reduces the risk of overfitting ("Bayesian posteriors are calibrated by definition").[3] This is relevant in the present study as meaningful priors could be created based on the 4 previously published reports which provide guidance as to what values for a and b can be reasonably expected.[4]

Parameterization of Bayesian models

We applied Bayesian statistics to analyse the role of gender in two different models both based on earlier simulation work done by our group,[4] evaluating existing literature in this area.[5–8] Firstly, a hierarchical model was created where "partial pooling" was allowed across gender for intercept a and slope b. As shown in Supplementary Figure 1a, this model was parameterized based on the allometric equation as relating log LVM (m) to log body height (h) in subjects i = 1, 2 ... 229 given gender $s \in$ {male, female} as $m_{i|s} \sim$ Normal (μ, σ), where $\mu = a_s + b_s \cdot x_{i|s}$, and $\sigma \sim$ Gamma (3.4, 0.2), to accommodate residuals with an expected value of approximately 30 g.[4] The model parameterized priors for intercept (a) and slope (b) as $a_s \sim$ Normal (ω, λ), and $b_s \sim$ Normal (κ, δ), where $\lambda \sim$ Gamma (1.0, 0.2) and $\delta \sim$ Gamma (1.0, 0.2). Hyperpriors were parameterized based on earlier work in this area as $\kappa \sim$ Normal (3.8, 3.0), and $\omega \sim$ Normal (1.7, 3.0). Concretely, the hyperprior for intercept a was designed to accommodate values of approximately 44 – 60 g, and the hyperprior for b was given a mean of 1.7 as informed by available data.[4,8]

Secondly, a “zero pooling” model was created where 2 separate regression lines were fitted to the data. As shown in Supplementary Figure 1b, this model was parameterized as $m_{i|s} \sim$ Normal ($a_s + b_s \cdot x_{i|s}$, σ), where $\sigma =$ Gamma (3.4, 0.2), $a_s \sim$ Normal (1.7, 3.0), and $b_s \sim$ Normal (3.4, 3.0). This created separate priors for males and females in the case of both *a* and *b*, but no hyperpriors and a single posterior draw for the output of the allometric equation and its error term.

Bayes Factor Analysis

Bayesian models were compared by calculating the ratio of their relative credibility—the so-called Bayes factor (BF). This approach can be viewed as an extension of Bayesian hierarchical models where a top-level parameter is added, representing an index for the models. A comparison may be performed of probability for each model conditional on the data after marginalising across the parameters within models. In practical terms, the magnitude of the BF is typically converted to a discrete decision about models at a threshold value of 3.0, above which “substantial evidence” exists in favour of one model over another.[9] Model comparison based on BF has been conceptualised by some as a Bayesian alternative to null hypothesis significance testing—with some similarities in application but also important differences. Of note, the BF is determined by differences in how well the available data fit models including their priors. In contrast, while null hypothesis significance testing in conventional statistics is also influenced by model fit, it relies heavily on the size of the population under study. A type 2 error arises if power is insufficient which leads researchers to reject a conventional model even if it fits the data better.

Simulation

As explained under Results in the main manuscript, when the slope (*b*) was entered as a random effect variable in a conventional mixed model, there was no significant improvement over a model with a random effect only for the intercept and only a fixed effect of *b*. Nonetheless, the Bayesian model comparison based on BF showed that a hierarchical model allowing partial mixing for *b* through a hyperprior (and zero mixing for *a*) gave a considerably better fit to data than a model with zero mixing for both *a* and *b*. To reconcile this apparent contradiction, we performed a simulation to examine the effect of population size on the p-value of the random effect term in the mixed model. The covariance matrix for the study population was used and the proportion of males:females was maintained, with randomly drawn study populations scaled up by multiples of the original study size (numbers of males and females, respectively, in simulated runs was {96, 133}, {192, 266}, {288, 399} etc).

Supplementary Discussion

Regularization and Overfitting

Diagrams in Supplementary Figures 5a – 5b show the fits achieved with different models, both frequentist and Bayesian. The better a model reduces dependence of LVM on body size, the closer its regression line passes to the origin of the chart. Models with best fit were generally the "zero pooling" models and the mixed models, as compared to e.g. Bayesian models with "partial pooling". Two observations that may be pointed out are as follows. (i) Firstly, while the random effect term of mixed models for slope (*b*) was statistically non-significant, the model nonetheless achieved a very close fit to data and minimized dependence of LVM on body size. (ii) Parameterization with zero pooling appeared to provide a closer fit to data, despite the BF analysis favoring the Bayesian model with "partial pooling" across gender. This begs the question of what role overfitting plays for these results and what the impact is of regularization in Bayesian analysis.

When a dataset is divided into 2 subsets and each given its own regression line, the ratio of free parameters available to the amount of information contained in the data can be reasonably expected to increase. The zero pooling model should therefore be able to better capture the structure of the dataset and give a closer fit. However, this does not consider the presence of a prior. The fact that between-model comparison using BF favored the clustered model over the zero pooling model illustrates a particular strength of Bayesian regression: as long as meaningful priors are provided, Bayesian models are less prone to overfitting than frequentist models. Firstly, Bayesian models may be conceptualized as incorporating training and testing within a single model: its priors can be reasonably viewed as the result of training on previous data, with the "present data" (or likelihood) representing a new, held-out dataset for analysis.[10] Secondly, frequentist estimates represent a single, intrinsically optimal fit to the data at hand. E.g. maximum likelihood estimation solves for the unbiased value for population probability that makes the observed data the most likely to have occurred. Confidence limits are subsequently obtained by combining population size with distributional assumptions on the maximum likelihood estimate. As the point estimates obtained may have limited external validity when applied to out-of-sample data, constraints are commonly placed on these to regularize them, in order to avoid optimism and overfitting (e.g. ridge regression or—in its general form—penalized maximum likelihood estimation).[11] While regularization is thus often employed in frequentist statistics to achieve conservative estimates, it may be shown that imposing a Gaussian prior on a regression parameter in a Bayesian model (as done in the present study) is in fact mathematically equivalent to performing ridge regression: the prior acts as a natural regularizer.[12] We believe the conservative estimates obtained in this study using Bayesian statistics are a strength in themselves, given the range of possible optimal values for *b* published in earlier papers in this field. As such, our interpretation of Supplementary Figures 5a – 5b is as follows. (i) Firstly, while conventional mixed models were underpowered to detect a statistically significant random effect of gender on the slope *b*, these models nonetheless achieved a very close fit to the available data which effectively eliminated dependence on body size as regression lines passed through the origin of the chart. Point estimates thus closely reflected the information contained in the dataset. (ii) Secondly, regression lines of Bayesian models with "partial pooling" did not pass as close to the origin of the chart owing to a design choice: shrinkage and avoidance of overfitting. We believe the adoption of allometric indexing into clinical practice will be facilitated if publications reporting estimates for *b* seek to achieve external validity, and Bayesian analysis as used in the present project is attractive in that respect. Future work seeking to reconcile earlier reports should consider using some form of regularization to achieve estimates with out-of-sample applicability.[4]

Supplementary References

1. Yap J, Lim WK, Sahlén A, et al. Harnessing technology and molecular analysis to understand the development of cardiovascular diseases in Asia: a prospective cohort study (SingHEART). *BMC Cardiovasc Disord*. 2019;19(1):259. doi:10.1186/s12872-019-1248-3

2. Self S, Liang KY. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *J Am Stat Assoc*. 1987;82(398):605-610.

3. Carpenter, B. Bayesian Posteriors are Calibrated by Definition. 2017. https://statmodeling.stat.columbia.edu/2017/04/12/bayesian-posteriors-calibrated/.

4. Sahlén A, Ewe SH, Ding ZP. Meta-Analysis of Optimal Allometric Exponent for Indexing of Echocardiographic Left Ventricular Mass. *J Am Soc Echocardiogr Off Publ Am Soc Echocardiogr*. 2020;33(3):401-403.e1. doi:10.1016/j.echo.2019.10.006

5. Seng MC, Shen X, Wang K, et al. Allometric Relationships for Cardiac Size and Longitudinal Function in Healthy Chinese Adults: Normal Ranges and Clinical Correlates. *Circ J Off J Jpn Circ Soc*. 2018;82(7):1836-1843. doi:10.1253/circj.CJ-18-0134

6. de Simone G, Daniels SR, Devereux RB, et al. Left ventricular mass and body size in normotensive children and adults: assessment of allometric relations and impact of overweight. *J Am Coll Cardiol*. 1992;20(5):1251-1260. doi:10.1016/0735-1097(92)90385-z

7. Lauer MS, Anderson KM, Larson MG, Levy D. A new method for indexing left ventricular mass for differences in body size. *Am J Cardiol*. 1994;74(5):487-491. doi:10.1016/0002-9149(94)90909-1

8. Chirinos JA, Segers P, De Buyzere ML, et al. Left ventricular mass: allometric scaling, normative values, effect of obesity, and prognostic performance. *Hypertens Dallas Tex 1979*. 2010;56(1):91-98. doi:10.1161/HYPERTENSIONAHA.110.150250

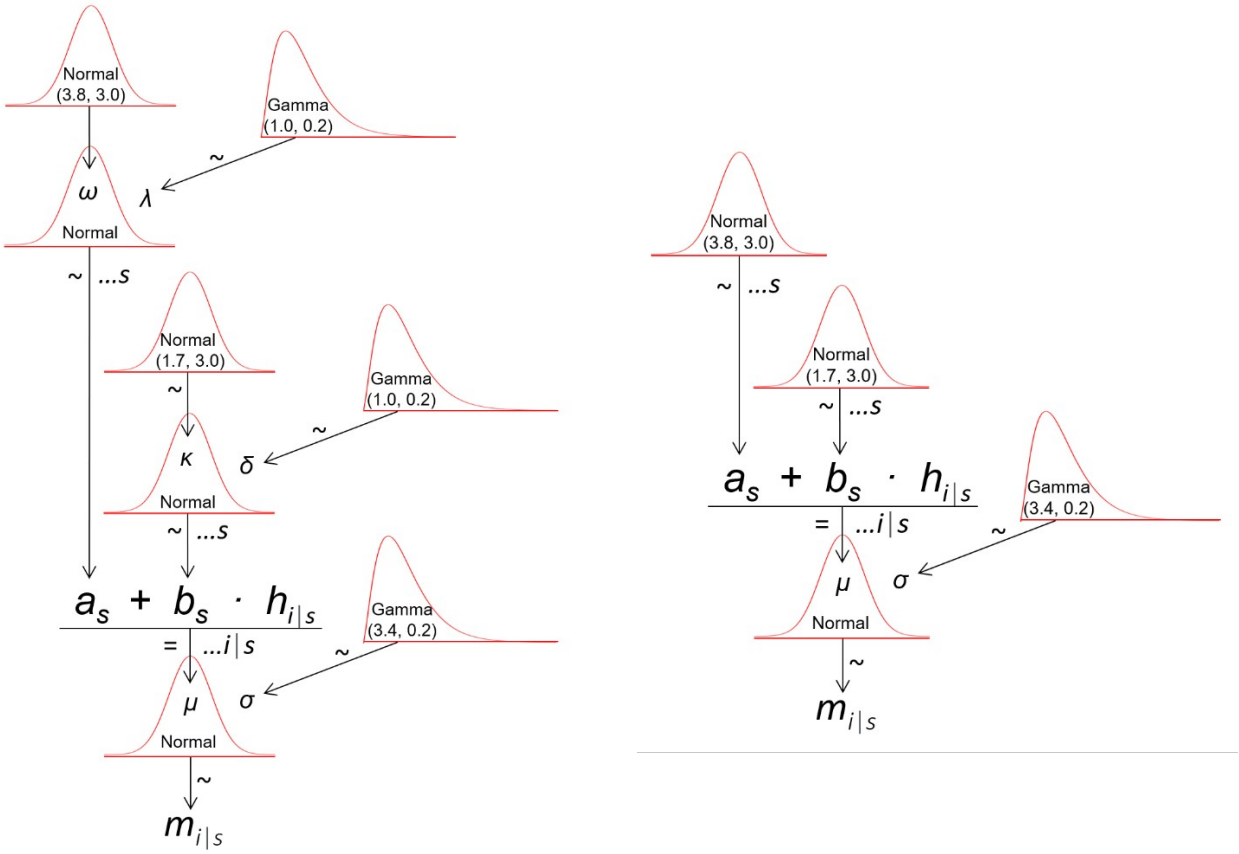
9. Kruschke JK. Model Comparison and Hierarchical Modeling. In: *Doing Bayesian Data Analysis. London, England.* 2nd ed. Elsevier; 2015:265-296.

10. Domingo P. Bayesian Averaging of Classifiers and the Overfitting Problem. In: *Proceedings of the 17th International Conference on Machine Learning*. Morgan Kaufmann; 2000:223-230.

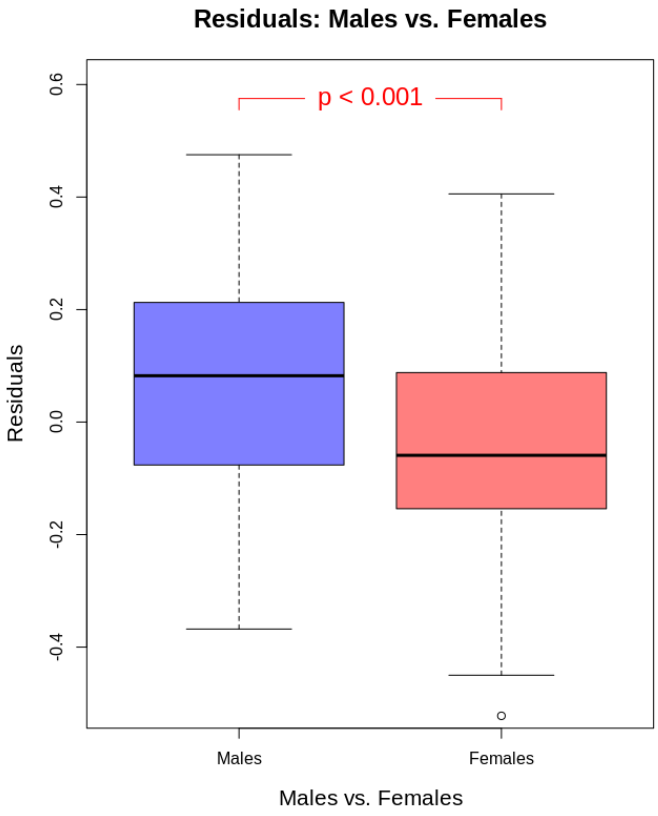
11. Harrell FE. Overview of Maximum Likelihood Estimation. In: *Regression Modeling Strategies*. 2nd ed. Springer Verlag; 2015:181-218.

12. Oman SD. A Different Empirical Bayes Interpretation of Ridge and Stein Estimators. *J R Stat Soc Ser B Methodol*. 1984;46(3):544-557.

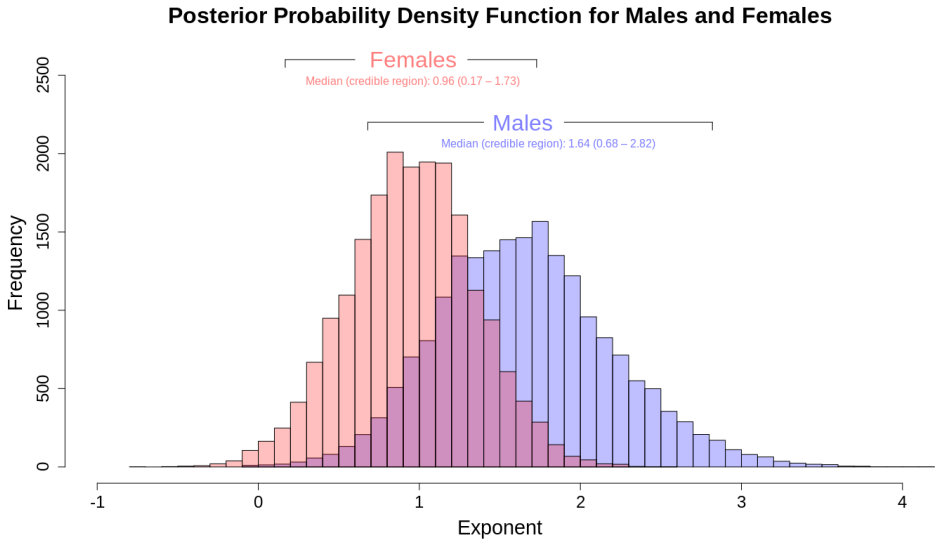
Supplementary Figures



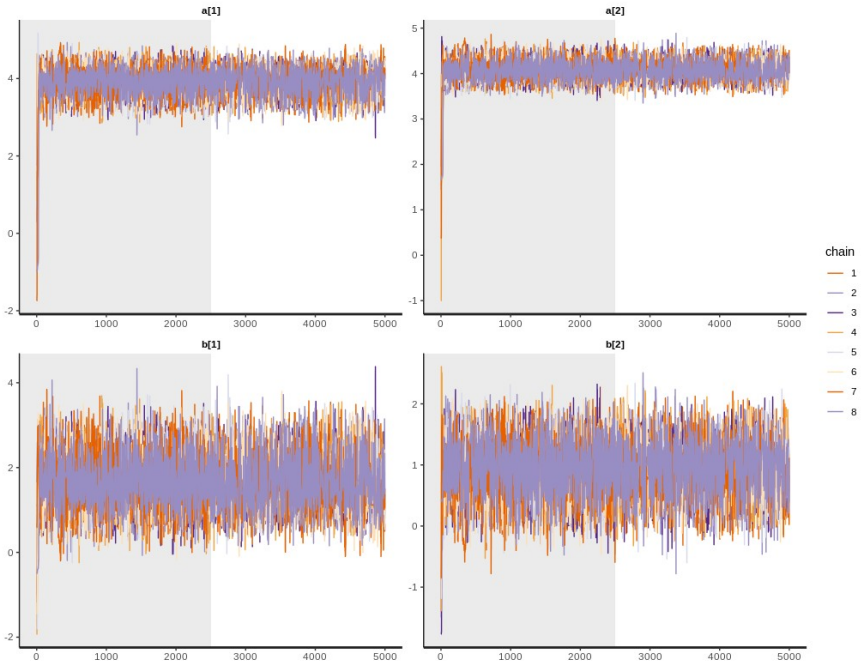
Supplementary Figure 1 Schematic representations of Bayesian models: “partially pooled” (left panel; a) and “zero pooled” (right panel; b).



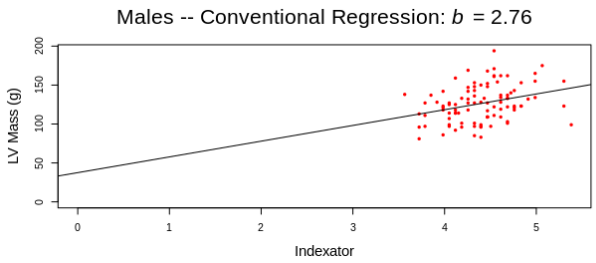
Supplementary Figure 2 Residuals in males (blue) and females (red) showing clustering on sex in the regression model, leading to violation of assumptions of independence of observations.

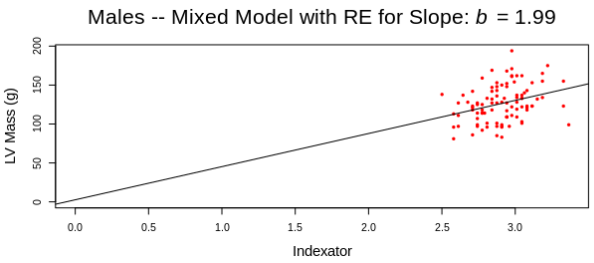


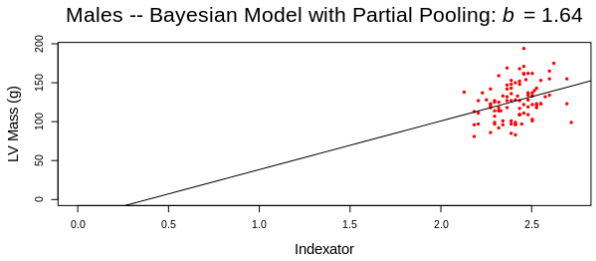
Supplementary Figure 3 Posterior probability densities in the “partially pooled” Bayesian model.

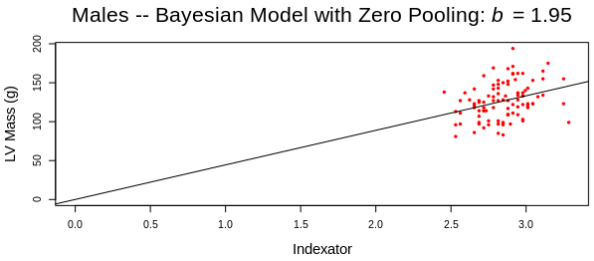


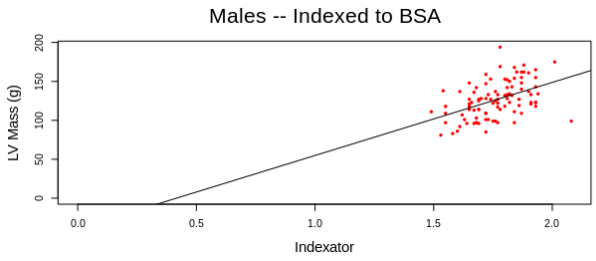
Supplementary Figure 4 Trace plot showing final 8 chains from Hamiltonian Monte Carlo Markov Chains including burn-in period (on the left; shaded in grey). Top panels show the intercept (*a*) in males (left) and females (right), bottom panels show slope (*b*) in males (left) and females (right).

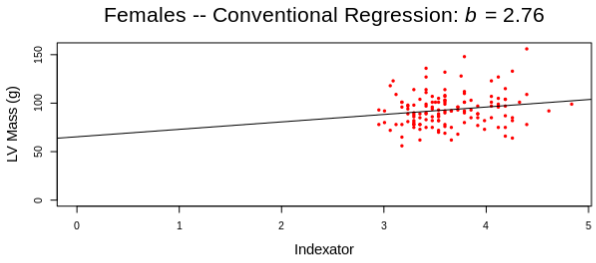

Males -- Conventional Regression: b = 2.76
LV Mass (g)
Indexator

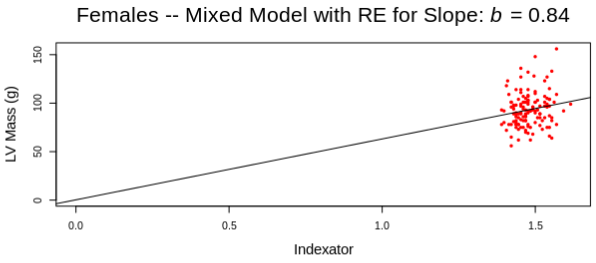

Males -- Mixed Model with RE for Slope: b = 1.99
LV Mass (g)
Indexator

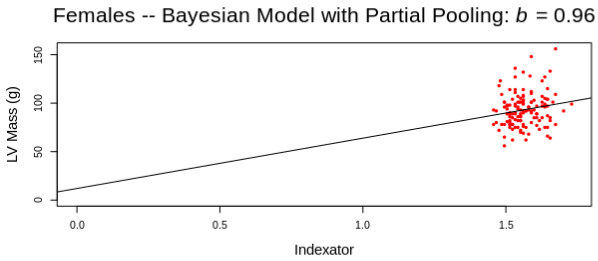

Males -- Bayesian Model with Partial Pooling: b = 1.64
LV Mass (g)
Indexator

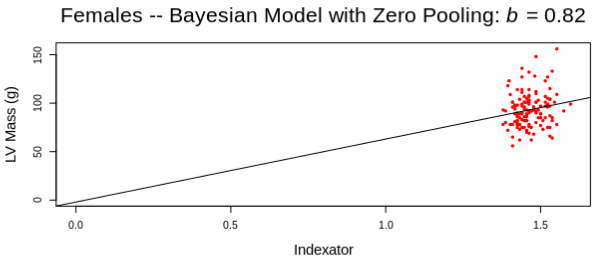

Males -- Bayesian Model with Zero Pooling: b = 1.95
LV Mass (g)
Indexator

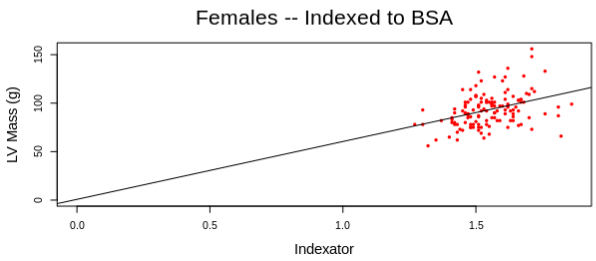

Males -- Indexed to BSA
LV Mass (g)
Indexator


Females -- Conventional Regression: b = 2.76
LV Mass (g)
Indexator

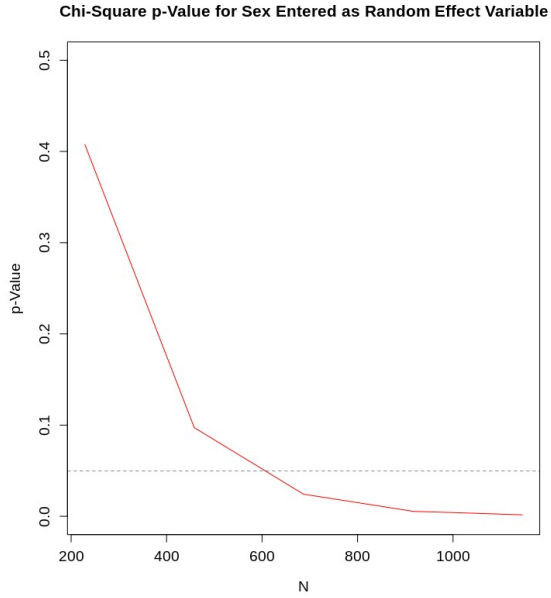

Females -- Mixed Model with RE for Slope: b = 0.84
LV Mass (g)
Indexator


Females -- Bayesian Model with Partial Pooling: b = 0.96
LV Mass (g)
Indexator


Females -- Bayesian Model with Zero Pooling: b = 0.82
LV Mass (g)
Indexator


Females -- Indexed to BSA
LV Mass (g)
Indexator

Supplementary Figure 5 Validation plots for male (upper panel; 6a) and female (lower panel; 6b) subjects showing how well coefficients from models were able to minimize dependence of LVM on body size, judged by how close the regression line passes to the origin of the diagram (0, 0). Of note, while the random effect (RE) term for gender in the conventional mixed model was statistically non-significant, the point estimate identified by the restricted maximum likelihood method nonetheless achieved a close fit to the data very. See Supplementary Discussion for details and interpretation.



Supplementary Figure 6 Exploratory simulation to examine the role of sample size for the p-value of the random effect term for gender in the conventional mixed model. A p-value below the commonly applied threshold value of 0.05 was reached at a population of approximately n = 600 subjects.