Analysis

outcome sensitivity questions
Is this outcome sensitive to choices? How sensitive?

association questions connecting outcomes to parameters/options
What is the source (cause) of the sensitivity? What else can we say about the connection between outcomes and choices?

Category 1: Understand Composition of the Multiverse

data and processes that are input to the multiverse

about analysis calculations

"VoE is estimated by computing the hazard ratio (HR) and P-value for a variable of interest while adjusting for all possible combinations of adjustments from a finite set of adjustment variables. Our algorithm for computing the VoE for a variable x (e.g., serum vitamin D) is shown in Fig. 1."
[patel2015, fig1, ref2]

"Next, we describe the VoE methodology for the association between serum vitamin D and all-cause mortality (Fig. 1B). The total number of combinations of adjusting variables from the set of n=13 total adjustments is 8,192 (or, in general, 2^n models, Fig. 1C). ..."
[patel2015, fig1, ref4]

"Next, we describe the VoE methodology for the association between serum vitamin D and all-cause mortality (Fig. 1B). ... The HR and the respective P-value for the association of that variable with all-cause mortality are estimated for all 8,193 models with different combinations of 13 adjusting variables using Cox proportional hazards time-to-event regression (Fig. 1D). ... "
[patel2015, fig1, ref4]

"We created metrics to express the distributions of VoE for a variable (Fig. 1F). The first was the RHR, the ratio of the 99th percentile and 1st percentile HR. The RHR connotes the spread of HRs for different combinations of adjustments. ..."
[patel2015, fig1, ref5]

"We created metrics to express the distributions of VoE for a variable (Fig. 1F). ... The second was the RP, which is the difference between the 99th and 1st percentile of -log10(P-value). The RP measures the range of P-values over all estimates. ... "
[patel2015, fig1, ref5]

"The 417 variables included 179 serum or urine biomarkers of environmental exposures (e.g., serum cadmium, mercury, or pesticide level), 9 self-reported behavioral factors such as smoking and physical activity, 84 self-reported nutritional intake information (from a food frequency questionnaire), 27 self-reported health conditions (e.g., diabetes), 92 clinical factors (e.g., BMI and cholesterol), and 13 sociodemographic variables (e.g., income). All continuous variables were log transformed and z standardized for comparison. Appendix (at www.jclinepi.com) describes these 417 variables."
[patel2015, fig1, ref7]

"Deriving the multiverse of statistical results. After constructing the data multiverse, the analysis of interest (in this case, an ANOVA or a logistic regression) is performed across all the alternatively constructed data sets. The results are shown in Panels A–F of Figure 1, each showing a histogram of the p values of the Fertility × Relationship interaction effect."
[steegen2016, fig1, ref2]

about data filtering or inclusion criteria

"... We chose variables of interest that had data on at least 1,000 participants and at least 100 death events during follow-up."
[patel2015, fig1, ref3]

"... Mortality information was collected from the date of the survey participation through December 31, 2006, and ascertained via a probabilistic match between NHANES and National Death Index (NDI) death certificate information [21]. ..."
[patel2015, fig1, ref3]

about source of data

"First, we downloaded 417 self-reported, clinical, and molecular measures with linked all-cause mortality information in participants from NHANES 1999-2004 (Fig. 1A). ..."
[patel2015, fig1, ref3]

Category 2: Assess Outcome Sensitivity

outcomes
(output results from analysis universes)

identify insensitive outcomes
(no variation across universes, or less than some threshold)
(Synonyms: "robust")

(quantified insensitivity?)

"The average sampling standard error is 2.4, and the modeling standard error is 2.5—uncertainty about the estimate derives equally from the data and from the model. The combined total (sampling and modeling) standard error is 3.5.6 The robustness ratio—the mean estimate divided by the total standard error—is 4.05. By the standard of a t-test, this would be considered a strongly robust result, which agrees with the 100 percent sign stability and significance rates."
[young2017, fig1, ref4]

"Table 3 shows that there are 1,024 unique combinations of the control variables. Running each of these models and storing all of the estimates, we graph the modeling distribution in Figure 1. The result appears strongly robust. The estimated coefficient on union membership is positive and significant in every possible combination of the control variables: both the sign stability and the significance rate are 100 percent. With this list of possible controls, and using OLS, it is not possible to find an opposite signed or even nonsignificant estimate."
[young2017, fig1, ref3]

"What were the consequences of this variability in analytic approaches? ... Twenty teams (69%) found a significant positive relationship, p < .05, and nine teams (31%) found a nonsignificant relationship. No team reported a significant negative relationship."
[silberzahn2017, fig2, ref2]

"Next, we report the robustness of this finding to the choice of control variables in the model. Does this finding hinge on sets of control variables, or do the findings hold regardless of what assumptions are made over the control variables?"
[young2017, fig1, ref3]

(qualitative insensitivity?)

"In this first pattern, one concludes that although adjustment weakens the magnitude relationship between vitamin D levels and mortality, inferences regarding the relationship are similar throughout all scenarios of adjustment."
[patel2015, fig2, ref2]

"Figure 4. Robustness checks for predictors. Coefficient plot showing a consistent effect of the fertility predictor among naturally cycling women (red) but not hormonal contraception users (black) across several predictor and model specifications (explained in further detail in the text)."
[arslan2018, fig4, ref1]

"The modeling distribution is relatively normal: There are no critically important modeling decisions that generate bimodality in the estimates. As shown in Figure 4, the significant estimates reported in Table 7 above are extreme outliers in the modeling distribution."
[young2017, fig4, ref5]

identify insensitive outcomes

identify sensitive outcomes
(ex: substantial variation between universes)

(quantified sensitivity?)

"The mean estimate from all models is 2.29 and the average sampling standard error is 1.61—indicating that the mean estimate is not statistically significant. In addition, the modeling standard error is 1.60—the estimates vary across models just as much as would be expected from drawing new samples. The total standard error—incorporating both sampling and modeling variance is 2.27, roughly the same size as the estimate itself, yielding a robustness ratio of 1.01."
[young2017, fig2, ref4]

"For religiosity, 7 out of the 120 choice combinations lead to a significant interaction effect, whereas the remaining 94% lead to p values ranging from .05 to 1.0. For fiscal political attitudes, 8% of the 210 choice combinations lead to a significant interaction (p < .05), whereas the remaining choice combinations lead to p values across the entire range from .05 to 1.0."
[steegen2016, fig1, ref3]

"For the remaining four variables, roughly half of the choice combinations lead to a significant interaction effect. ..."
[steegen2016, fig1, ref4]

"The figure shows that the majority of specifications lead to estimates of the sign predicted by the original authors (feminine hurricanes produce more deaths), though a very small minority of all estimates are statistically significant (p<.05). The point estimates range from -1 to +12 additional deaths."
[simonsohn2015, fig2, ref4]

"Across the 256 possible combinations of controls, the effect of gender is typically positive but only 25 percent of the estimates are statistically significant. And 12 percent of the estimates have the opposite sign (though none of those estimates are significant)."
[young2017, fig2, ref3]

"Among the few statistically significant results, the great majority are wrong signed: estimates with negative signs indicate migration toward higher tax states. Only 0.2 percent of estimates are significantly positive compared to 1.3 percent that are significant and wrong signed. ."
[young2017, fig4, ref5]

"For two variables—religiosity in Study 1 (Panel A) and fiscal political attitudes (Panel C)—the multiverse analysis reveals a near-uniform distribution, indicating that the p value for the interaction effect between fertility and relationship varies widely across the multiverse."
[steegen2016, fig1, ref3]

(qualitative sensitivity?)

"The teams also varied in their approaches to handling the nonindependence of players and referees, and this variability also influenced both median estimates of the effect size and the rates of significant results."
[silberzahn2017, fig3, ref3]

"It seems that all arbitrary choices for data processing can have an impact on whether the obtained data set will lead to a significant or a nonsignificant outcome."
[steegen2016, fig2, ref5]

"Next, we report the robustness of this finding to the choice of control variables in the model. Does this finding hinge on sets of control variables, or do the findings hold regardless of what assumptions are made over the control variables?"
[young2017, fig1, ref3]

"Both models are highly sensitive to adding or deleting insignificant controls, and this set of controls is the only combination among many thousands that yields a significant result in both the ACS and IRS data."
[young2017, fig4, ref3]

identify sensitive outcomes

Category 3: Connect Parameters to Outcome Values to Identify Sources of Sensitivity

parameters
(groups of analysis options, ex: exclusion criteria)

identify insensitive parameters (or meta-parameters)
(those that do not cause substantial variation in outcomes)

(meta-parameters)

"Other patterns included VoE where all P-values were >0.05 and the strength of the association decreased (n = 50, 12%), increased (n = 27, 6%), or showed no dependence (n = 15, 4%) with increasing number of adjustment variables k (see Table S3/Appendix at www.jclinepi.com)."
[patel2015, fig2, ref6]

"Other patterns included VoE where all P-values were >0.05 and the strength of the association decreased (n = 50, 12%), increased (n = 27, 6%), or showed no dependence (n = 15, 4%) with increasing number of adjustment variables k (see Table S3/Appendix at www.jclinepi.com)."
[patel2015, fig2, ref6]

(parameters)

"The different exclusion criteria and cycle day estimation options do not seem to have a large impact on fluctuation in the statistical conclusion."
[steegen2016, fig2, ref3]

identify insensitive parameters (or meta-parameters)

identify sensitive parameters (or meta-parameters)
(those that do cause substantial variation in outcomes)

(meta-parameters)

"All the HR estimates are <1.00, indicating that higher levels of vitamin D tend to be associated with longer survival (all HR <0.76); however, the magnitude of the estimated effect is dependent on the number of adjustment variables, and the association is attenuated when adjusting for more variables, from HR = 0.64 with no adjustment (k = 0) to 0.75 with all 13 adjustment variables included (k = 13)."
[patel2015, fig2, ref2]

"Higher thyroxine levels tend to be associated with longer survival, but P-values become greater than the nominal level of statistical significance (P = 0.05) with nine or more adjustment variables on average."
[patel2015, fig2, ref3]

"In the last pattern, as exemplified by a-tocopherol (Fig. 2D), the estimated HRs can be both greater and less than the null value (HR > 1 and HR < 1) depending on what adjustments were made. We call this the Janus effect after the two-headed representation of the ancient Roman god."
[patel2015, fig2, ref5]

(parameters)

"The teams also varied in their approaches to handling the nonindependence of players and referees, and this variability also influenced both median estimates of the effect size and the rates of significant results."
[silberzahn2017, fig3, ref3]

"All other controls push the tax migration estimate toward a zero or wrong-signed result, and virtually must be excluded to support the hypothesis."
[young2017, fig4, ref6]

"In general, deflating the multiverse involves developing a better and more complete theorizing of the constructs of interest and improving their measurement. ... A first approach involves improving the experimental material and design. For example, the detailed multiverse examination shown in Figure 2 revealed that a lot of fluctuation hinged on the different choices for relationship status assessment. Thus, apparently, this type of research could benefit from a better way of assessing relationship status."
[steegen2016, fig2, ref9]

"Why do these estimates vary so much? Why is the distribution so non-normal? What combinations of control variables are critical to finding a positive and significant result?"
[young2017, fig2, ref5]

"It seems that all arbitrary choices for data processing can have an impact on whether the obtained data set will lead to a significant or a nonsignificant outcome."
[steegen2016, fig2, ref5]

identify sensitive parameters (or meta-parameters)

options
(specific analysis options)

identify insensitive options

"For social political attitudes (Panel B), the statistical conclusion is highly robust for the first and second option for relationship status assessment (significant for R1 and nonsignificant for R2)."
[steegen2016, fig2, ref4]

"For religiosity in Study 2 (Panel A), most data sets constructed under the second option for relationship assessment (R2) yield a nonsignificant interaction effect."
[steegen2016, fig2, ref3]

"For religiosity [...] The multiverse analysis revealed that almost all choice combinations for data processing lead to large p values."
[steegen2016, fig2, ref6]

"Figure 4. Robustness checks for predictors. Coefficient plot showing a consistent effect of the fertility predictor among naturally cycling women (red) but not hormonal contraception users (black) across several predictor and model specifications (explained in further detail in the text)."
[arslan2018, fig4, ref1]

"Two control variables clearly stand out as most influential: marital status and race. ... The other controls have much less impact on the estimate and have little model influence."
[young2017, fig3, ref2]

identify insensitive options

identify sensitive options

"In these cases, it is informative to display the multiverse in greater detail by showing which constellation of choices corresponds to which statistical result. This allows to identify the key choices in data processing that are most consequential in the fluctuation of the statistical results."
[steegen2016, fig1, ref4]

"Using the third option for relationship status assessment (R3) leads to more fluctuation, depending on the choices for the other processing steps."
[steegen2016, fig2, ref4]

"Fifteen teams used logistic models, and 11 of these teams found a significant effect (median OR = 1.34; median absolution deviation, or MAD = 0.07). Six teams used Poisson models, and 4 of these teams found a significant effect (median OR = 1.36, MAD = 0.08). Of the 6 teams that used linear models, 3 found a significant effect (median OR = 1.21, MAD = 0.05). The final 2 teams used models classified as miscellaneous, and both of these teams reported significant effects (ORs = 1.71 and 2.88, respectively)."
[silberzahn2017, fig3, ref2]

"For most other variables, there was considerable ambiguity: The interaction seemed to be significant across about half of the arbitrary choice combinations. In these cases, the conclusion on the effect of fertility strongly depends on the evaluation of the different processing options."
[steegen2016, fig2, ref7]

"4.3. Identifying "multimodality of effects" with VoE: By empirically estimating the VoE, it is also possible to detect whether one or more adjustment variables make a marked difference in the results, leading to multiple modes (Fig. 3) which we call multimodality of effects. ... "
[patel2015, fig3, ref2]

"Both models are highly sensitive to adding or deleting insignificant controls, and this set of controls is the only combination among many thousands that yields a significant result in both the ACS and IRS data."
[young2017, fig4, ref3]

"Two control variables clearly stand out as most influential: marital status and race. The influence estimate for marriage shows that, all else equal, when controlling for marital status the coefficient on female increases by 2.47, more than doubling the mean estimate across all models. Controlling for race (with the dummy variable "black") also increases the effect size of gender by 1.91, a full 83 percent higher than the mean estimate."
[young2017, fig3, ref2]

identify sensitive options

Category 4: Connect Combinations of Parameters to Outcome Values to Identify Complex Relationships that lead to Sensitivity

identify association between specific outcomes and specific parameters/options
(ex: which options lead to large effect sizes?)

"The specification "curve" shows the estimated effect size across all specifications, sorted by magnitude, accompanied below by a "dashboard chart" indicating the operationalizations behind each result (see e.g., Figure 2)."
[simonsohn2015, fig2, ref2]

"The bottom panel of the figure tells us which analytic decisions produce different estimates. For example, we can see that obtaining a negative point estimate requires a fairly idiosyncratic combination of operationalizations ... "
[simonsohn2015, fig2, ref5]

"In these cases, it is informative to display the multiverse in greater detail by showing which constellation of choices corresponds to which statistical result."
[steegen2016, fig1, ref4]

"Using the third option for relationship status assessment (R3) leads to more fluctuation, depending on the choices for the other processing steps."
[steegen2016, fig2, ref4]

"Teams that employed logistic or Poisson models tended to report estimates that were larger than those of teams that used linear models (see the effect sizes in Fig. 3, in which the teams are clustered according to the distribution used for analyses)."
[silberzahn2017, fig3, ref2]

"We observed three modes in the association between triglyceride levels and mortality (Fig. 3D-F). The multimodal plots indicated that total cholesterol and diabetes were driving these modes."
[patel2015, fig3, ref4]

"In essence, there are two distinct modeling distributions to consider which are plotted in Figure 3. In one set of models, the controls for race and marital status are always excluded but all other controls are allowed in the model space (which gives 128 models). [...] In order to draw robust conclusions from these data, one must make a substantive judgment about two key modeling assumptions: the inclusion of race and marital status. None of the other model ingredients affect the basic conclusion. These two model assumptions determine the results."
[young2017, fig3, ref3]

"Both models are highly sensitive to adding or deleting insignificant controls, and this set of controls is the only combination among many thousands that yields a significant result in both the ACS and IRS data."
[young2017, fig4, ref3]

"We found that including adjustments for menstruation and pre-menstruation (M_c3) reduced effect sizes for the fertile window predictor."
[arslan2018, fig4, ref2]

identify association between specific outcomes and specific parameters/options

identify combinations of parameters/options associated with specific outcomes
(sensitivity conditioned upon some portion of parameter space, interactions between options)

"For example, we can see that obtaining a negative point estimate requires a fairly idiosyncratic combination of operationalizations: (i) not taking into account the year of the storm, (ii) operationalizing severity of the storm by the log of damages, (iii) conducting an OLS regression, etc."
[simonsohn2015, fig2, ref5]

"Using the third option for relationship status assessment (R3) leads to more fluctuation, depending on the choices for the other processing steps."
[steegen2016, fig2, ref4]

"For example, we can see that obtaining a negative point estimate requires a fairly idiosyncratic combination of operationalizations: (i) not taking into account the year of the storm, (ii) operationalizing severity of the storm by the log of damages, (iii) conducting an OLS regression, etc."
[simonsohn2015, fig2, ref5]

"The first and third options (R1 and R3) consistently lead to a significant interaction effect in combination with the first and second option for fertility assessment (F1 and F2) and to a nonsignificant interaction effect in combination with F5, whereas data sets constructed under R1 or R3 in combination with F3 or F4 lead to more fluctuating conclusions, depending on the other choices for data processing."
[steegen2016, fig2, ref3]

"For k = 5-13 number of adjustment variables, the association tends to become stronger in HR and statistical significance; however, the trend is less clear for k = 0-4, where the median P-values increase."
[patel2015, fig2, ref4]

"We also assessed whether associations appeared on both sides of the null (HR <1 and HR >1): depending on what adjustments are chosen, the results may suggest that the variable of interest is associated with either increased or decreased mortality."
[patel2015, fig1, ref5]

"In the last pattern, as exemplified by a-tocopherol (Fig. 2D), the estimated HRs can be both greater and less than the null value (HR > 1 and HR < 1) depending on what adjustments were made. We call this the Janus effect after the two-headed representation of the ancient Roman god."
[patel2015, fig2, ref5]

"Why do these estimates vary so much? Why is the distribution so non-normal? What combinations of control variables are critical to finding a positive and significant result?"
[young2017, fig2, ref5]

identify combinations of parameters/options associated with specific outcomes

identify no definite pattern of association between parameters/options and outcomes
(unclear, inconsistent)

"Finally, for voting and donation preferences (Panels C and D, respectively), it is hard to extract a consistent pattern of fluctuation across the different choice combinations."
[steegen2016, fig2, ref5]

"It seems that all arbitrary choices for data processing can have an impact on whether the obtained data set will lead to a significant or a nonsignificant outcome."
[steegen2016, fig2, ref5]

"Rarer patterns included variables where all P-values were <0.05 and there was an increasing strength of association (n = 5, 1%) or no clear relationship with increasing k (n = 4, 1%), and those having P-values with a range less than and greater than 0.05 with no clear relationship with k (n = 15, 4%)."
[patel2015, fig2, ref6]

"We found that including adjustments for menstruation and pre-menstruation (M_c3) reduced effect sizes for the fertile window predictor. We could not always adjust for (pre-)menstruation when using a narrow window predictor because of model convergence problems. After taking this into account, we found no systematic pattern in which certain predictors (narrow or broad window, forward or backward counted) had larger effect sizes than others across outcomes (see Figure 4)."
[arslan2018, fig4, ref2]

identify no definite pattern of association between parameters/options and outcomes

identify quantitative relationship/association between parameters/options and outcomes
(correlation, step-wise function, others?)

"The first pattern is exemplified by the association between serum levels of vitamin D and mortality (Fig. 2A). All the HR estimates are <1.00, indicating that higher levels of vitamin D tend to be associated with longer survival (all HR <0.76); however, the magnitude of the estimated effect is dependent on the number of adjustment variables, and the association is attenuated when adjusting for more variables, from HR = 0.64 with no adjustment (k = 0) to 0.75 with all 13 adjustment variables included (k = 13)."
[patel2015, fig2, ref2]

"The second pattern is exemplified by the relationship between thyroxine and mortality, displays how increasing adjustment might change inference (Fig. 2B). Higher thyroxine levels tend to be associated with longer survival, but P-values become greater than the nominal level of statistical significance (P = 0.05) with nine or more adjustment variables on average."
[patel2015, fig2, ref3]

"The third pattern, as exemplified by an indicator of kidney function, urinary creatinine, and mortality, shows an opposite trend (Fig. 2C). For k = 5-13 number of adjustment variables, the association tends to become stronger in HR and statistical significance; however, the trend is less clear for k = 0-4, where the median P-values increase. Twenty-six (6%) of the 417 variables exhibited similar behavior to urinary creatinine where the effect sizes increased and P-values decreased for larger k."
[patel2015, fig2, ref4]

"Rarer patterns included variables where all P-values were <0.05 and there was an increasing strength of association (n = 5, 1%) or no clear relationship with increasing k (n = 4, 1%), and those having P-values with a range less than and greater than 0.05 with no clear relationship with k (n = 15, 4%)."
[patel2015, fig2, ref6]

"We also visualized trends corresponding to the number of adjusting variables (k), plotting the median effect size and P-value for each k from 0 to 13. We recorded the proportion of estimates that achieved different levels of nominal statistical significance (P < 0.05, 0.0001)."
[patel2015, fig1, ref6]

identify quantitative relationship/association between parameters/options and outcomes

identify distinctive patterns of association between parameters/options and outcomes
(would this be non-quantitative? is it clear if a pattern is one or the other?)

"4.2. Prototypical patterns of the VoE: We describe four prototypical patterns from the set of 417 variables (Fig. 2, see Appendix at www.jclinepi.com for all 417 variables)."
[patel2015, fig2, ref2]

"In the last pattern, as exemplified by a-tocopherol (Fig. 2D), the estimated HRs can be both greater and less than the null value (HR > 1 and HR < 1) depending on what adjustments were made. We call this the Janus effect after the two-headed representation of the ancient Roman god. For a-tocopherol, most of the HR and P-values were concentrated around 1 and nonsignificance, respectively. However, 1% of the models had an HR < 0.875 (12.5% decreased risk of death for 1SD increase in exposure) with a nominally significant P-value (P < 0.05), whereas another 1% of the models had HR > 1.05 (5% increased risk for death for 1SD increase of exposure), albeit without reaching nominal significance. The Janus effect is common: 131 (31%) of the 417 variables had their 99th percentile HR > 1 and their 1st percentile HR < 1."
[patel2015, fig2, ref5]

"The multiverse analysis revealed that almost all choice combinations for data processing lead to large p values. As such nonsignificant findings in general represent nothing more than uncertainty, this pattern of results clearly raises serious questions regarding the finding on the effect of fertility found in the single data set analysis, and should make a researcher hesitant to trust the single data set finding."
[steegen2016, fig2, ref6]

"We observed three modes in the association between triglyceride levels and mortality (Fig. 3D-F). The multimodal plots indicated that total cholesterol and diabetes were driving these modes. For example, in models that did not contain these two adjustments, the associations had smaller P-values and a smaller range of HR. Furthermore, in models containing diabetes, HR were attenuated."
[patel2015, fig3, ref4]

"Why do these estimates vary so much? Why is the distribution so non-normal? What combinations of control variables are critical to finding a positive and significant result?"
[young2017, fig2, ref5]

identify distinctive patterns of association between parameters/options and outcomes

Category 5: Validate the Multiverse

compare reliability metrics between sets of universes

To further investigate model quality, Emma drills down to individual universes by clicking a dot in the outcome view. She sees in the model fit view (Fig. 1e) that a model gives largely mismatched predictions.
[liu2020, fig1, ref2]

Clicking a result in the outcome view populates the model fit view with visual predictive checks, which show how well predictions from a given model replicate the empirical distribution of observed data [14], allowing users to further assess model quality (T5). ...
[liu2020, fig1, ref3]

... Users may further drill down to assess the fit quality of individual models (e) by comparing observed data (pink) with model predictions (teal).
[liu2020, fig1, ref1]

... The model fit visualization juxtaposes violin plots of the observed data and model predictions to facilitate comparison of the two distributions (see Fig. 1e). Within the violin plots, we overlay observed and predicted data points as centered density dot plots to help reveal discrepancies in approximation due to kernel density estimation. ...
[liu2020, fig1, ref3]

Now that Emma understands what decisions lead to null effects, she wonders if these results are from trustworthy models. She changes the color-by field to get an overview of model fit quality (Fig. 8a) and sees that the universes around zero have a poorer fit. She then uses a slider to remove universes that fail to meet a quality threshold (Fig. 8b).
[liu2020, fig8, ref2]

Boba enables an overview of model fit quality across all universes (T5) by coloring the outcome view with a model quality metric (Fig. 8a). We use normalized root mean squared error (NRMSE) to measure model quality and map NRMSE to a single-hue colormap of blue shades where a darker blue indicates a better fit.
[liu2020, fig8, ref3]

But do we have evidence that certain outcomes are less trustworthy? We toggle the color-by drop-down menu so that each universe is colored by its model quality metric (Fig. 11b). The large estimates are almost exclusively coming from models with a poor fit. We further verify the model fit quality by picking example universes and examining the model fit view (Fig. 11c). The visual predictive checks confirm issues in model fit, for example the models fail to generate predictions smaller than 3 deaths, while the observed data contains plenty such cases.
[liu2020, fig11, ref2]

T5: Model Fit – assess the model fit quality of individual universes to distinguish trustworthy models from questionable ones.
[liu2020, tasks, T5]

... What does make a difference is whether fatigue is treated as a collider and excluded as a covariate (Model 1), or treated as a mediator and controlled for in the analysis (Model 2). Making an informed decision between these models would require additional empirical evidence (e.g., experimental or quasiexperimental studies), theoretical developments, or both.
[delgiudice2020, fig6, ref3]

compare multiverse to null

A more robust inference utilizes the null distribution – the expected distribution of outcomes when the null hypothesis of no effect is true. In this case, the inference view shows an aggregate plot followed by a detailed plot (Fig. 9ab). The aggregate plot (Fig. 9a) compares the null distribution (red) to possible outcomes of the actual multiverse (blue) across sampling and decision variations. The detailed plot (Fig. 9b) shows point estimates (colored dots) against 95% confidence intervals representing null distributions (gray lines) for each universe. Each point estimate is orange if it is outside the range, or blue otherwise. ...
[liu2020, fig9, ref4]

Now that we have reasons to be skeptical of the large estimates, the remaining universes still support a small, positive effect. How reliable is the effect? We proceed to the inference view to compare the possible outcomes in the observed multiverse and the expected distribution under the null hypothesis (Fig. 11d). The two distributions are different in terms of mode and shape, yet they are highly overlapping, which suggests the effect is not reliable. The detail plot depicting individual universes (supplemental Fig. 1) further confirms this observation. ...
[liu2020, fig11, ref3]

aggregate uncertainty across universes

Fig. 5. Decision view and outcome view. (a) The decision view shows analytic decisions as a graph with order and dependencies between them, and highlights more sensitive decisions in darker colors. (b) The outcome view visualizes outputs from all analyses, including individual point estimates and aggregated uncertainty.
[liu2020, fig5, ref1]

In addition, Boba enables users to propagate concerns in model fit quality to the inference view in two possible ways. The first way employs a model averaging technique called stacking [58] to take a weighted combination of the universes according to their model fit quality. ... As an alternative, Boba excludes the universes below the model quality cutoff users provide in Sect. 5.4. ...
[liu2020, fig9, ref5]

Finally, the uncertainty and model fit visualizations in Boba surface potential issues that previous work might have overlooked. First, though the point estimates in Fig. 10b fall into three distinct clusters, the aggregated uncertainty distribution appears unimodal despite a wider spread. The PDF plot (Fig. 10e) shows that sampling distribution from one analysis typically spans the range of multiple peaks, thus explaining why the aggregated uncertainty is unimodal. These observations suggest that the multimodal patterns exhibited by point estimates are not robust when we take sampling variations into account. ...
[liu2020, fig10, ref4]

T4: Uncertainty – assess the end-to-end uncertainty as well as uncertainty associated with individual universes.
[liu2020, tasks, T4]

Inspect individual universes

Besides aggregated uncertainty, Boba allows users to examine uncertainty from individual universes (Fig. 7). ...
[liu2020, fig7, ref2]

T4: Uncertainty – assess the end-to-end uncertainty as well as uncertainty associated with individual universes.
[liu2020, tasks, T4]

Multiverse Interpretation support?

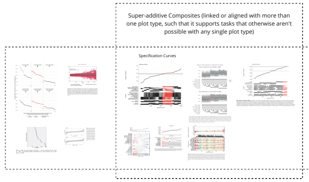
... (b). To interpret the results from all analyses, users start with a graph of analytic decisions (c), where sensitive decisions are highlighted in darker blues. Clicking a decision node allows users to compare point estimates (d, blue dots) and uncertainty distributions (d, gray area) between different alternatives. Users may further drill down to assess the fit quality of individual models (e) by comparing observed data (pink) with model predictions (teal).
[liu2020, fig1, ref1]

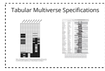
After an in-depth exploration, Emma proceeds to the final step, asking "given the multiverse, how reliable is the effect?" ...
[liu2020, fig9, ref2]

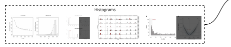
To support users in making inference and judging how reliable the hypothesized effect is (T6), Boba provides an inference view at the end of the analysis workflow, after users have engaged in exploration. Once in the inference view, all earlier views and interactions are inaccessible to avoid multiple comparison problems [60] arising from repeated inference. ...
[liu2020, fig9, ref3]

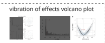
T6: Inference – perform statistical inference to judge how reliable the hypothesized effect is, while accounting for model quality.
[liu2020, tasks, T6]

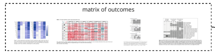
Analysis

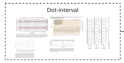


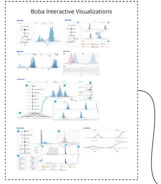

Tabular Multiverse Specifications

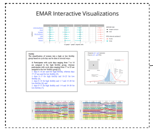












Category 1 only Coding Legend

Detail of information provided (assuming you know how to use the plot):
0: not supported, no information
1: some information provided, but partial, implicit, or only as lookup codes
2: at least basic information provided on some options or parameters, but not all
3: at least basic information provided on all included options and parameters

Coding Legend (all other Categories)

Ease of visual task (assuming you know how to use the plot):
0: not supported
1: information is there (technically possible), but requires mental calculations and is time consuming/difficult
2: some sub-tasks are easy, some difficult - it depends
3: some clear visual features or direct encoding of the information makes the task fast and easy

Category 1: Understand Composition of the Multiverse → **Category 2: Assess Outcome Sensitivity** → **Category 3: Connect Parameters to Outcome Values to Identify Sources of Sensitivity** → **Category 4: Connect Combinations of Parameters to Outcome Values to Identify Complex Relationships that lead to Sensitivity** → **Category 5: Validate the Multiverse** → **Category 6: Interpret the Multiverse [Direction for future work]**

identify sensitive outcomes
(ex: substantial variation between universes)

identify sensitive parameters/meta-parameters/options
(those that cause substantial variation in outcomes)

identify association between specific outcomes and specific parameters/options
(ex: which options lead to large effect sizes?)

These tasks are are about understanding what a multiverse (or visualization) is made of

These tasks are across all universes, without subsetting

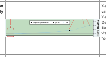
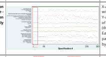
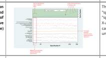
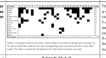
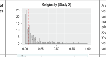
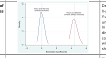
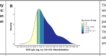
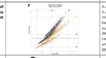
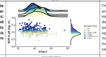
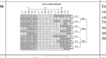
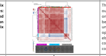
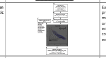
These tasks compare across specific subsets of universes (by simple criteria of parameter/option or outcome values)

These tasks compare across specific subsets of universes (by more complex criteria of parameter/option combinations, or in conjunction with outcome values)

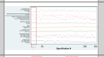
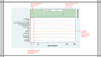
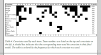
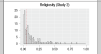
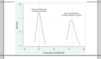
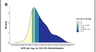
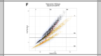
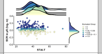
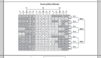
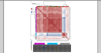
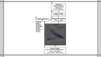
These tasks examine the validity of individual universes or sets of universes (defined by parameters, options, etc.)

Tasks to interpret the multiverse are not well defined, understood, or supported visually, so we stop here as it becomes a matter of our own opinions rather than things for which we have specific examples in published literature as applied to visualizations.

Archetype Coding Grid

Archetype Name	Archetype Representative Image	Short description of what the visualization shows, in the language of this framework	Notess about the visualization that are not task-specific	Domain specific	Variables of Interest	Uncertainty Measures	Universe Scalability	Parameter and Option Scalability
							(1s, 10s, 100s, 100s)	(1s, 10s, 100s, 100s)
specification curve - curve only		X-axis: universes ordered by outcome measure value. Y-axis: The outcome measure, which here is Extra Deaths predicted by the model within each universe. Each blue dot represents 1 universe. This type of visualization is sometimes linked with the "descriptive panel" type, but is often shown alone.	This is the most commonly reported basis for the specification curve (at least as we have encountered them), and notably contains no connection to parameters/options when given by itself.		1	stat. sig.	100s * (100s visible, random selection from Unlimited universes)	0
specification curve - specification panel only		X-axis: universes ordered by outcome measure, but with the outcome measure value not shown here. Y-axis: Parameters (groups of options) share a color of dot, while each row represents a single option (dotted if present in this universe, blank otherwise). Each column is one universe, with one dot per parameter to indicate how that universe is defined by analytical options.	Shows parameters and the options that universes are made of, along with a short text description of each option. Without the upper panel one could know that the specifications are ranked by effect size, but otherwise this visualization is coded as if there was no upper panel shown.		1	NA	100s * (100s visible, random selection from Unlimited universes)	10s
specification curve - combined (Example of Super-additive Composite)		This is a linked composite of two types: the "specification curve - curve only" and the "specification curve - specification panel only". The X-axis is linked so that outcome values of a universe can be tied to the options that makeup that universe.	Is composed of a discrete upper and lower panel, which are aligned and work together to define the full version of this visualization.		1	stat. sig.	100s * (100s visible, random selection from Unlimited universes)	10s
Tabular Multiverse Specification		This tabular visualization shows the specifications in a sparse multiverse - one where not every combination of parameters and options is analyzed. Y-axis: "Covariate" items are binary parameters, with each row being a different parameter, and the final row of "N Covariates" is a meta-parameter. X-axis: Each numbered column is a universe, while the final column "% used" is a meta-parameter. No outcomes are encoded in this visualization.			1	0	10s	10s
histogram of outcomes		A conventional histogram that shows the outcome values of a single outcome measure across universes in a given multiverse. In this case the name of the outcome measure (Religiosity) is in the plot title. X-axis: the outcome measure value, binned. Y-axis: the count or proportion of that measure value compared across all universes, with each universe counting once.	Original used faceting across variables of interest		1	p-value	Unlimited	0
density plot of outcomes		Density plot of outcome values of universes. X-axis: outcome value. Y-axis: density, which is relative proportion of universes with a given outcome value. In this case color is used to compare two distributions of universes, defined by the combination of two analytical options. Universes with "Race" and "Married" included in the model are shown in red, and all other universes in blue.	Some versions of this density plot showed only the distribution overall, so parameters/options were not shown at all.		1	0	Unlimited	1s
Varietal of density plot of outcomes: multiple option shaded variation		This density plot encodes individual observations within the full dataset that is used to construct the multiverse. The "Exclusion Group" is an analytical parameter, with each associated option being represented as colored portions of the distribution. This plot thus connections analytical options to the observations in the dataset they effect, yet does not encode any universe outcomes.			1	NA	Unlimited	1s
Archetype Name	Archetype Representative Image	Short description of what the visualization shows, in the language of this framework	Notess about the visualization that are not task-specific	Domain specific	Variables of Interest	Uncertainty Measures	Universe Scalability	Parameter and Option Scalability
vibration of effects volcano plot		X-axis: outcome value of effect size (hazard ratio). Y-axis: statistical significance (in this case using inverse scaled p-values, which is negative log scale). Each dot is a universe, thus encoding two types of outcome value at once. Saturation is used to show density of values (to resolve over-plotting). In this case color is also used to distinguish two sets of universes defined by a single parameter/option (gold and black for with and without that option),	There are many variations on a theme of this plot within the original paper, including some that show no parameters/options. This plot type notably encodes effect size and statistical significance simultaneously, as well as a meta-parameter (red line).		1	p-value	Unlimited	1s
Composite Varietal of scatter plot of outcomes: parameter shaded density plots along axis margins		The base component of this figure is a scatter plot encoding two different outcome values, but with each dot representing a single participant in the full dataset. Color is used to encode an analytical parameter ("Exclusion Group") and its associated options. Additionally, each outcome value axis has its own associated density plot added along the axis margin, which is also colored to encode the the same analytical options.	Composite of two plot types: scatter plot and density plot		2	0	Unlimited	1s
Outcome Matrix		Each cell is a single universe, with the text of each cell being the value of an outcome measure. Cells can be colored or shaded by outcome value. The X and Y axes are analytical options, with each level of the tree being a a parameter. Ex: The upper-left-most gray cell with a zero inside is a universe defined by options F1, R1, EC1, ECL1, and NM01.	Original used faceting across variables of interest		1	p-value	100s	10s
Varietal of matrix of outcomes: clustered outcome-option correlation matrix		The base component of this plot is a matrix of outcome values, with the value encoded as color, only here the correlation between outcome values between universes is being shown. Each row is a universe with an assigned ID, with the analytical options that define each ID needing to be looked-up elsewhere. As this is a correlation matrix, the columns are a repeat of the rows. The tree axes connect universes according to cluster similarity.	Complex figure, and though the lower-panel is domain-specific (just a color-coded variety of brain spatial map), the upper panel is more broadly relevant as it shows multiple clustering techniques.	?	1	stat. sig	Unlimited	10s
computation schematic		Each letter-labeled box is a step of the analytical process or important aspect of the definition of the multiverse(s) overall, with arrows showing the the order of operations. This visualization type does not encode outcome values itself, though can be combined with any other visualization type by embedding it as one of the insets.	This visualization is not about encoding outcomes, but about explaining the process or composition of the multiverse. * Panel E is treated as an empty box for coding of this visualization.		Unlimited	NA	Unlimited	10s
Archetype Name	Archetype Representative Image	Short description of what the visualization shows, in the language of this framework	Notess about the visualization that are not task-specific	Domain specific	Variables of Interest	Uncertainty Measures	Universe Scalability	Parameter and Option Scalability

Archetype Name	**Composition: Process** understand the process that defines and creates the universes being considered	**Composition: Parameters** understand the definition and composition of universe parameters and parameter values	**Outcome: Range** assess range or spread of outcome values across all universes	**Outcome: Frequency** assess overall frequency of outcome values across all universes	**Connect: Outcome Range** Connect parameters to outcomes by comparing similarity or range of outcome values across a subset of universes defined by a specific parameter value	**Connect: Outcome Frequency** Connect parameters to outcomes by comparing frequency of outcome values across a subset of universes defined by a specific parameter value	**Connect: Specific Outcomes** Connect parameters to outcomes by examining specific outcome values of interest and identifying parameter values that lead to those outcomes
	(what gets included in the multiverse overall, or this specific visualization?)	(what are the parameters, with some information on their relationship to each other or their definition)	(does not require frequency measure)	(requires outcome value and frequency information to check for modality, normality, uniformity, special shapes, etc.)	(what would the multiverse look like if no outliers were excluded, as opposed to if all were?)		e.g. what parameters have options that lead to negative effect sizes?
specification curve - curve only	Detail: 0 Not supported	Detail: 0 Not supported	Ease: 3 The full range of outcome values is encoded explicitly, along with displaying all occurring outcome values (or a random sample of the mid-range, for larger multiverses).	Ease: 2 Frequency is supported at this universe scale by using sampling along the center of the distribution, and precision of frequency estimation is thus made more difficult.	Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported
specification curve - specification panel only	Detail: 0 Not supported	Detail: 2 All parameters and associated options are explicitly encoded on the y-axis, with the full specification of each universe shown (one universe per column). However, as there is only space for a few words to describe each element, the majority of the description must be provided elsewhere.	Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported
specification curve - combined (Example of Super-additive Composite)	Detail: 0 Not supported	Detail: 2 All parameters and associated options are explicitly encoded on the y-axis, with the full specification of each universe shown (one universe per column). However, as there is only space for a few words to describe each element, the majority of the description must be provided elsewhere.	Ease: 3 The full range of outcome values is encoded explicitly, along with displaying all occurring outcome values (or a random sample of the mid-range, for larger multiverses).	Ease: 2 Frequency is supported at this universe scale by using sampling along the center of the distribution, and precision of frequency estimation is thus made more difficult.	Ease: 3 As the upper panel is sorted by outcome value and aligned with the analytical options in the lower panel, this task becomes straightforward.	Ease: 2 It is relatively easy to estimate frequency of outcomes for some analytical parameters, such as "Model" (orange dots), as the dots cluster together in terms of their connected outcome values. By contrast, it is far more challenging to do with parameters with more options and less consistent patterns, as with "Name femininity" (pink dots).	Ease: 3 This task is illustrated by the original authors with the red box on the left side of the figure. This task is made more difficult with larger multiverses, such that specifications are packed closer together and harder to visually differentiate. However, this could be trivially addressed with visual guides such as grid lines or alternating shading, so we do not deduct from the score for this minor issue.
Tabular Multiverse Specification	Detail: 0 Not supported	Detail: 1 Only a single word, code, or very short description, can be given for each parameter or option.	Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0	Ease: 0	Ease: 0
histogram of outcomes	Detail: 0 Not supported	Detail: 0 Not supported	Ease: 3 The histogram explicitly shows the full distribution of outcome values and their frequencies.	Ease: 3 The histogram explicitly shows the full distribution of outcome values and their frequencies.	Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported
density plot of outcomes	Detail: 0 Not supported	Detail: 1 A very-short, partial description of some select parameters and options can be encoded as text or color-coded, but there is space for little else.	Ease: 3 The full distribution of outcome values and their relative frequencies are directly encoded in the density plot, just as with the histogram.	Ease: 3 The full distribution of outcome values and their relative frequencies are directly encoded in the density plot, just as with the histogram.	Ease: 2* This task is easy if the specific parameter/option is given its own density line, but otherwise the task would not be possible.	Ease: 2* This task is easy if the specific parameter/option is given its own density line, but otherwise the task would not be possible.	Ease: 1 Minimally supported, as only the parameters/options that define the density lines can be connected to their associated outcomes.
Varietal of density plot of outcomes: multiple option shaded variation	Detail: 0 Not supported	Detail: 1 Only supports a very short description or codes of a single parameter and its associated options.	Ease: 3* The full distribution of outcome values and their relative frequencies are directly encoded in the density plot, just as with the histogram. Note that this example's axis is not an outcome value, but this any continuous outcome could be visualized this way.	Ease: 3* The full distribution of outcome values and their relative frequencies are directly encoded in the density plot, just as with the histogram. Note that this example's axis is not an outcome value, but any continuous outcome could be visualized this way.	Ease: 2* This task is easy if the specific parameter/option is given its own density line or color-coding, but otherwise the task would not be possible. Note that this example's axis is not an outcome value, but any continuous outcome could be visualized this way.	Ease: 2* This task is easy if the specific parameter/option is given its own density line or color-coding, but otherwise the task would not be possible. Note that this example's axis is not an outcome value, but any continuous outcome could be visualized this way.	Ease: 1 Minimally supported, as only the parameters/options that define the density lines of the color coding can be connected to their associated outcomes. Note that this example's axis is not an outcome value, but any continuous outcome could be visualized this way.
Archetype Name	**Composition: Process** understand the process that defines and creates the universes being considered	**Composition: Parameters** understand the definition and composition of universe parameters and parameter values	**Outcome: Range** assess range or spread of outcome values across all universes	**Outcome: Frequency** assess overall frequency of outcome values across all universes	**Connect: Outcome Range** Connect parameters to outcomes by comparing similarity or range of outcome values across a subset of universes defined by a specific parameter value	**Connect: Outcome Frequency** Connect parameters to outcomes by comparing frequency of outcome values across a subset of universes defined by a specific parameter value	**Connect: Specific Outcomes** Connect parameters to outcomes by examining specific outcome values of interest and identifying parameter values that lead to those outcomes
vibration of effects volcano plot	Detail: 0 Not supported	Detail: 0 Not supported	Ease: 3 As a type of scatter plot, it is easy to see the full range of outcome values that occur as it is directly encoded along the x-axis.	Ease: 2 When multiverses are dense over-plotting occurs, as with this example, and so color or transparency is used to allow a rough estimate of frequency to be determined. When effect sizes are spread out also along the y-axis, it is hard to estimate the relative frequencies of some outcomes compared to others.	Ease: 2* This task is easy if the specific parameter/option is given a color coding, as otherwise the task is not possible with this visualization.	Ease: 2* This task is sometimes easy if the specific parameter/option is given a color coding, as otherwise the task is not possible with this visualization. Even with color coding, over-plotting and outcomes spread across the y-axis make this task considerably harder to complete for some multiverses.	Ease: 1 Minimally supported as only a meta-parameter (red line and dots) and single color-coded parameter can be connected to outcomes. For other (non-shown) parameters, the task is impossible.
Composite Varietal of scatter plot of outcomes: parameter shaded density plots along axis margins	Detail: 0 Not supported	Detail: 1 Only supports a very short description or codes of a single parameter and its associated options.	Ease: 3* The full distribution of outcome values and their relative frequencies are directly encoded in the density plot, just as with the histogram.	Ease: 3* The full distribution of outcome values and their relative frequencies are directly encoded in the density plot, just as with the histogram. Note that this example's axes are not outcome values, but any continuous outcomes could be visualized this way.	Ease: 3* This task is easy if the specific parameter/option is given its own color-coding or axis, but otherwise the task would not be possible. Note that this example's axes are not outcome values, but any continuous outcomes could be visualized this way.	Ease: 3* This task is easy if the specific parameter/option is given its own color-coding or axis, but otherwise the task would not be possible. Note that this example's axes are not outcome values, but any continuous outcomes could be visualized this way.	Ease: 1 Minimally supported, as only the parameters/options given a color coding or axis can be connected to their associated outcomes. Note that this example's axes are not outcome values, but any continuous outcome could be visualized this way.
Outcome Matrix	Detail: 0 Not supported	Detail: 1 All parameters and options are shown only as codes, so further understanding of the meaning and exact relationships of each choice cannot be determined from the figure alone. However, the structure of relationships within and between parameters is at last shown.	Ease: 3 In this example outcome values are treated as binary, and thus the shading makes this trivial. If the outcome values were continuous and even if continuous coloring were used instead, the range/similarity would still be similarly easy to see.	Ease: 2 In this example outcome values are treated as binary, and thus the shading makes this trivial. If the outcome values were continuous and even if continuous coloring were used instead, the frequency would require more mental effort and be less precise.	Ease: 2 The ease with which this task can be performed depends on the tree axis level for the parameter. For example, it is easy to compare the outcomes derived from options R1 to R2, but it is considerably more difficult to compare ECL1 to ECL3.	Ease: 2 The ease with which this task can be performed depends on the tree axis level for the parameter. For example, it is easy to compare the outcomes (and their relative frequencies) derived from options R1 to R2, but it is considerably more difficult to compare ECL1 to ECL3.	Ease: 3 The tree-axes allows it to be relatively easy to pick a given outcome value as a cell from the matrix, then mentally connect it to the associated options by following the path along the axes.
Varietal of matrix of outcomes: clustered outcome-option correlation matrix	Detail: 0 Not supported	Detail: 0 Not supported	Ease: 3 The outcome is a correlation value (1 to -1), which is encoded as a bipolar color scale. This allows a user to quickly scan even a large matrix and see the full range of outcome values that occur.	Ease: 2 The particular clustering of the bipolar color scale of values in this matrix example allows for a relatively easy mental estimate of frequency of outcomes. However, if the structure of the matrix did not so cleanly cluster similar values together in this way, it would be considerably harder to mentally estimate the frequency of different outcome values.	Ease: 1* * Only full specifications by team are shown as codes, so identifying outcomes as being connected to specific parameters/options is limited.	Ease: 1* * Only full specifications by team are shown as codes, so identifying outcomes as being connected to specific parameters/options is limited.	Ease: 1* * Only full specifications by team are shown as codes, so identifying outcomes as being connected to specific parameters/options is limited.
computation schematic	Detail: 3 A flow diagram and short description of each step in the process allows the user to quickly get a general overview of the process that produces the multiverse.	Detail: 3 Each parameter and its associated options are listed in human-readable stand-alone form, with space for a short description where needed, and is shown in the context of multiverse generation process.	Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported
Archetype Name	**Composition: Process** understand the process that defines and creates the universes being considered	**Composition: Parameters** understand the definition and composition of universe parameters and parameter values	**Outcome: Range** assess range or spread of outcome values across all universes	**Outcome: Frequency** assess overall frequency of outcome values across all universes	**Connect: Outcome Range** Connect parameters to outcomes by comparing similarity or range of outcome values across a subset of universes defined by a specific parameter value	**Connect: Outcome Frequency** Connect parameters to outcomes by comparing frequency of outcome values across a subset of universes defined by a specific parameter value	**Connect: Specific Outcomes** Connect parameters to outcomes by examining specific outcome values of interest and identifying parameter values that lead to those outcomes

Archetype Name	Archetype Representative Image	**Connect Combo: Outcome Range** Connect combinations of parameters to outcomes by comparing similarity or range of outcome values across subsets of universes defined by parameter values	**Connect Combo: Outcome Frequency** Connect combinations of parameters to outcomes by comparing frequency of outcome values across subsets of universes defined by parameter values	**Connect Combo: Idiosyncratic** Connect combinations of parameters to outcomes according to idiosyncratic patterns particular to a given visualization or analysis	**Validate: Metrics** Assess validity metrics of universes or compare metrics across parameter values	**Validate: Details** Assess validity of universes by examining the underlying details of analyses in each universe to interrogate their validity	assess uncertainty within and across universes	aggregate multiple sources of uncertainty across universes	compare multiverse to another theoretical multiverse	summarize robustness/sensitivity of the multiverse overall
				(correlation, step-wise function, or other specialty pattern of some kind)	(ex: model fit metrics - are all universes equally valid, reasonable, likely, arbitrary?)	(are there expert-experience revealed details that suggest some universes are more/less valid or reasonable?)	(confidence intervals, etc.?)	(ex: combining sampling and model fit concerns)	(compare to a 'null' construction, etc.)	(possibly fallacious counting/proportions, use of summary metrics, null test results, etc.)
specification curve - curve only		Ease: 0 Not supported	Ease: 0 Not supported	No special features identified	Not supported	Not supported				
specification curve - specification panel only		Ease: 0 Not supported	Ease: 0 Not supported	No special features identified	Not supported	Not supported				
specification curve - combined (Example of Super-additive Composite)		Ease: 2 When some combinations of options result in very similar outcomes, as shown in the red box on the left of this example, this task can be performed with reasonable ease. However, when the connections are more ambiguous or complex, such as across three parameters, the task becomes impractical in this format.	Ease: 1 One could attempt to mentally tabulate the frequency of outcomes connected to each relevant option, but this would be impractical in all but the simplest of cases.	"Idiosyncratic Specifications" is the original author's term for when some set of outcomes only occur under a small subset of all available option combinations. This figure shows an example of this for negative effect sizes, highlighted with the red box on the left of the figure.	Not supported	Not supported				
Tabular Multiverse Specification		Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported	Not supported	Not supported				
histogram of outcomes		Ease: 0 Not supported	Ease: 0 Not supported	No special features identified	Not supported	Not supported				
density plot of outcomes		Ease: 2 Comparing combinations of outcomes is easy if the combinations themselves are explicitly encoded as individual density lines, as with this example. Otherwise the task would not be possible at all.	Ease: 2 Comparing combinations of outcomes is easy if the combinations themselves are explicitly encoded as individual density lines, as with this example. Otherwise the task would not be possible at all.	Modality/Multimodality of outcome distributions is described by the author as being an indicator that some options - or combinations of options - are driving divergent sets of outcomes. In the publication this example is from, all outcomes are shown as one distribution, and the appearance of modality is presented as a kind of diagnostic indicator for the possible presence of some especially consequential analytical options.	Not supported	Not supported				
Varietal of density plot of outcomes: multiple option shaded variation		Ease: 0 Not supported	Ease: 0 Not supported	The author does not mention modality of the distribution explicitly, but identifying modality/multimodality would be possible with this visualization as it is a variation of the density plot of outcomes.	Not supported	Not supported				
Archetype Name	Archetype Representative Image	**Connect Combo: Outcome Range** Connect combinations of parameters to outcomes by comparing similarity or range of outcome values across subsets of universes defined by parameter values	**Connect Combo: Outcome Frequency** Connect combinations of parameters to outcomes by comparing frequency of outcome values across subsets of universes defined by parameter values	**Connect Combo: Idiosyncratic** Connect combinations of parameters to outcomes according to idiosyncratic patterns particular to a given visualization or analysis	**Validate: Metrics** Assess validity metrics of universes or compare metrics across parameter values	**Validate: Details** Assess validity of universes by examining the underlying details of analyses in each universe to interrogate their validity	assess uncertainty within and across universes	aggregate multiple sources of uncertainty across universes	compare multiverse to another theoretical multiverse	summarize robustness/sensitivity of the multiverse overall
vibration of effects volcano plot		Ease: 1 This task is arguably supported in a limited way through the use of the color of the dots for some parameters and options, combined with the red line and points that encode a meta-parameter in this multiverse.	Ease: 1 This task is arguably supported in a limited way through the use of the color of the dots for some parameters and options, combined with the red line and points that encode a meta-parameter in this multiverse.	"VoE Modality" (VoE terminology) and "VoE Patterns" (ex: "Janus effect") are visual patterns described in detail by the author. VoE Modality refers to clusters of output results, and is comparable to modality/multimodality of outcomes in other plots. VoE Patterns show relationships between statistical significance and effect sizes (outcome values), and thus require both variables to be plotted together.	Not supported	Not supported				
Composite Varietal of scatter plot of outcomes: parameter shaded density plots along axis margins		Ease: 0 Not supported	Ease: 0 Not supported	The author does not mention modality of the distribution explicitly, but identifying modality/multimodality would be possible with this visualization as it is a variation of the density plot of outcomes.	Not supported	This example is one of few that directly shows observations from the underlying dataset, and thus allows a user to examine how some choices impact the handling of individual datapoints. This can allow a user to examine the presence and handling of outliers, for example.				
Outcome Matrix		Ease: 2 The ease with which this task can be performed depends on the tree axis level for the parameters being considered. This task is thus made easy in some cases, and difficult in others.	Ease: 2 The ease with which this task can be performed depends on the tree axis level for the parameters being considered. This task is thus made easy in some cases, and difficult in others.	The original author describes looking for patterns of ambiguity or consistency between option combinations. In this example, R2 always leads to 'non-significant' outcomes, R1 with F3-F5 is always 'significant', and there is variability in outcomes with all other combinations of options. The visual patterns of horizontal/vertical stripes and colored-clusters point to unnamed relationships among options in the multiverse.	Not supported	Not supported				
Varietal of matrix of outcomes: clustered outcome-option correlation matrix		Ease: 1* * Only full specifications by team are shown as codes, so identifying outcomes as being connected to specific parameters/options is limited.	Ease: 1* * Only full specifications by team are shown as codes, so identifying outcomes as being connected to specific parameters/options is limited.	"Clustering" similarities between outcomes and analytical choices is supported in multiple ways in this visualization, including visual clustering by the color scale on outcome values, binary tree axes, color-coded spatial clusters on the y-axis, and color-coded binary outcomes along the x-axis.	Not supported	Not supported				
computation schematic		Ease: 0 Not supported	Ease: 0 Not supported	Ease: 0 Not supported	Not supported	Not supported				
Archetype Name	Archetype Representative Image	**Connect Combo: Outcome Range** Connect combinations of parameters to outcomes by comparing similarity or range of outcome values across subsets of universes defined by parameter values	**Connect Combo: Outcome Frequency** Connect combinations of parameters to outcomes by comparing frequency of outcome values across subsets of universes defined by parameter values	**Connect Combo: Idiosyncratic** Connect combinations of parameters to outcomes according to idiosyncratic patterns particular to a given visualization or analysis	**Validate: Metrics** Assess validity metrics of universes or compare metrics across parameter values	**Validate: Details** Assess validity of universes by examining the underlying details of analyses in each universe to interrogate their validity	assess uncertainty within and across universes	aggregate multiple sources of uncertainty across universes	compare multiverse to another theoretical multiverse	summarize robustness/sensitivity of the multiverse overall

meaning of colors

Master multiverse visualization reference note (not on diagram)

Regular multiverse visualization reference note

Reference note marked as not containing a relevant topic

Reference note marked as containing only saturated topics

General topic category - organized into at least a partial hierarchy

Rough, initial, developing note topic category

Analysis Category

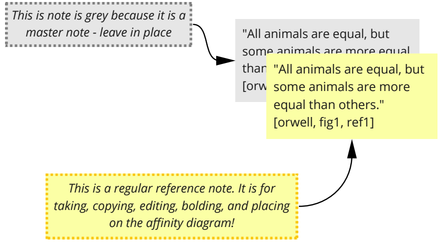
Inspection Category

Interpretation/Conclusion Category

The Big Question:
What are the goals and tasks that multiverse analysis visualizations support?

Rule 1:
Grey notes are stone - leave these master notes in place as a record.

Step 1:
Take a note from a figure frame.



This is note is grey because it is a master note - leave in place

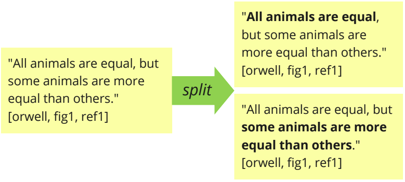
"All animals are equal, but some animals are more equal than [orw

"All animals are equal, but some animals are more equal than others." [orwell, fig1, ref1]

This is a regular reference note. It is for taking, copying, editing, bolding, and placing on the affinity diagram!

Rule 2:
When you take a note, make sure to split it fully so we don't miss something important.

Step 2:
Split the note fully into discrete goals or visualization tasks (as you see it), **bolding** the focus.



"All animals are equal, but some animals are more equal than others." [orwell, fig1, ref1]

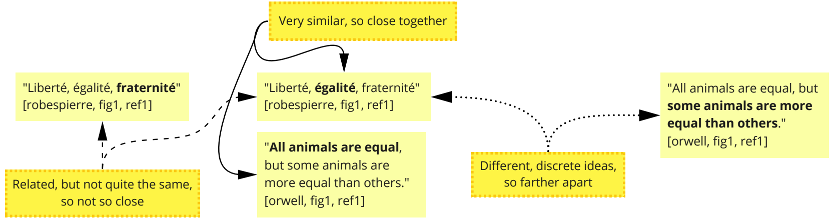
"**All animals are equal**, but some animals are more equal than others." [orwell, fig1, ref1]

"All animals are equal, but **some animals are more equal than others**." [orwell, fig1, ref1]

If you only split off a small piece of a note and think there may be more left to split, *italicize the entire un-split section* and leave it in the original figure frame to come back to later.

Rule 3:
Left-right distance on the affinity board has meaning.
Closeness = similarity

Step 3:
Place notes on the affinity diagram.
Almost touching = about the same
Separated a little = related
Wide space between = different topic



Very similar, so close together

"Liberté, égalité, **fraternité**" [robespierre, fig1, ref1]

"Liberté, **égalité**, fraternité" [robespierre, fig1, ref1]

"**All animals are equal**, but some animals are more equal than others." [orwell, fig1, ref1]

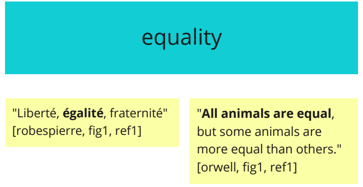
Related, but not quite the same, so not so close

Different, discrete ideas, so farther apart

"All animals are equal, but **some animals are more equal than others**." [orwell, fig1, ref1]

Rule 4:
Notes and clusters can be reorganized by anyone at any time. Move and arrange them whenever it makes sense.

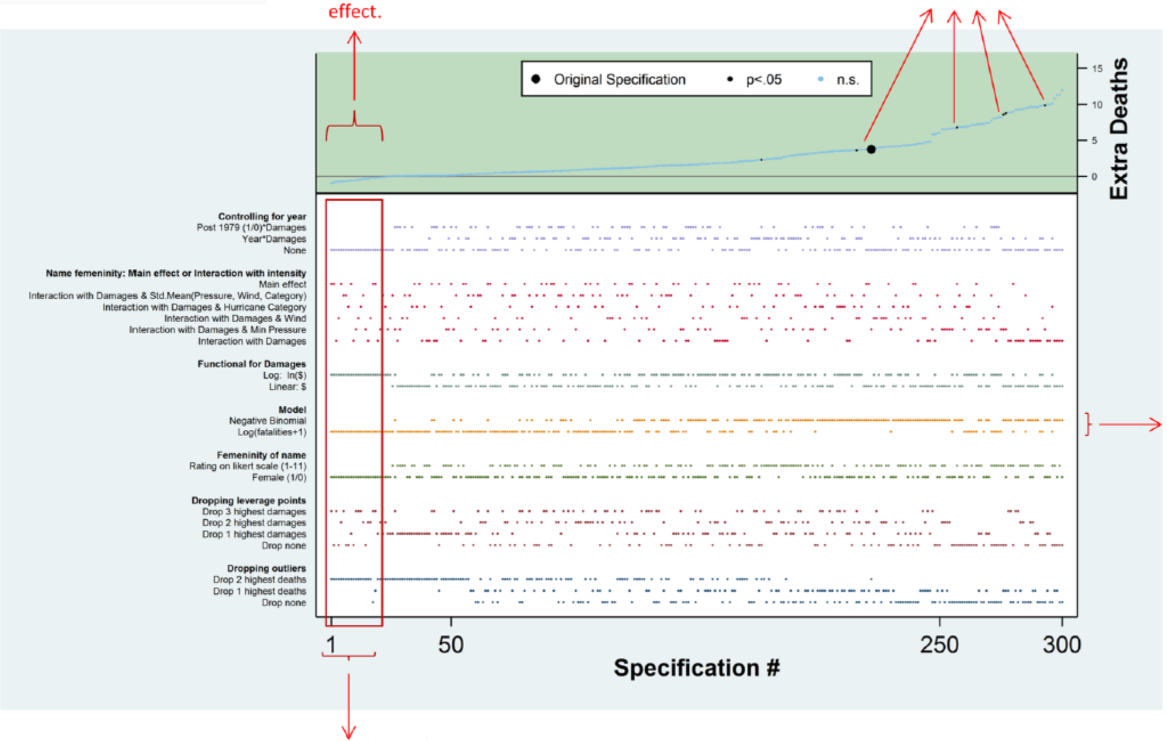
Step 4:
As clusters form, name what they have in common (fewer words are better). Put this name on a differently-colored note above the cluster of notes.



equality

"Liberté, **égalité**, fraternité" [robespierre, fig1, ref1]

"**All animals are equal**, but some animals are more equal than others." [orwell, fig1, ref1] simonsohn2015 - fig2



Only 20 specifications show a negative effect.

Of the 1728 specifications, 37 obtain *p*<.05

The largest estimates primarily involve negative binomial regressions

Negative point estimates requires idiosyncratic specifications.

"Figure 2. Descriptive Specification Curve. Each dot in the top panel (green area) depicts the marginal effect, estimated at sample means, of a hurricane having a female rather than male name; the dots vertically aligned below (white area) indicate the analytic decisions behind those estimates. A total of 1728 specifications were estimated; the figure depicts the 50 highest and lowest point estimates, and a random subset of 200 additional ones."
[simonsohn2015, fig2, ref1]

"The specification "curve" shows the estimated effect size across all specifications, sorted by magnitude, accompanied below by a "dashboard chart" indicating the operationalizations behind each result (see e.g., Figure 2). This enables readers to visually identify both the variation in effect size across specifications, and its covariation with operationalization decisions. Specification Curve analysis also includes an inferential component, which combines the results from all specifications into a joint statistical test. It assesses whether, in combination, all specifications reject the notion that the effect of interest does not exist."
[simonsohn2015, fig2, ref2]

"Among other differences with all of these approaches, Specification Curve Analysis: (i) helps identify the source of variation in results across specifications via a descriptive specification curve (see Figure 2), and (ii) provides a formal joint significance test for the family of alternative specifications, derived from expected distributions under the null. We are not aware of any existing approach that provides either feature."
[simonsohn2015, fig2, ref3]

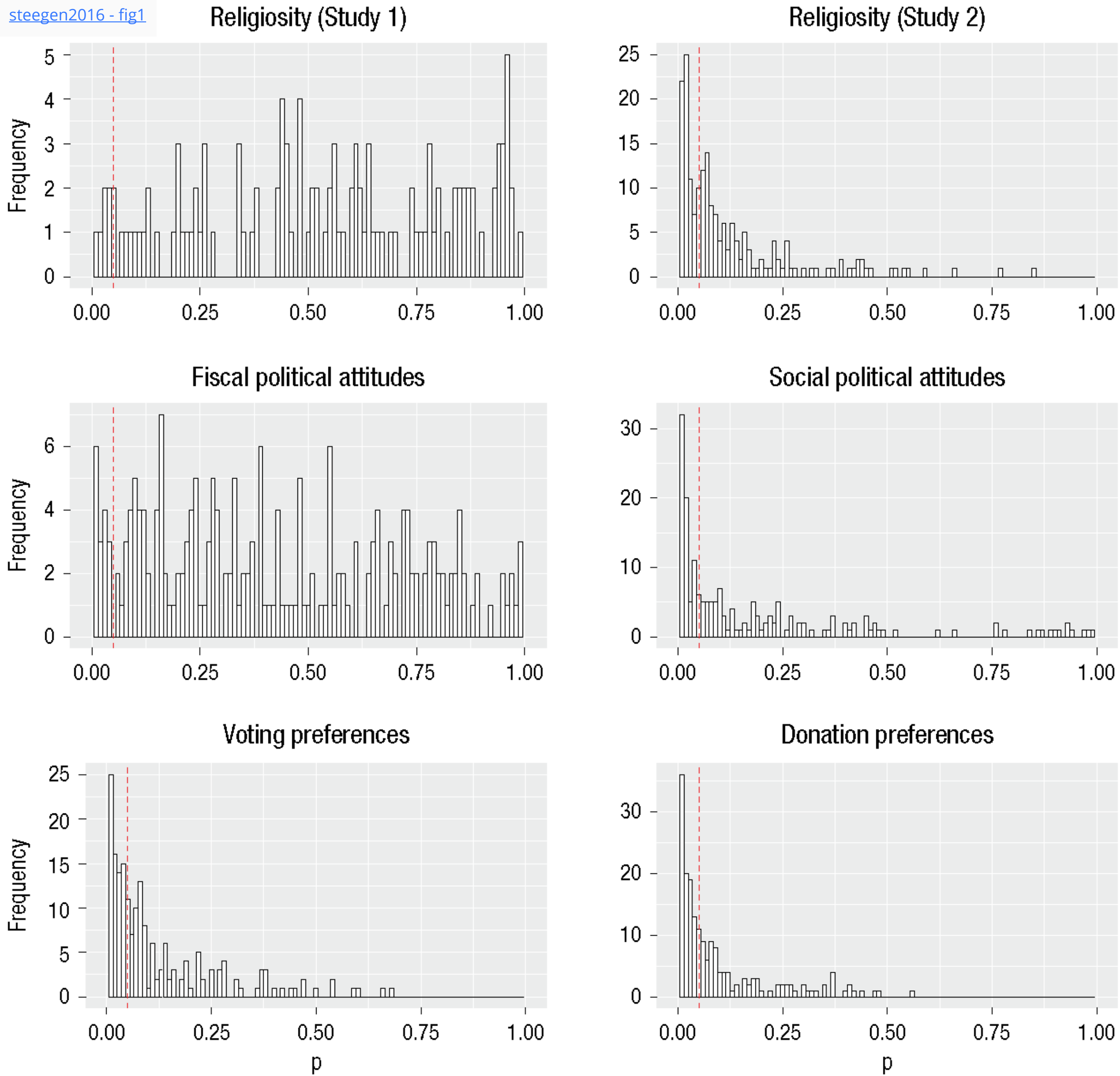
"Figure 2 reports the descriptive specification curve for the hurricanes example. The top panel depicts estimated effect size, in additional fatalities, of a hurricane having a feminine rather than masculine name. The figure shows that the majority of specifications lead to estimates of the sign predicted by the original authors (feminine hurricanes produce more deaths), though a very small minority of all estimates are statistically significant (p<.05). The point estimates range from -1 to +12 additional deaths."
[simonsohn2015, fig2, ref4]

"The bottom panel of the figure tells us which analytic decisions produce different estimates. For example, we can see that obtaining a negative point estimate requires a fairly idiosyncratic combination of operationalizations: (i) not taking into account the year of the storm, (ii) operationalizing severity of the storm by the log of damages, (iii) conducting an OLS regression, etc. A researcher motivated to show a negative point estimate would be able to report twenty different specifications that do so, but the specification curve shows that a negative point estimate is atypical."
[simonsohn2015, fig2, ref5]

"Returning to Figure 1, this appears to be a Panel C situation. Original authors and critics disagree on the set of valid specifications to run. The specification curve results from Figure 2 show that, while such disagreements may be legitimate and profound, we do not need to address them to determine what to make of the hurricanes data. In particular, the figure shows that even keeping the same set of observations as the original study and treating damages in the same way as treated in the original, modifying virtually any arbitrary analytical decision renders the original effect nonsignificant. Readers need not take a position on whether it does or does not make sense to include a damages x pressure interaction in the model to determine if the original findings are robust."
[simonsohn2015, fig2, ref6]

"Figure 2 shows that PNAS could have published nearly 1,700 letters showing individual specifications that make the effect go away (without deviating from the original red circle). It also could have published 37 responses with individual specifications showing the robustness of the findings. It would be better to publish a single specification curve in the original paper."
[simonsohn2015, fig2, ref7]

steegen2016 - fig1



"Fig. 1. Histogram of p values of the Fertility × Relationship status interaction on religiosity for the multiverse of 120 data sets in Study 1 and 210 data sets in Study 2 (Panels A and B), on fiscal and social political attitudes for the multiverse of 210 data sets in Study 2 (Panels C and D), and on voting and donation preferences for the multiverse of 210 data sets in Study 2 (Panels E and F). The dashed line indicates p = .05."
[steegen2016, fig1, ref1]

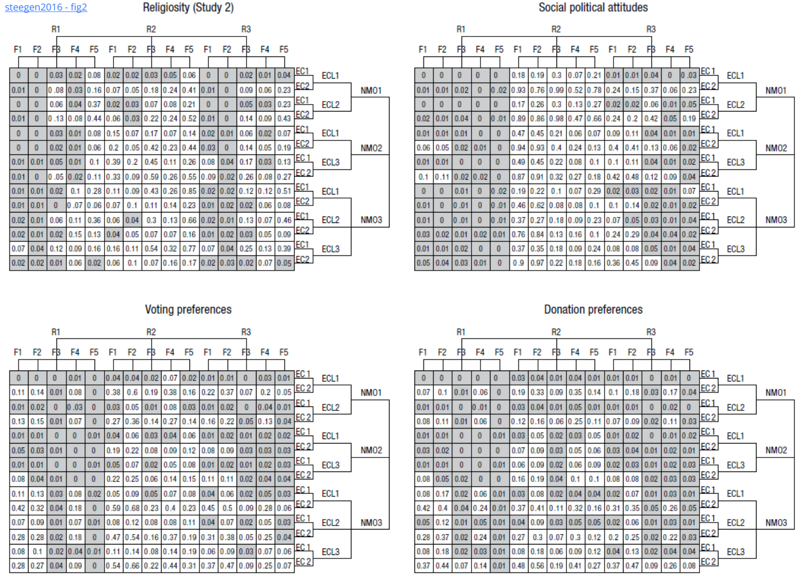
"Deriving the multiverse of statistical results. After constructing the data multiverse, the analysis of interest (in this case, an ANOVA or a logistic regression) is performed across all the alternatively constructed data sets. The results are shown in Panels A–F of Figure 1, each showing a histogram of the p values of the Fertility × Relationship interaction effect."
[steegen2016, fig1, ref2]

"For two variables—religiosity in Study 1 (Panel A) and fiscal political attitudes (Panel C)—the multiverse analysis reveals a near-uniform distribution, indicating that the p value for the interaction effect between fertility and relationship varies widely across the multiverse. For religiosity, 7 out of the 120 choice combinations lead to a significant interaction effect, whereas the remaining 94% lead to p values ranging from .05 to 1.0. For fiscal political attitudes, 8% of the 210 choice combinations lead to a significant interaction ($p < .05$), whereas the remaining choice combinations lead to p values across the entire range from .05 to 1.0."
[steegen2016, fig1, ref3]

"For the remaining four variables, roughly half of the choice combinations lead to a significant interaction effect. In particular, for religiosity in Study 2 (Panel B), 88 out of the 210 choice combinations (42%) lead to a p value smaller than .05. Regarding social political attitudes (Panel D), 49% of the p values is smaller than .05. Finally, 46% and 57% of the p values are smaller than .05 for voting (Panel E) and donation (Panel F) preferences, respectively. In these cases, it is informative to display the multiverse in greater detail by showing which constellation of choices corresponds to which statistical result. This allows to identify the key choices in data processing that are most consequential in the fluctuation of the statistical results."
[steegen2016, fig1, ref4]

"The multiverse analysis does not produce a single value summarizing the evidential value of the data, nor does it imply a threshold for an effect to reach to be declared robustly significant. Nevertheless, one might try to summarize the multiverse analysis more formally. One reasonable first step is to simply average the p values in the multiverse, in this case averaging all the numbers displayed in Figure 1 or 2. This mean value can be considered as the p value of a hypothetical preregistered study with conditions chosen at random among the possibilities in the multiverse and seems like a fair measurement in a setting where all of the possible data processing choices seem plausible (as in the example presented here, where the different options are drawn from other papers in the relevant literature)."
[steegen2016, fig1, ref5]

steegen2016 - fig2



Religiosity (Study 2)

			R1					R2					R3				
			F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
NMO1	ECL1	EC1	0	0	0.03	0.02	0.08	0.02	0.02	0.03	0.05	0.06	0	0	0.02	0.01	0.04
		EC2	0.01	0	0.08	0.03	0.16	0.07	0.05	0.18	0.24	0.41	0.01	0	0.09	0.06	0.23
	ECL2	EC1	0	0	0.06	0.04	0.37	0.02	0.03	0.07	0.08	0.21	0	0	0.05	0.03	0.23
		EC2	0.01	0	0.13	0.08	0.44	0.06	0.03	0.22	0.24	0.52	0.01	0	0.14	0.09	0.43
NMO2	ECL1	EC1	0	0	0.03	0.01	0.08	0.15	0.07	0.17	0.07	0.14	0.02	0.01	0.06	0.02	0.07
		EC2	0	0	0.02	0.01	0.06	0.2	0.05	0.42	0.23	0.44	0.03	0	0.14	0.05	0.19
	ECL3	EC1	0.01	0.01	0.05	0.01	0.1	0.39	0.2	0.45	0.11	0.26	0.08	0.04	0.17	0.03	0.13
		EC2	0.01	0	0.05	0.02	0.11	0.33	0.09	0.59	0.26	0.55	0.09	0.02	0.26	0.08	0.27
NMO3	ECL1	EC1	0.01	0.01	0.02	0.1	0.28	0.11	0.09	0.43	0.26	0.85	0.02	0.02	0.12	0.12	0.51
		EC2	0.01	0.01	0	0.07	0.06	0.07	0.1	0.11	0.14	0.23	0.01	0.02	0.02	0.06	0.08
	ECL2	EC1	0.02	0.01	0.06	0.11	0.36	0.06	0.04	0.3	0.13	0.66	0.02	0.01	0.13	0.07	0.46
		EC2	0.02	0.01	0.02	0.15	0.13	0.04	0.05	0.07	0.07	0.16	0.01	0.02	0.03	0.05	0.09
	ECL3	EC1	0.07	0.04	0.12	0.09	0.16	0.16	0.11	0.54	0.32	0.77	0.07	0.04	0.25	0.13	0.39
		EC2	0.02	0.02	0.01	0.06	0.02	0.06	0.1	0.07	0.16	0.17	0.02	0.03	0.02	0.07	0.05

Social political attitudes

			R1					R2					R3				
			F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
NMO1	ECL1	EC1	0	0	0	0	0	0.18	0.19	0.3	0.07	0.21	0.01	0.01	0.04	0	0.03
		EC2	0.01	0.01	0.02	0	0.02	0.93	0.76	0.99	0.52	0.78	0.24	0.15	0.37	0.06	0.23
	ECL2	EC1	0	0	0	0	0	0.17	0.26	0.3	0.13	0.27	0.02	0.02	0.06	0.01	0.05
		EC2	0.02	0.01	0.04	0	0.01	0.89	0.86	0.98	0.47	0.66	0.24	0.2	0.42	0.05	0.19
NMO2	ECL1	EC1	0.01	0.01	0.01	0	0	0.47	0.45	0.21	0.06	0.07	0.09	0.11	0.04	0.01	0.01
		EC2	0.06	0.05	0.02	0.01	0	0.94	0.93	0.4	0.24	0.13	0.4	0.41	0.13	0.06	0.02
	ECL3	EC1	0.01	0.01	0	0	0	0.49	0.45	0.22	0.08	0.1	0.1	0.11	0.04	0.01	0.02
		EC2	0.1	0.11	0.02	0.02	0	0.87	0.91	0.32	0.27	0.18	0.42	0.48	0.12	0.09	0.04
NMO3	ECL1	EC1	0	0	0.01	0	0.02	0.19	0.22	0.1	0.07	0.29	0.02	0.03	0.02	0.01	0.07
		EC2	0.01	0	0.01	0	0.01	0.46	0.62	0.08	0.08	0.1	0.1	0.14	0.02	0.01	0.02
	ECL2	EC1	0.01	0	0.01	0	0.01	0.37	0.27	0.18	0.09	0.23	0.07	0.05	0.03	0.01	0.04
		EC2	0.03	0.02	0.01	0.02	0.01	0.76	0.84	0.13	0.16	0.1	0.24	0.29	0.04	0.04	0.02
	ECL3	EC1	0.01	0.01	0.02	0	0	0.37	0.35	0.18	0.09	0.24	0.08	0.08	0.05	0.01	0.04
		EC2	0.05	0.04	0.03	0.01	0	0.9	0.97	0.22	0.18	0.16	0.36	0.45	0.09	0.04	0.02

Voting preferences

			R1					R2					R3				
			F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
NMO1	ECL1	EC1	0	0	0	0.01	0	0.04	0.04	0.02	0.07	0.02	0.01	0.01	0	0.03	0.01
		EC2	0.11	0.14	0.01	0.08	0	0.38	0.6	0.19	0.38	0.16	0.22	0.37	0.07	0.2	0.05
	ECL2	EC1	0.01	0.02	0	0.03	0	0.03	0.05	0.01	0.08	0.03	0.01	0.02	0	0.04	0.01
		EC2	0.13	0.15	0.01	0.07	0	0.27	0.36	0.14	0.27	0.14	0.16	0.22	0.05	0.13	0.04
NMO2	ECL1	EC1	0.01	0.01	0	0	0.01	0.04	0.06	0.03	0.04	0.06	0.01	0.02	0.01	0.02	0.02
		EC2	0.05	0.03	0.01	0	0	0.19	0.22	0.08	0.09	0.12	0.08	0.09	0.03	0.03	0.03
	ECL3	EC1	0.01	0.01	0	0	0.01	0.05	0.07	0.02	0.05	0.08	0.01	0.02	0.01	0.02	0.03
		EC2	0.08	0.04	0.01	0	0	0.22	0.25	0.06	0.14	0.15	0.11	0.11	0.02	0.04	0.04
NMO3	ECL1	EC1	0.11	0.13	0.03	0.08	0.02	0.05	0.09	0.05	0.07	0.08	0.04	0.06	0.02	0.05	0.03
		EC2	0.42	0.32	0.04	0.18	0	0.59	0.68	0.23	0.4	0.23	0.45	0.5	0.09	0.28	0.06
	ECL2	EC1	0.07	0.09	0.01	0.07	0.01	0.08	0.12	0.08	0.08	0.11	0.04	0.07	0.02	0.05	0.03
		EC2	0.28	0.28	0.02	0.18	0	0.47	0.54	0.16	0.37	0.19	0.31	0.38	0.05	0.25	0.04
	ECL3	EC1	0.08	0.1	0.02	0.04	0.01	0.11	0.14	0.08	0.14	0.19	0.06	0.09	0.03	0.07	0.06
		EC2	0.28	0.27	0.04	0.09	0	0.54	0.66	0.22	0.44	0.31	0.37	0.47	0.09	0.25	0.07

Donation preferences

			R1					R2					R3				
			F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
NMO1	ECL1	EC1	0	0	0	0	0	0.03	0.04	0.01	0.04	0.01	0.01	0.01	0	0.01	0
		EC2	0.07	0.1	0.01	0.06	0	0.19	0.33	0.09	0.35	0.14	0.1	0.18	0.03	0.17	0.04
	ECL2	EC1	0.01	0.01	0	0.01	0	0.03	0.04	0.01	0.05	0.01	0.01	0.01	0	0.02	0
		EC2	0.08	0.11	0.01	0.06	0	0.12	0.16	0.06	0.25	0.11	0.07	0.09	0.02	0.11	0.03
NMO2	ECL1	EC1	0.01	0.01	0	0	0.01	0.03	0.05	0.02	0.03	0.05	0.01	0.02	0	0.01	0.02
		EC2	0.03	0.02	0	0	0	0.07	0.09	0.03	0.05	0.06	0.03	0.04	0.01	0.02	0.01
	ECL3	EC1	0.01	0.01	0	0	0.01	0.06	0.09	0.02	0.06	0.09	0.02	0.03	0.01	0.02	0.03
		EC2	0.08	0.05	0.02	0	0	0.16	0.19	0.04	0.1	0.1	0.08	0.08	0.02	0.03	0.03
NMO3	ECL1	EC1	0.08	0.17	0.02	0.06	0.01	0.03	0.08	0.02	0.04	0.04	0.02	0.07	0.01	0.03	0.01
		EC2	0.42	0.4	0.04	0.24	0.01	0.37	0.41	0.11	0.32	0.16	0.31	0.35	0.05	0.26	0.05
	ECL2	EC1	0.05	0.12	0.01	0.05	0.01	0.04	0.09	0.03	0.05	0.05	0.02	0.06	0.01	0.03	0.01
		EC2	0.28	0.37	0.02	0.24	0.01	0.27	0.3	0.07	0.3	0.12	0.2	0.25	0.02	0.22	0.03
	ECL3	EC1	0.08	0.18	0.02	0.03	0.01	0.08	0.18	0.06	0.09	0.12	0.04	0.13	0.02	0.04	0.04
		EC2	0.37	0.44	0.07	0.14	0.01	0.48	0.56	0.19	0.41	0.27	0.37	0.47	0.09	0.26	0.08

"Fig. 2. Visualization of the multiverse of p values of the Fertility × Relationship status interaction on religiosity (Panel A), on social political attitudes (Panel B), on voting preferences (Panel C), and on donation preferences (Panel D) in Study 2, showing the dependence of the results on data processing choices. See Table 1 for an explanation of the acronyms."
[steegen2016, fig2, ref1]

"Such a closer inspection is provided in Figure 2, showing a grid of p values for each of these four variables. In each panel, the cells show the different p values that can be obtained across all choice combinations for data processing. Depending on whether the p value is smaller or larger than the α level, the cells are colored gray or white, respectively."
[steegen2016, fig2, ref2]

"For religiosity in Study 2 (Panel A), most data sets constructed under the second option for relationship assessment (R2) yield a nonsignificant interaction effect. The first and third options (R1 and R3) consistently lead to a significant interaction effect in combination with the first and second option for fertility assessment (F1 and F2) and to a nonsignificant interaction effect in combination with F5, whereas data sets constructed under R1 or R3 in combination with F3 or F4 lead to more fluctuating conclusions, depending on the other choices for data processing. The different exclusion criteria and cycle day estimation options do not seem to have a large impact on fluctuation in the statistical conclusion."
[steegen2016, fig2, ref3]

"For social political attitudes (Panel B), the statistical conclusion is highly robust for the first and second option for relationship status assessment (significant for R1 and nonsignificant for R2). Using the third option for relationship status assessment (R3) leads to more fluctuation, depending on the choices for the other processing steps."
[steegen2016, fig2, ref4]

"Finally, for voting and donation preferences (Panels C and D, respectively), it is hard to extract a consistent pattern of fluctuation across the different choice combinations. It seems that all arbitrary choices for data processing can have an impact on whether the obtained data set will lead to a significant or a nonsignificant outcome."
[steegen2016, fig2, ref5]

"In our demonstration, we started from a single set of raw data and performed both a single data set analysis as well as a multiverse analysis. Comparison of both types of analysis highlights the dramatic impact of going beyond an N = 1 sample from the multiverse. For religiosity in Study 1, the arbitrary data processing choices made in the single data set analysis led to a significant result. Placing this significant result in the multiverse of statistical results illustrates the risk of running a single data set analysis. The multiverse analysis revealed that almost all choice combinations for data processing lead to large p values. As such nonsignificant findings in general represent nothing more than uncertainty, this pattern of results clearly raises serious questions regarding the finding on the effect of fertility found in the single data set analysis, and should make a researcher hesitant to trust the single data set finding. The effect of fertility on religion seems too sensitive to arbitrary choices and thus too fragile to be taken seriously."
[steegen2016, fig2, ref6]

"For most other variables, there was considerable ambiguity: The interaction seemed to be significant across about half of the arbitrary choice combinations. In these cases, the conclusion on the effect of fertility strongly depends on the evaluation of the different processing options. Both the authors performing the multiverse analysis and the readers of the research can construct arguments in favor or against certain choices, and the validity of these arguments will help drawing the conclusion. For example, if additional information suggests that the fifth option of assessing fertility is clearly superior, then Panel A of Figure 2 indicates that there is little evidence for an effect of fertility on religiosity in Study 2. On the other hand, if additional information suggests that the second option of assessing fertility is clearly superior, then most choice combinations lead to a significant interaction effect."
[steegen2016, fig2, ref7]

"If no strong arguments can be made for certain choices, we are left with many branches of the multiverse that have large p values. In these cases, the only reasonable conclusion on the effect of fertility is that there is considerable scientific uncertainty. One should reserve judgment and acknowledge that the data are not strong enough to draw a conclusion on the effect of fertility. The real conclusion of the multiverse analysis is that there is a gaping hole in theory or in measurement, and that researchers interested in studying the effect of fertility should work hard to deflate the multiverse. The multiverse analysis gives useful directions in this regard."
[steegen2016, fig2, ref8]

"In general, deflating the multiverse involves developing a better and more complete theorizing of the constructs of interest and improving their measurement. Both routes for deflating the multiverse are illustrated in our case study. A first approach involves improving the experimental material and design. For example, the detailed multiverse examination shown in Figure 2 revealed that a lot of fluctuation hinged on the different choices for relationship status assessment. Thus, apparently, this type of research could benefit from a better way of assessing relationship status. Looking at the alternative options for assessing relationship status, it seems that the ambiguous Option 2 in the relationship status question could be formulated more precisely, so that relationship status assessment is no longer an arbitrary choice. This would have narrowed down the multiverses to 40 and 70 choice combinations in Study 1 and 2, respectively."
[steegen2016, fig2, ref9]

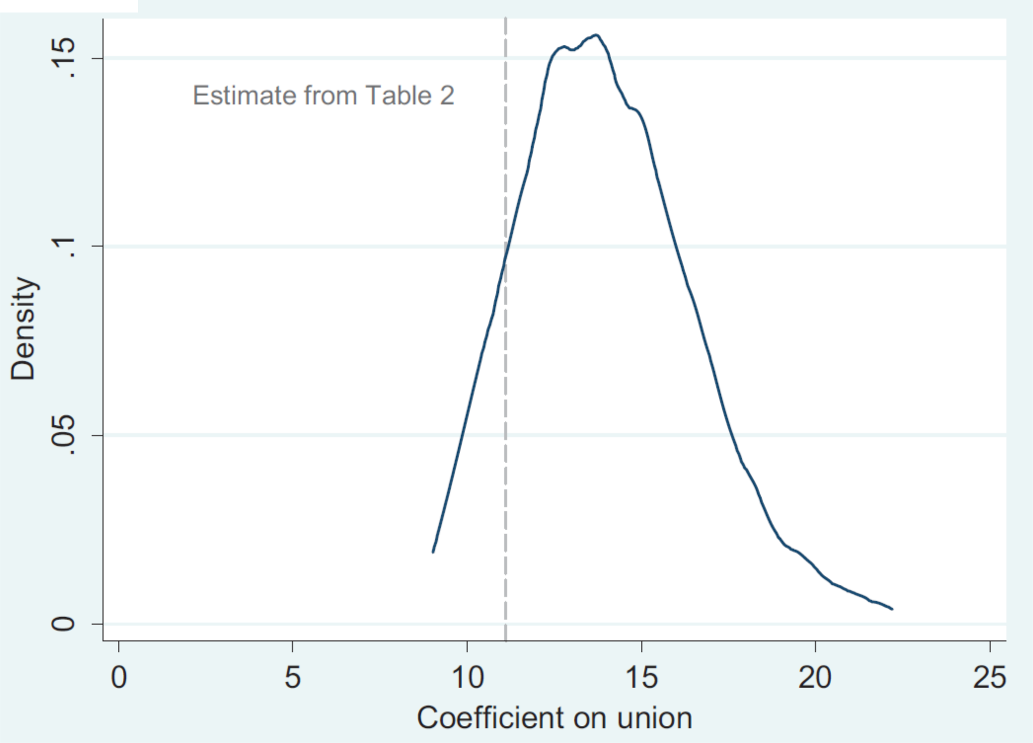
"A second approach for deflating the multiverse involves developing more complete and more precise theory in such a way that some options are theoretically superior than others, and it should be preferred when constructing data sets. For example, a great deal of variation in the results appeared to be driven by the different options for assessing fertility. Clearly, for this type of research, developing and applying a more precise way of assessing fertility should become a research priority. The availability of different reasonable options for estimating next menstrual onset or for classifying women into a high or low fertility group based on their cycle day stems from the fact that a precise theoretical foundation is lacking (Harris, 2013). The development of elaborated theories concerning these issues would narrow down the number of alternative options and deflate fluctuation. Recently, Gangestad et al. (2016) have recommended assessing fertility based on the detection of surges in luteinizing hormone, ideally in a within-subjects design. It is of note that this alternative strategy of assessing fertility was used in several papers by Durante (e.g., Durante et al., 2011; Durante et al., 2012)."
[steegen2016, fig2, ref10]

"As is evident from our demonstration, a multiverse analysis is highly context-specific and inherently subjective. Listing the alternative options for data construction requires judgment about which options can be considered reasonable and will typically depend on the experimental design, the research question, and the researchers performing the research. Whereas this subjectivity may seem undesirable, presenting results given only a single combination of reasonable options is much more misleading; indeed, one of the sources of the current crisis in scientific replication is that researchers traditionally have taken p values at face value without considering the multiplicity of choices in data construction."
[steegen2016, fig2, ref11]

"A related point is that not all options are necessarily exactly interchangeable. Some options might seem better than others, at least for some researchers. If such is the case, this knowledge can be used to construct arguments for interpreting results such as those shown in Figure 2. However, a multiverse analysis should involve all plausible construction alternatives, not just the most plausible ones. When only one choice is clearly and unambiguously the most appropriate one, variation across this choice is uninformative."
[steegen2016, fig2, ref12]

"The richness of possibilities for different data processing choices present in the raw data made the case study exceptionally suitable for the demonstration of a multiverse analysis. We do not expect that all multiverses will consist of such a numerous amount of data sets. The fact that more typical multiverses will tend to be smaller does not make a multiverse analysis less necessary. Even when confronted with only one arbitrary data processing choice, researchers should be transparent about it and reveal the sensitivity of the result to this choice."
[steegen2016, fig2, ref13]

young2017 - fig1



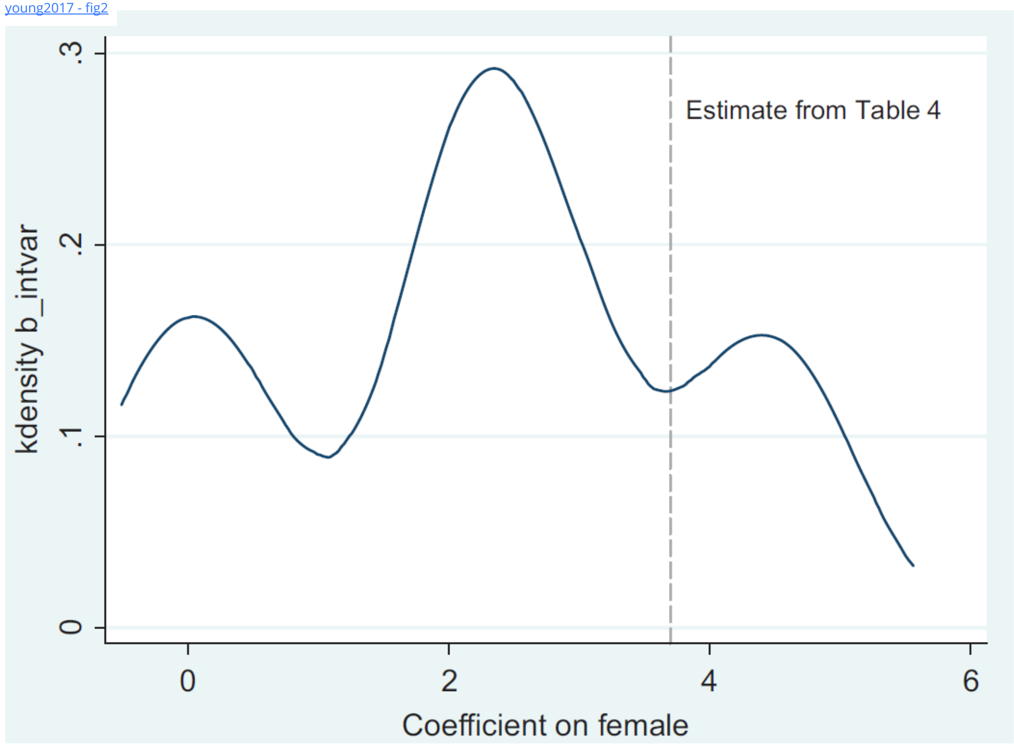
"Figure 1. Modeling distribution of union wage premium. Note: Kernel density graph of estimates from 1,024 models. Vertical line indicates the preferred estimate of an 11 percent union wage premium as reported in Table 2."
[young2017, fig1, ref1]

"Application 1: The Union Wage Premium. Before proceeding to more detailed aspects of model robustness, we illustrate the basic approach—robustness to the choice of controls—using a data set included in Stata, the 1988 wave of the National Longitudinal Survey of Women. We estimate the effect of union membership on wages (i.e., the union wage premium) controlling for 10 other variables that may be correlated with hourly wages (and union membership; (see Table 2). The coefficient on union, 11.1, means that union members earn about 11 percent more than nonunion members. This is on the low side of conventional estimates, which center around a 15 percent premium (Hirsch 2004)."
[young2017, fig1, ref2]

"Next, we report the robustness of this finding to the choice of control variables in the model. Does this finding hinge on sets of control variables, or do the findings hold regardless of what assumptions are made over the control variables? Table 3 shows that there are 1,024 unique combinations of the control variables. Running each of these models and storing all of the estimates, we graph the modeling distribution in Figure 1. The result appears strongly robust. The estimated coefficient on union membership is positive and significant in every possible combination of the control variables: both the sign stability and the significance rate are 100 percent. With this list of possible controls, and using OLS, it is not possible to find an opposite signed or even nonsignificant estimate. Figure 1 shows the modeling distribution as a density graph of all the estimates calculated; the vertical line marks the 11 percent wage premium estimate from Table 2. Estimates as low as 9 percent and as high as over 20 percent are possible in the model space."
[young2017, fig1, ref3]

"As shown in Table 3, the average estimate across all of these models is 14.0. This simply represents the average coefficient across all models and is not necessarily the most theoretically defensible. The average sampling standard error is 2.4, and the modeling standard error is 2.5—uncertainty about the estimate derives equally from the data and from the model. The combined total (sampling and modeling) standard error is 3.5.6 The robustness ratio—the mean estimate divided by the total standard error—is 4.05. By the standard of a t-test, this would be considered a strongly robust result, which agrees with the 100 percent sign stability and significance rates. Our conclusion is that, within the scope of these model ingredients, the positive union wage premium is a clear and strongly robust result. This suggests that the decline of unionization in America may well have contributed to middle-class wage stagnation—and not just for male workers (Rosenfeld 2014)."
[young2017, fig1, ref4]

young2017 - fig2



"Figure 2. Modeling distribution of the gender effect on mortgage lending. Note: Kernel density graph of estimates from 256 models. See Table 5 for more information about the distribution. The vertical line shows the preferred estimate from Table 4 (3.7 percent higher acceptance rate for women)."
[young2017, fig2, ref1]

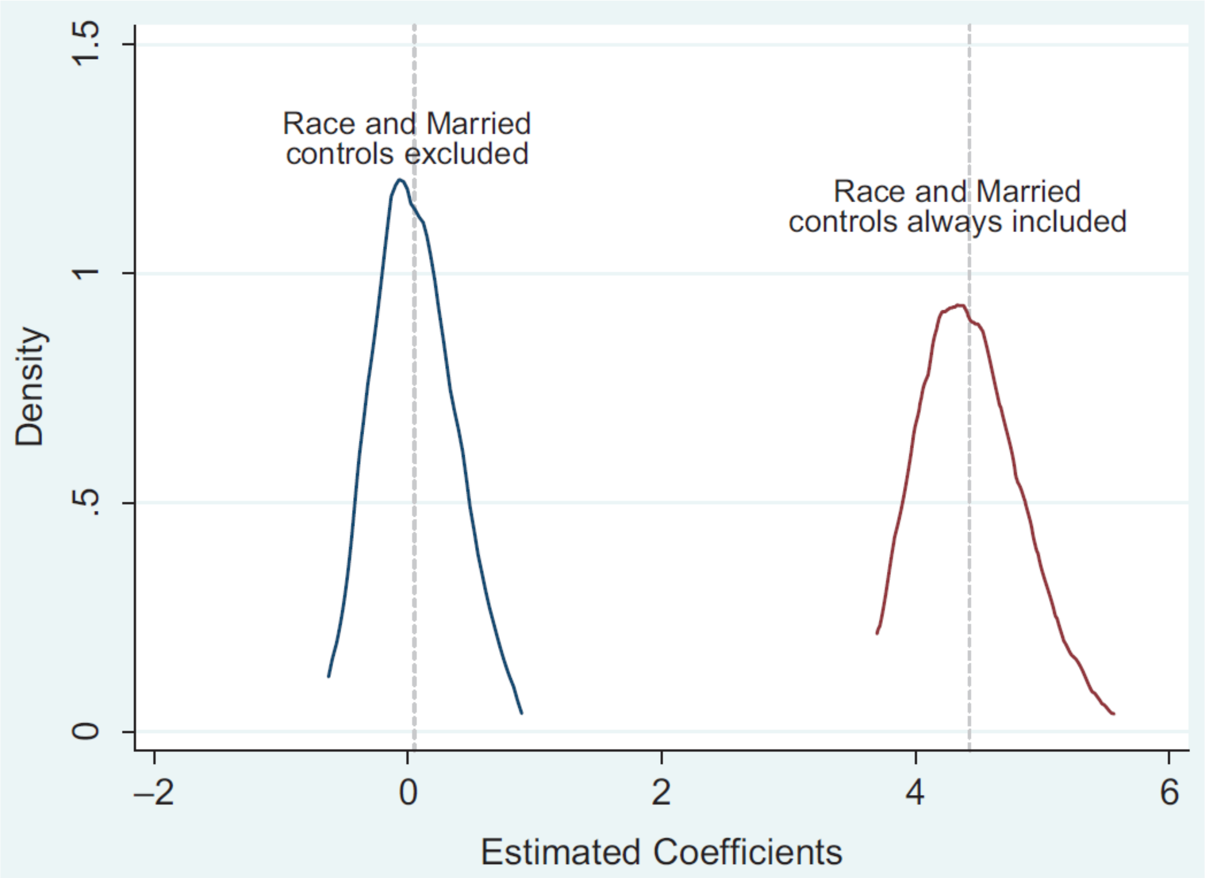
"Application 2: Mortgage Lending by Gender. Next, we draw on an influential study of discrimination in mortgage lending conducted by the Federal Reserve Bank of Boston (Munnell et al. 1996). What factors lead banks to approve an individual's mortgage application? The initial study focused on race, showing compelling evidence of discrimination against black applicants. In this application, we focus on the effect of an applicant's gender. We regress the mortgage application acceptance rate on a dummy for female as well as other variables capturing the demographic and financial characteristics of applicants. The results (Table 4) interestingly show that women are 3.7 percent more likely to be approved for a mortgage, suggesting banks favor female applicants—perhaps because women are seen as more prudent and responsible with household finances."
[young2017, fig2, ref2]

"However, when we relax the assumption that any one of these control variables must be in the model—allowing us to consider all possible combinations of the controls—there is much uncertainty about the estimate. Table 5 reports the model robustness results. Across the 256 possible combinations of controls, the effect of gender is typically positive but only 25 percent of the estimates are statistically significant. And 12 percent of the estimates have the opposite sign (though none of those estimates are significant)."
[young2017, fig2, ref3]

"The mean estimate from all models is 2.29 and the average sampling standard error is 1.61—indicating that the mean estimate is not statistically significant. In addition, the modeling standard error is 1.60—the estimates vary across models just as much as would be expected from drawing new samples. The total standard error—incorporating both sampling and modeling variance is 2.27, roughly the same size as the estimate itself, yielding a robustness ratio of 1.01."
[young2017, fig2, ref4]

"Figure 2 shows the distribution of estimates from all the 256 models with a vertical line showing the "preferred estimate" of 3.7 percent from Table 6. The modeling distribution is multimodal with clusters of estimates around zero, 2.3, and 4.5 percent. It seems hard to draw substantive conclusions from the evidence without knowing more about the modeling distribution. Why do these estimates vary so much? Why is the distribution so non-normal? What combinations of control variables are critical to finding a positive and significant result? These questions lead us to the next stage in our analysis: understanding model influence."
[young2017, fig2, ref5]

young2017 - fig3

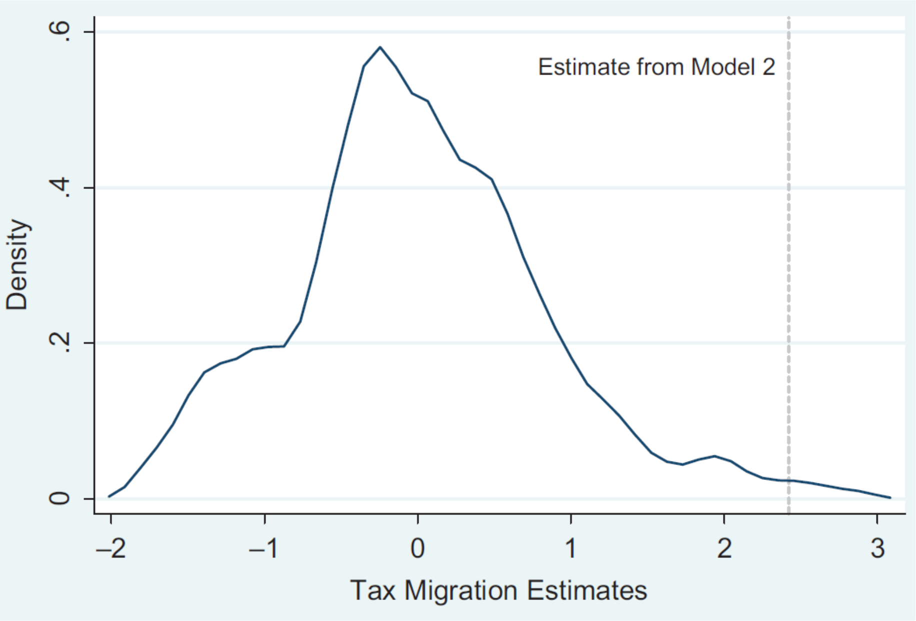


"Figure 3. Modeling distributions for the gender effect under different assumptions." [young2017, fig3, ref1]

"Influence Analysis of the Gender Effect in Mortgage Lending: For the mortgage lending analysis, Table 6 shows the influence of control variables on the coefficient of interest (female). The Delta-Beta effect of controls is reported in order of absolute magnitude influence. To aid interpretation, we also report Delta-Beta as a percent change in the estimate from the mean of the modeling distribution (2.29 as in Table 7). Two control variables clearly stand out as most influential: marital status and race. The influence estimate for marriage shows that, all else equal, when controlling for marital status the coefficient on female increases by 2.47, more than doubling the mean estimate across all models. Controlling for race (with the dummy variable “black”) also increases the effect size of gender by 1.91, a full 83 percent higher than the mean estimate. The other controls have much less impact on the estimate and have little model influence."
[young2017, fig3, ref2]

"In essence, there are two distinct modeling distributions to consider which are plotted in Figure 3. In one set of models, the controls for race and marital status are always excluded but all other controls are allowed in the model space (which gives 128 models). Under these assumptions, the estimates of the gender effect are tightly centered around zero, with an almost even split between positive (52 percent) and negative (48 percent) estimates, none of which are statistically significant. Here, there is no evidence at all for a gender effect. In contrast, the second distribution is defined by the opposite assumption: race and marital status must be in the model, but all combinations of the other controls are possible. Under these assumptions, the estimates cluster around a 4.5 percent higher mortgage acceptance rate for women. Both the significance rate and the sign stability are 100 percent— complete robustness. In order to draw robust conclusions from these data, one must make a substantive judgment about two key modeling assumptions: the inclusion of race and marital status. None of the other model ingredients affect the basic conclusion. These two model assumptions determine the results."
[young2017, fig3, ref3]

young2017 - fig4



"Figure 4. Modeling distribution of tax migration estimates. Note: Kernel density graph of estimates from 24,576 models."
[young2017, fig4, ref1]

"In Table 7, we show our main analysis. Model 1 includes just the base populations of the origin and destination states and the income tax differences between them. When the income tax rate in the origin state is higher, there tends to be more migration from the origin state to other (lower tax) destinations. Migration flows are 1.4 percent higher for each percentage point difference in income tax, but the estimate is not statistically significant. Model 2 adds in controls for contiguity, distance, the sales and property tax rates, state income, and a measure of natural amenities (topographical/landscape variability). The tax effect is now larger and statistically significant. For each one point difference in the tax rate, migration flows are 2.4 percent higher. Finally, in model 3, when using an IRS migration data with the same set of controls, we find a similar significant effect. This gives seemingly compelling evidence that high income taxes cause migration to lower tax states."
[young2017, fig4, ref2]

"What this fails to show, however, is the extreme model dependence in this conclusion. Models 2 and 3 are knife-edge specifications, carefully selected to report statistically significant results, and remarkably unrepresentative of the overall modeling distribution. Both models are highly sensitive to adding or deleting insignificant controls, and this set of controls is the only combination among many thousands that yields a significant result in both the ACS and IRS data."
[young2017, fig4, ref3]

"We embrace a wide robustness analysis that relaxes assumptions about possible controls, possible data sources for migration, and alternative estimation commands. There are two controls that we see as absolutely critical to the gravity model: base populations of the origin and destination states. Combinatorially including or excluding these variables produces models that we regard as nonsense, so we impose the assumption that they must be in all models. However, we leave as debatable the controls for distance, contiguity, other tax rates, economic performance of the states, and a rich set of natural amenities which have been previously shown to influence migration (McGranahan 1999). All possible combinations of these controls give 4,096 models. Moreover, we test these models across the two alternative data sets for migration and population (ACS and IRS), and across three different estimation strategies (Poisson, negative binomial, and OLS log-linear). For each data set, there are three possible estimation commands, and for each (data set X estimation command), there are 4,096 possible sets of controls. This robustness analysis, therefore, runs 24,576 plausible models."
[young2017, fig4, ref4]

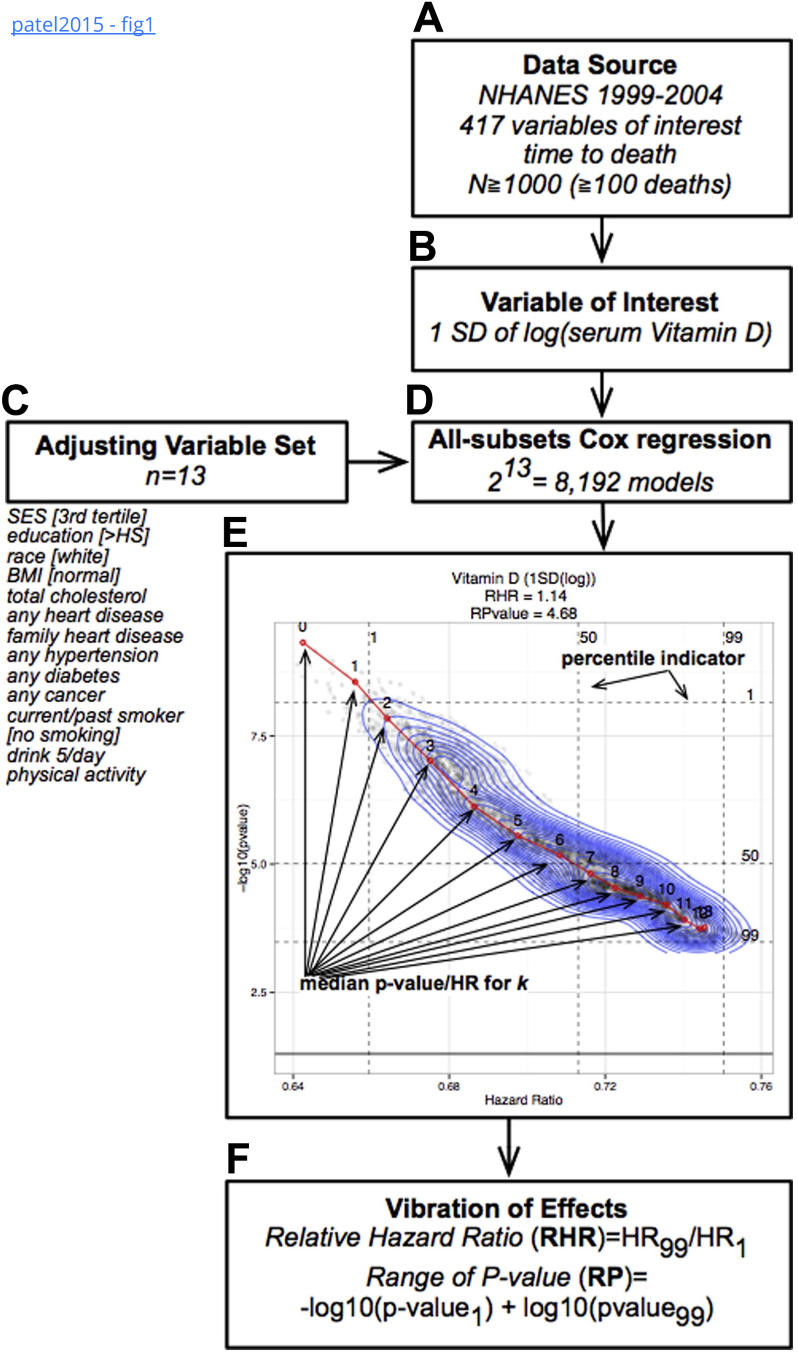
"As shown in Table 8, the tax coefficient is statistically significant in only 1.5 percent of all models. The mean estimate is almost exactly zero, and estimates are evenly split between positive tax flight estimates (48.9 percent) and wrong-signed negative estimates (51.1 percent). Among the few statistically significant results, the great majority are wrong signed: estimates with negative signs indicate migration toward higher tax states. Only 0.2 percent of estimates are significantly positive compared to 1.3 percent that are significant and wrong signed. The robustness ratio—the mean estimate divided by the total standard error—is 0.01. The modeling distribution is relatively normal: There are no critically important modeling decisions that generate bimodality in the estimates. As shown in Figure 4, the significant estimates reported in Table 7 above are extreme outliers in the modeling distribution."
[young2017, fig4, ref5]

"In this case, when the robustness analysis is so overwhelmingly nonsupportive, the influence analysis has less to work with. However, there are a few informative points. Compared to Poisson, the negative binomial and OLS log-linear models give less positive estimates. Estimates from themodels using IRS rather than ACS data are more positive. This suggests that the most supportive evidence will come from using Poisson with the IRS data (reported as model 3 above), and the least supportive evidence will come from using OLS log-linear models with ACS data. Yet, even when we narrow our robustness testing to the most supportive estimator (Poisson) and data set (IRS), there is weak support: while the sign stability is 100 percent, the income tax effect is significant in only 1 percent of those models.13 By control variables, the sales tax rate, average income, and the property tax rate have the most positive influence—generating more positive estimates of tax flight when these controls are included. (Note, however, that none of these controls were significant in model 3.) All other controls push the tax migration estimate toward a zero or wrong-signed result, and virtually must be excluded to support the hypothesis."
[young2017, fig4, ref6]

"In these results, we see another case where the most significant control has among the least model influence. In the main regression models 2 and 3, distance between the states is a powerful predictor of migration flows, showing t-statistics greater than 10. Yet, including distance in the model has almost no influence on the tax migration estimate (-6.3 percent in Table 9)."
[young2017, fig4, ref7]

"While it is possible to support the tax flight hypothesis with a few knife-edge model specifications, there is remarkably little support even in a more narrow and supportive robustness analysis. This shows how extreme the difference can be between a curated selection of regression results (Table 7) and a rigorous robustness analysis (Table 8). While one offers an existence proof that a significant result can be found, the weight of the evidence frommany crediblemodels gives scant support to the tax migration hypothesis. It remains technically possible that the one-in-a-thousand specifications of Table 7 present the best, most theoretically compelling estimates. If so, authorswould need to carefully explain to readers why such painstakingly exact model assumptions are required, and why virtually any departure from model 2 or 3 fails to support the conclusions."
[young2017, fig4, ref8]

patel2015 - fig1



"Fig. 1. Vibration of effects (VoE) computation schematic. (A) Data source. (B) Choose a variable of interest. (C) Construct a set of adjustment variables from a set of 13 socioeconomic, demographic, or health-related variables. Reference level is in the square brackets. (D) All subsets Cox regression for each 8,193 models. (E) Visualization ("volcano plot") of -log10(P-value) vs. effect size (HR). The median HR and P-value of the number of adjustment variables (k) in the model is in red. The 1st, median, and 99th percentile of the -log10(P-value) and HR are depicted in the dotted line. (F) Compute VoE summary statistics, the relative hazard ratio (RHR) and relative P-value (RP)."
[patel2015, fig1, ref1]

"VoE is estimated by computing the hazard ratio (HR) and P-value for a variable of interest while adjusting for all possible combinations of adjustments from a finite set of adjustment variables. Our algorithm for computing the VoE for a variable x (e.g., serum vitamin D) is shown in Fig. 1."
[patel2015, fig1, ref2]

"First, we downloaded 417 self-reported, clinical, and molecular measures with linked all-cause mortality information in participants from NHANES 1999-2004 (Fig. 1A). Mortality information was collected from the date of the survey participation through December 31, 2006, and ascertained via a probabilistic match between NHANES and National Death Index (NDI) death certificate information [21]. We chose variables of interest that had data on at least 1,000 participants and at least 100 death events during follow-up."
[patel2015, fig1, ref3]

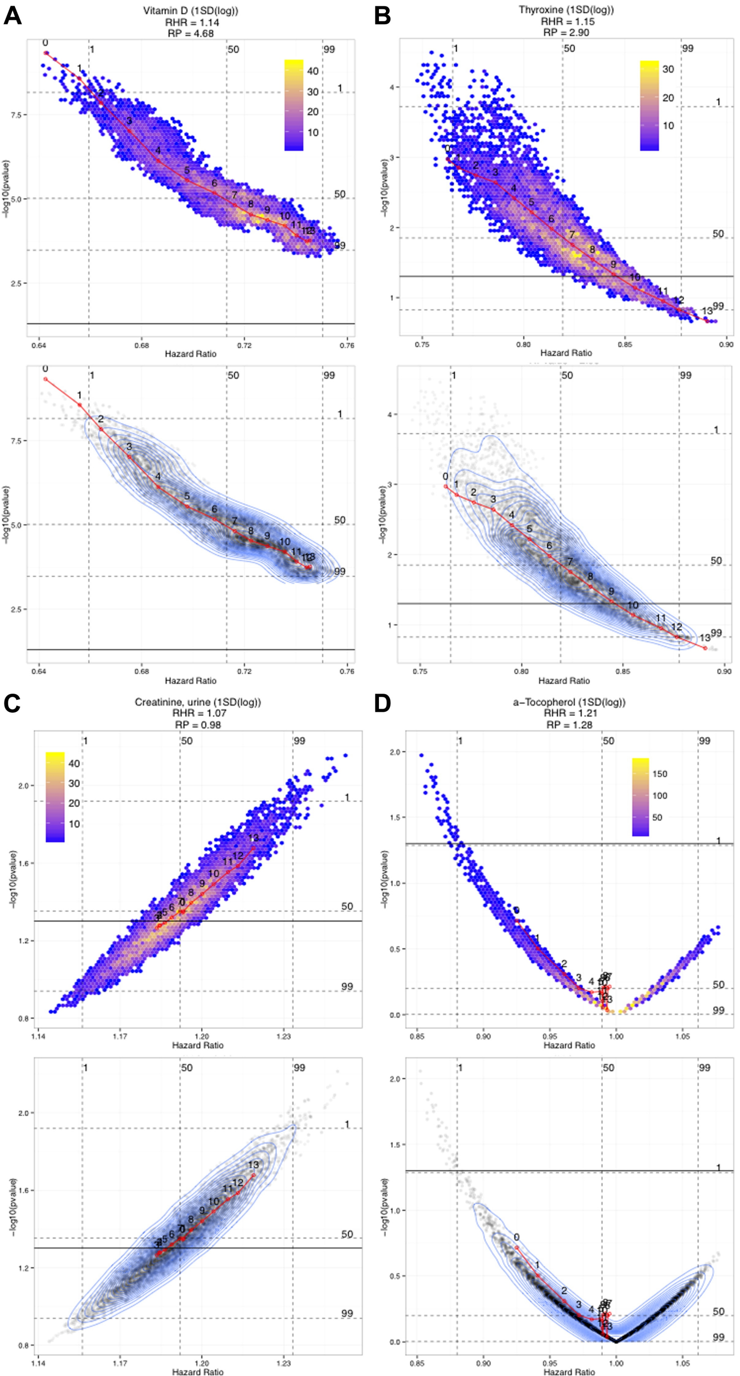
"Next, we describe the VoE methodology for the association between serum vitamin D and all-cause mortality (Fig. 1B). The total number of combinations of adjusting variables from the set of n=13 total adjustments is 8,192 (or, in general, 2^n models, Fig. 1C). We chose a set of 13 variables as the set of possible adjustments (Fig. 1B, C, Appendix at www.jclinepi.com). Because there is no consensus on what variables should (or should not) be included as adjustments in association with all-cause mortality,we based the selection of these 13 variables on a large meta-analysis of 80 studies of physical activity on all-cause mortality [29]. The most common adjustment variables in these 80 investigations included (in decreasing order of frequency) age, smoking, BMI, hypertension, diabetes, cholesterol, alcohol consumption, education, income, sex, family history of heart disease, heart disease, and any cancer. Because age and sex are wellknown factors related to mortality, we chose to keep these in all models ("baseline" variables). The HR and the respective P-value for the association of that variable with all-cause mortality are estimated for all 8,193 models with different combinations of 13 adjusting variables using Cox proportional hazards time-to-event regression (Fig. 1D). We visualized the VoE for a given variable by plotting the HR vs. -log10(P-value) as two-dimensional histogram and a contour plot (Fig. 1E)."
[patel2015, fig1, ref4]

"We created metrics to express the distributions of VoE for a variable (Fig. 1F). The first was the RHR, the ratio of the 99th percentile and 1st percentile HR. The RHR connotes the spread of HRs for different combinations of adjustments. The second was the RP, which is the difference between the 99th and 1st percentile of -log10(P-value). The RP measures the range of P-values over all estimates. We also assessed whether associations appeared on both sides of the null (HR <1 and HR >1): depending on what adjustments are chosen, the results may suggest that the variable of interest is associated with either increased or decreased mortality."
[patel2015, fig1, ref5]

"We also visualized trends corresponding to the number of adjusting variables (k), plotting the median effect size and P-value for each k from 0 to 13. We recorded the proportion of estimates that achieved different levels of nominal statistical significance (P < 0.05, 0.0001)."
[patel2015, fig1, ref6]

"The 417 variables included 179 serum or urine biomarkers of environmental exposures (e.g., serum cadmium, mercury, or pesticide level), 9 self-reported behavioral factors such as smoking and physical activity, 84 self-reported nutritional intake information (from a food frequency questionnaire), 27 self-reported health conditions (e.g., diabetes), 92 clinical factors (e.g., BMI and cholesterol), and 13 sociodemographic variables (e.g., income). All continuous variables were log transformed and z standardized for comparison. Appendix (at www.jclinepi.com) describes these 417 variables."
[patel2015, fig1, ref7]

patel2015 - fig2



"Fig. 2. Volcano plots visualizing the vibration of effects (VoE) for four examples, (A) serum vitamin D, (B) serum thyroxine, (C) urinary creatinine, (D) serum a-tocopherol. Two-dimensional histogram representation in upper panel and contour scatter plot is in lower panel. All effects are for a 1SD change in logged level of variable interest."
[patel2015, fig2, ref1]

"4.2. Prototypical patterns of the VoE: We describe four prototypical patterns from the set of 417 variables (Fig. 2, see Appendix at www.jclinepi.com for all 417 variables). The first pattern is exemplified by the association between serum levels of vitamin D and mortality (Fig. 2A). All the HR estimates are <1.00, indicating that higher levels of vitamin D tend to be associated with longer survival (all HR <0.76); however, the magnitude of the estimated effect is dependent on the number of adjustment variables, and the association is attenuated when adjusting for more variables, from HR = 0.64 with no adjustment (k = 0) to 0.75 with all 13 adjustment variables included (k = 13). In contrast, the P-values are less than the nominal level of statistical significance (P = 0.05, black line). Most of the results are centered on HR ~ 0.72 and P ~ 10^-4 (two-dimensional mode). In this first pattern, one concludes that although adjustment weakens the magnitude relationship between vitamin D levels and mortality, inferences regarding the relationship are similar throughout all scenarios of adjustment. Of the 417 variables, 53 (13%) exhibited similar behavior to vitamin D, where all associations were beyond the level of nominal statistical significance, but the association was attenuated with a greater number of adjustment variables (see Fig. S1/Appendix at www.jclinepi.com)."
[patel2015, fig2, ref2]

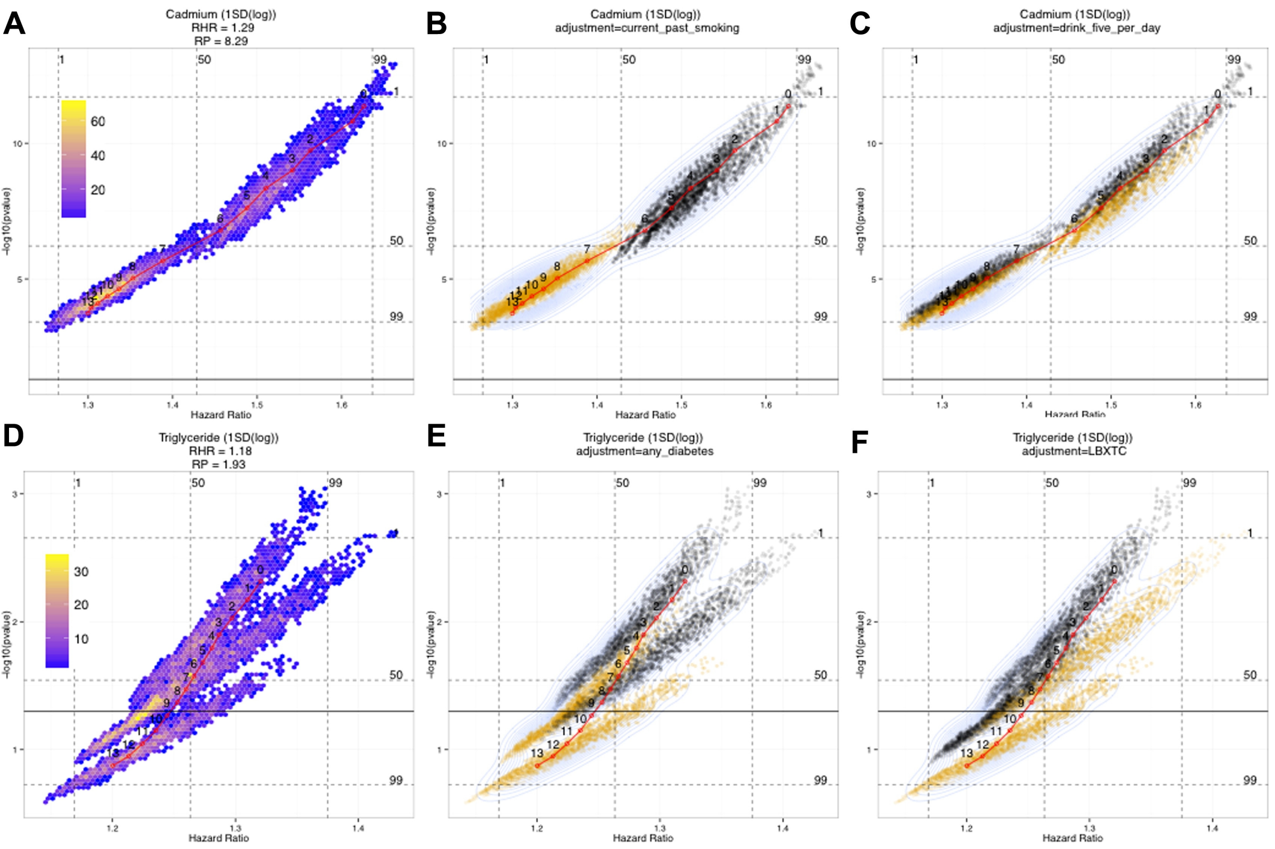
"The second pattern is exemplified by the relationship between thyroxine and mortality, displays how increasing adjustment might change inference (Fig. 2B). Higher thyroxine levels tend to be associated with longer survival, but P-values become greater than the nominal level of statistical significance (P = 0.05) with nine or more adjustment variables on average. Of the 417 variables, 91 (22%) variables had similar behavior to thyroxine in which HR were attenuated and the P-values rose above the nominal level of significance (P > 0.05) as the number of adjusting variables, k, increased (see Fig. S1 and Table S3/ Appendix at www.jclinepi.com)."
[patel2015, fig2, ref3]

"The third pattern, as exemplified by an indicator of kidney function, urinary creatinine, and mortality, shows an opposite trend (Fig. 2C). For k = 5-13 number of adjustment variables, the association tends to become stronger in HR and statistical significance; however, the trend is less clear for k = 0-4, where the median P-values increase. Twenty-six (6%) of the 417 variables exhibited similar behavior to urinary creatinine where the effect sizes increased and P-values decreased for larger k."
[patel2015, fig2, ref4]

"In the last pattern, as exemplified by a-tocopherol (Fig. 2D), the estimated HRs can be both greater and less than the null value (HR > 1 and HR < 1) depending on what adjustments were made. We call this the Janus effect after the two-headed representation of the ancient Roman god. For a-tocopherol, most of the HR and P-values were concentrated around 1 and nonsignificance, respectively. However, 1% of the models had an HR < 0.875 (12.5% decreased risk of death for 1SD increase in exposure) with a nominally significant P-value (P < 0.05), whereas another 1% of the models had HR > 1.05 (5% increased risk for death for 1SD increase of exposure), albeit without reaching nominal significance. The Janus effect is common: 131 (31%) of the 417 variables had their 99th percentile HR > 1 and their 1st percentile HR < 1."
[patel2015, fig2, ref5]

"Examples such as those in Fig. 2A-Drepresented theVoE patterns for 72% of the 417 associations. Other patterns included VoE where all P-values were >0.05 and the strength of the association decreased (n = 50, 12%), increased (n = 27, 6%), or showed no dependence (n = 15, 4%) with increasing number of adjustment variables k (see Table S3/Appendix at www.jclinepi.com). Rarer patterns included variables where all P-values were <0.05 and there was an increasing strength of association (n = 5, 1%) or no clear relationship with increasing k (n = 4, 1%), and those having P-values with a range less than and greater than 0.05 with no clear relationship with k (n = 15, 4%)."
[patel2015, fig2, ref6]

patel2015 - fig3



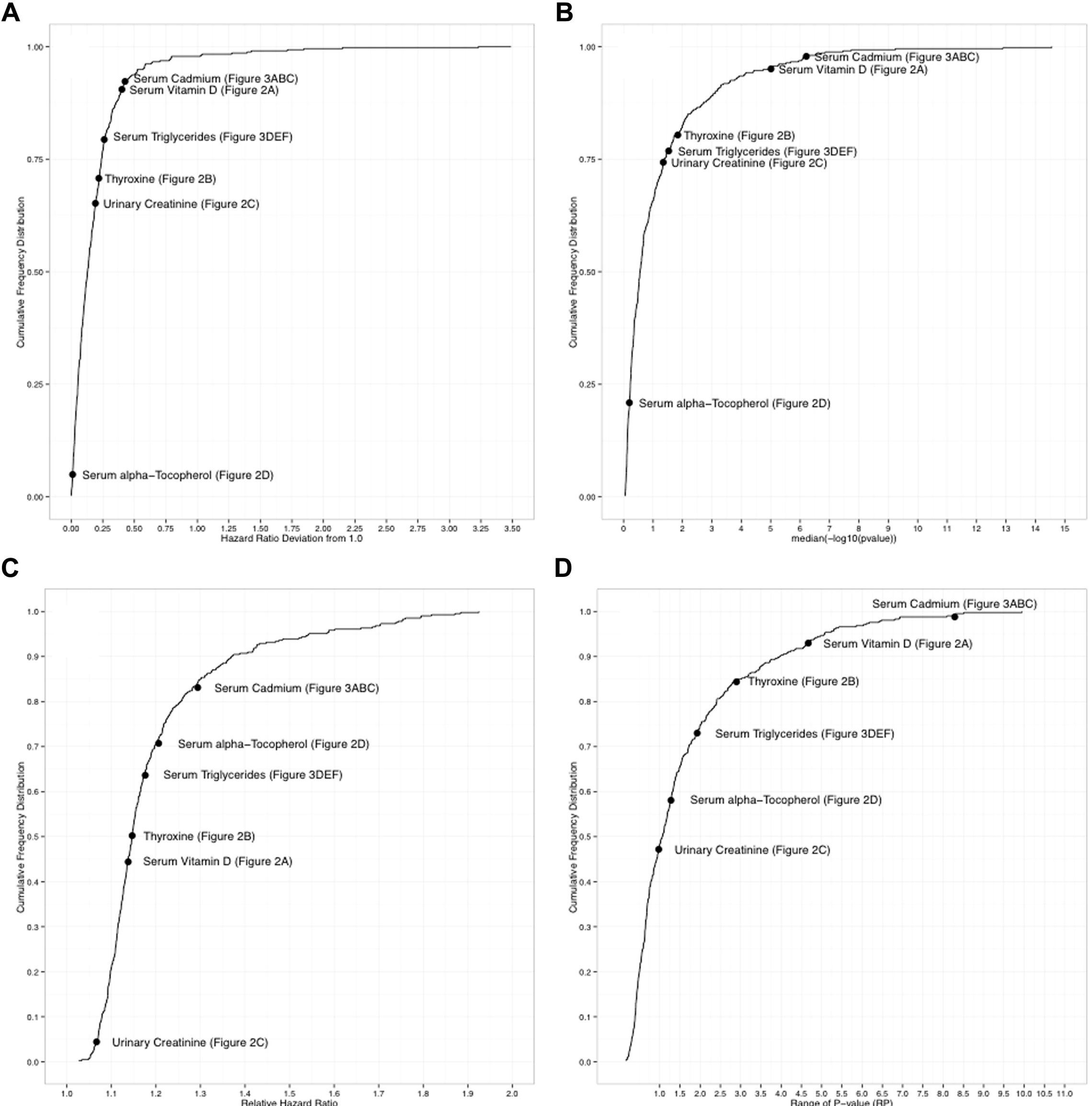
"Fig. 3. Volcano plots visualizing vibration of effects (VoE) for three examples with multiple "modes." (A) The 2D histogram for 1SD increase of the logarithm of serum cadmium, (B) volcano scatter plot with of serum cadmium if smoking was included in the model (yellow) or smoking not included in model (black). (C) Volcano scatter plot for serum cadmium models with drink five per day (yellow) or models without drink five per day (black). (D) The 2D histogram for 1SD increase of the logarithm of serum triglycerides, (E) with total cholesterol included in the model (yellow) or total cholesterol not included in model (black). (F) With any diabetes (yellow) or models without any diabetes (black)."
[patel2015, fig3, ref1]

"4.3. Identifying "multimodality of effects" with VoE: By empirically estimating the VoE, it is also possible to detect whether one or more adjustment variables make a marked difference in the results, leading to multiple modes (Fig. 3) which we call multimodality of effects. Multimodality of effects was clearly seen in 71 of the 417 (17%) assessed variables. For example, the overall VoE for serum cadmium on mortality indicates strong association with mortality (Fig. 3A); all of the HRs are >1.2 per 1 SD change in serum cadmium levels, and P-values in all analytical scenarios are <0.05. However, two modes are visually evident (Fig. 3A)."
[patel2015, fig3, ref2]

"To identify the key variable(s) that separated these different distributions, we visualized the VoE by coloring each point on whether it included (or did not include) each one of the 13 adjustment variables in the model, leading to 13 separate visualizations. In serum cadmium, we observed the two distinct modes were indicative of models that contained or did not contain current or past smoking (Fig. 3B). Specifically, models that contained the smoking adjustment variable (Fig. 3B, yellow points) had HR lower than the models without the smoking adjustment and lower -log10(P-values) (Fig. 3B, black points). One source of cadmium exposure includes smoking, and we concluded that the correlation between smoking and exposure to cadmium might be driving the multimodal behavior of VoE. Furthermore, we observed that models that included (or did not include) alcohol drinking also resulted in separate modes in P-values (Fig. 3C)."
[patel2015, fig3, ref3]

"We observed three modes in the association between triglyceride levels and mortality (Fig. 3D-F). The multimodal plots indicated that total cholesterol and diabetes were driving these modes. For example, in models that did not contain these two adjustments, the associations had smaller P-values and a smaller range of HR. Furthermore, in models containing diabetes, HR were attenuated. The multimodal pattern seems reasonable in light of the high correlation between triglyceride levels and total cholesterol levels/risk for diabetes. We observed a similar pattern for other cardiometabolic indicators, including fasting blood glucose and insulin (see Fig. S1/Appendix at www.jclinepi.com)."
[patel2015, fig3, ref4]

patel2015 - fig4



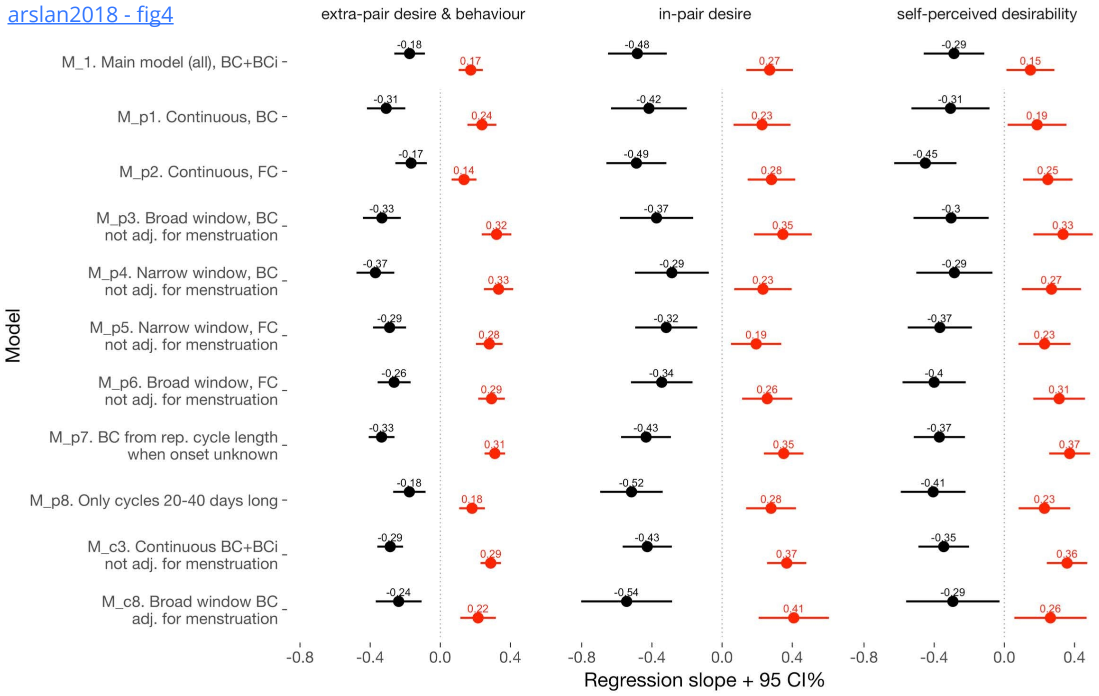
"Fig. 4. Cumulative distributions of vibration of effects (VoE) for 417 variables. (A) Absolute deviation of HR from 1, (B) log10(P-value), (C) relative hazard ratio (RHR), (D) relative P-value (RP). Examples shown in Figures 1-3 are shown in the distribution."
[patel2015, fig4, ref1]

"4.4. Summary of common patterns of the VoE: Figure 4 shows the distribution of the fold deviation of HR from the null (HR = 1.00), the -log10(P-value), RHR, and RP for all 417 variables considered. The “fold deviation” is the difference of the median VoE-estimated HR from 1 (the null value). The median fold deviation was 1.13-fold (25th percentile: 1.05-fold, 75th percentile: 1.24-fold, Fig. 4A). Moreover, 50% of the variables had a median P-value less than or greater than 0.27 (25th percentile: 0.04, 75th percentile: 0.59, Fig. 4B). The median RHR was 1.15 (5th percentile: 1.07, 25th percentile: 1.11, 75th percentile: 1.22, 95th percentile: 1.70, Fig. 4C). The median RP was 1.07 (5th percentile: 0.31, 25th percentile: 0.589, 75th percentile: 2.03, 95th percentile: 5.09). We observed that most associations could vary by at least 1.15-fold in the magnitude of the HR and by one order of magnitude [log10(P-value)] in the level of statistical significance, and much larger changes were not uncommon. We observed a weak correlation between RHR and RP (see Fig. S2/ Appendix at www.jclinepi.com, p = 0.09, P = 0.06)."
[patel2015, fig4, ref2]

"Returning to the prototypical examples that we discussed previously, the RHR for vitamin D and thyroxine was moderate 1.14 (44th percentile) and 1.15 (51st percentile; Figs. 4C, 2A, and B). However, their RPs were among the largest and equal to 4.7 (93rd percentile) and 2.90 (84th percentile), respectively (Figs. 4D, 2A, and B). For urinary creatinine, the scenarios of adjustment had less prominent VoE. The RHR and RP for urinary creatinine was 1.07 (5th percentile) and 0.98 (47th percentile; Fig. 4C and D)."
[patel2015, fig4, ref3]

"The RHR for a-tocopherol (with the Janus effect) was higher (1.21, 71st percentile, Fig. 4C). Variables that demonstrated multimodality, such as serum cadmium and triglycerides, tended to have larger VoE. For example, serum cadmium had an RHR of 1.29 (82nd percentile) and one of the highest RPs, 8.29 (99th percentile). Serum triglycerides had an RHR of 1.18 (64th percentile) and an RP of 1.93 (73rd percentile)."
[patel2015, fig4, ref4]

arslan2018 - fig4

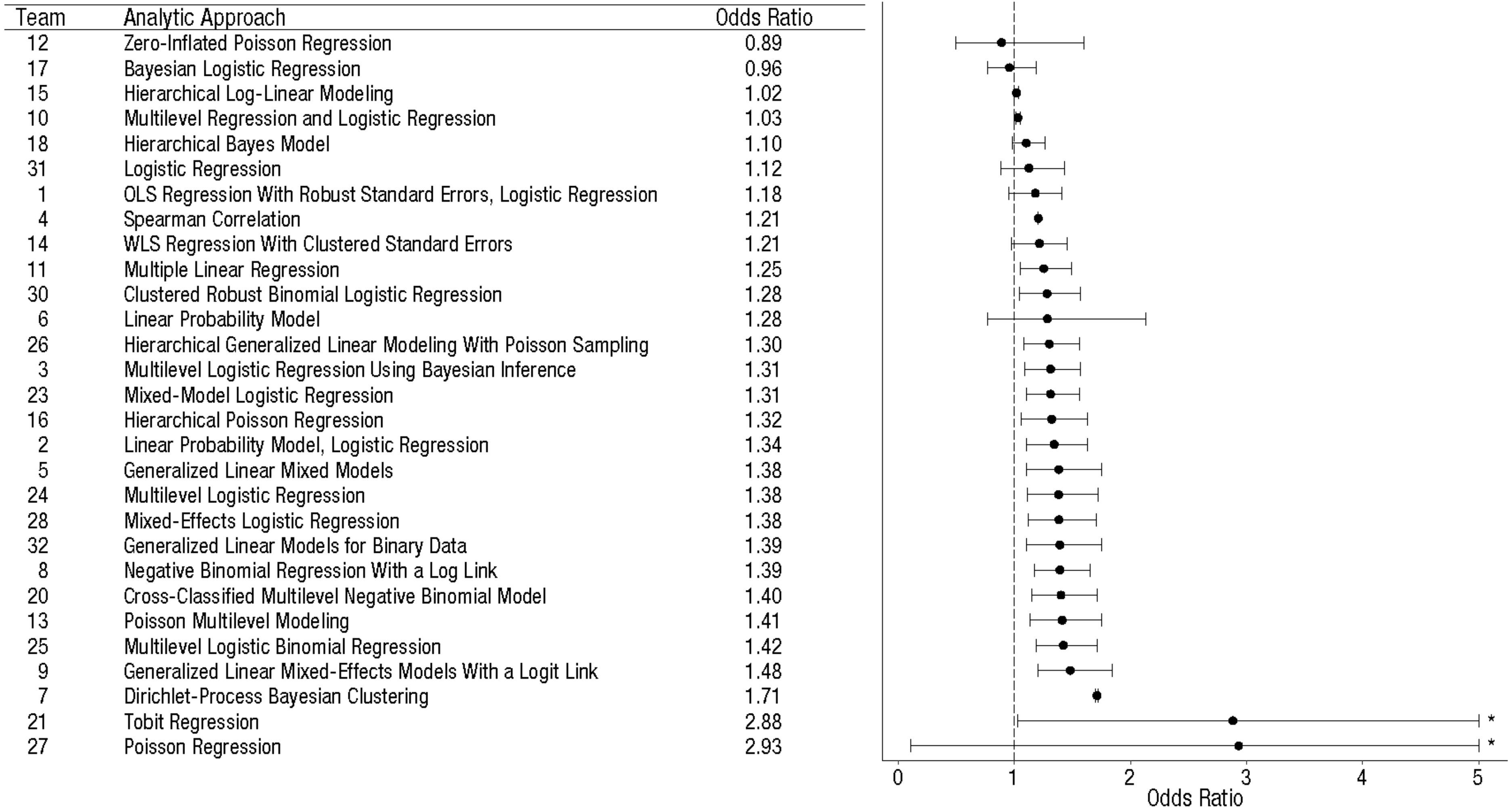


"Figure 4. Robustness checks for predictors. Coefficient plot showing a consistent effect of the fertility predictor among naturally cycling women (red) but not hormonal contraception users (black) across several predictor and model specifications (explained in further detail in the text). FC = forward-counted from last menstrual onset, BC = backward-counted from observed next menstrual onset, BCi = backward-counted from inferred next menstrual onset."
[arslan2018, fig4, ref1]

"In models M_p1 to M_p11, we tested different estimates of the fertile window as our predictor to address the concerns about varying standards described in Methodological issues. We compared all combinations of a narrow window, broad window, continuous estimates, and backward- and forwardcounting. When we used a continuous fertile window predictor, we also adjusted for premenstrual and menstrual days. We found that including adjustments for menstruation and pre-menstruation (M_c3) reduced effect sizes for the fertile window predictor. We could not always adjust for (pre-)menstruation when using a narrow window predictor because of model convergence problems. After taking this into account, we found no systematic pattern in which certain predictors (narrow or broad window, forward or backward counted) had larger effect sizes than others across outcomes (see Figure 4). However, continuous curves over backward-counted days (Figure 3) matched the predicted pattern more closely than curves over forward-counted days (see supportive website, osf.io/pbef2)."
[arslan2018, fig4, ref2]

"Although it is difficult to compute an equivalent of Cohen's d for multilevel models, our comparable effect size estimates ranged from 0.12 to 0.43. These effect sizes are disattenuated for measurement error in the predictor, but not in the outcome. Some were hence only a quarter of the smallest effect size (0.4) considered in Gangestad et al.'s (2016) simulations and sample size recommendations. Empirically, had we used sample sizes like the studies we were replicating, none of the effects reported here would have been significant. Whether the fertility predictor was formed based on forward- or backward-counting, narrow, broad, or continuous fertile phases seemed to make less of a difference (Figure 4), except that predictors using more data are preferable and that (pre-)menstruation should be adjusted for. While the absolute sizes of the effects we found were not huge, their practical implications might still be noteworthy. The effects on in-pair desire are, for instance, comparable with reported effects of hormonal contraceptive use on sexual desire in a randomised controlled trial (Zethraeus et al., 2016). Moreover, we found evidence for substantial inter-individual variation, so that effects that are small on average might be substantial for some women."
[arslan2018, fig4, ref3]

silberzahn2017 - fig2

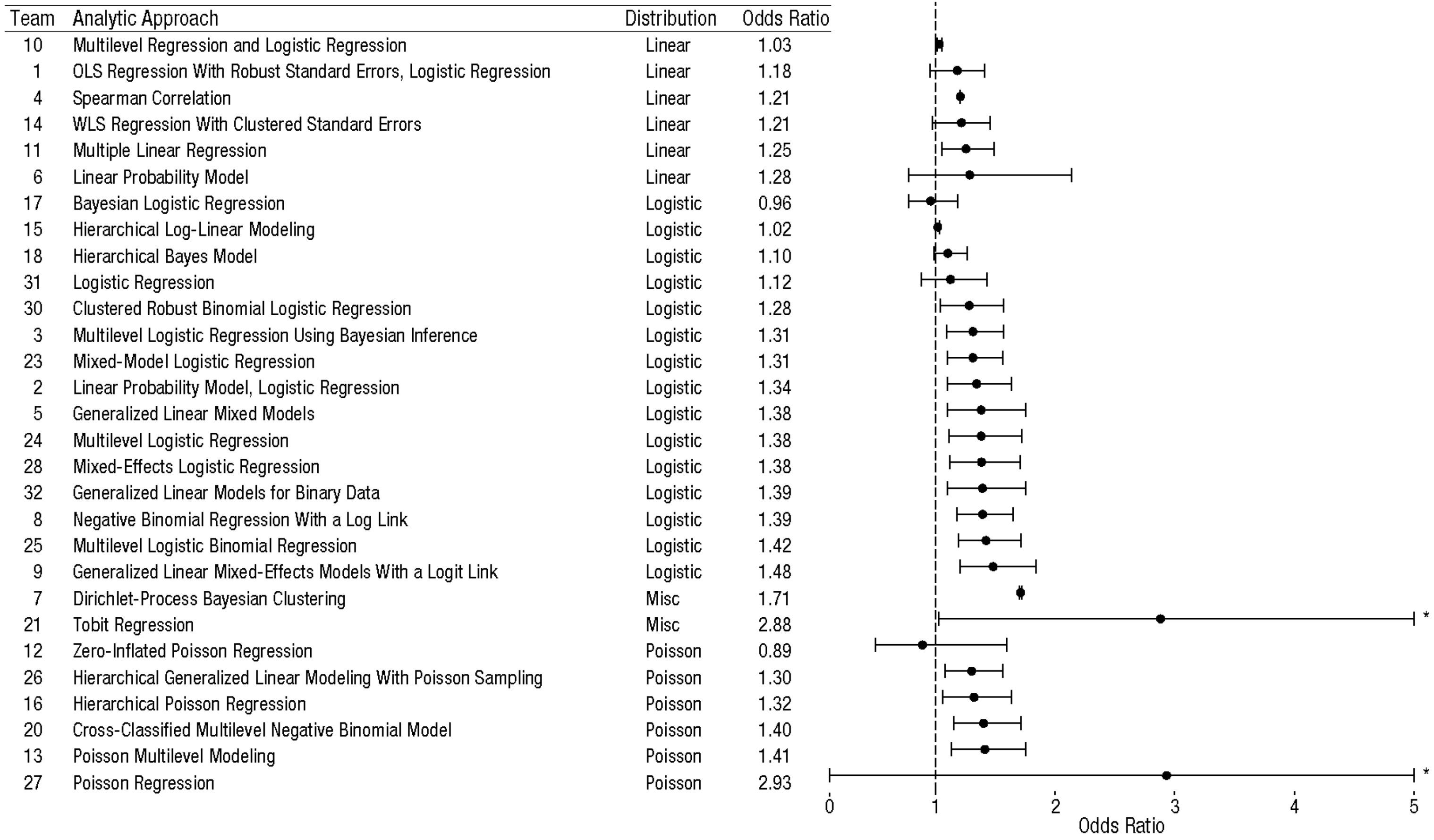


Team	Analytic Approach	Odds Ratio
12	Zero-Inflated Poisson Regression	0.89
17	Bayesian Logistic Regression	0.96
15	Hierarchical Log-Linear Modeling	1.02
10	Multilevel Regression and Logistic Regression	1.03
18	Hierarchical Bayes Model	1.10
31	Logistic Regression	1.12
1	OLS Regression With Robust Standard Errors, Logistic Regression	1.18
4	Spearman Correlation	1.21
14	WLS Regression With Clustered Standard Errors	1.21
11	Multiple Linear Regression	1.25
30	Clustered Robust Binomial Logistic Regression	1.28
6	Linear Probability Model	1.28
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
3	Multilevel Logistic Regression Using Bayesian Inference	1.31
23	Mixed-Model Logistic Regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear Probability Model, Logistic Regression	1.34
5	Generalized Linear Mixed Models	1.38
24	Multilevel Logistic Regression	1.38
28	Mixed-Effects Logistic Regression	1.38
32	Generalized Linear Models for Binary Data	1.39
8	Negative Binomial Regression With a Log Link	1.39
20	Cross-Classified Multilevel Negative Binomial Model	1.40
13	Poisson Multilevel Modeling	1.41
25	Multilevel Logistic Binomial Regression	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	1.48
7	Dirichlet-Process Bayesian Clustering	1.71
21	Tobit Regression	2.88
27	Poisson Regression	2.93

"Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that the smallest reported effect size is at the top and the largest is at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot; the actual upper bounds of the confidence intervals were 11.47 for Team 21 and 78.66 for Team 27. OLS = ordinary least squares; WLS = weighted least squares."
[silberzahn2017, fig2, ref1]

"What were the consequences of this variability in analytic approaches? Figure 2 shows each team's estimated effect size, along with its 95% confidence interval (CI). As this figure and Table 3 show, the estimated effect sizes ranged from 0.89 (slightly negative) to 2.93 (moderately positive) in odds-ratio (OR) units; the median estimate was 1.31. The confidence intervals for many of the estimates overlap, which is expected because they are based on the same data. Twenty teams (69%) found a significant positive relationship, p < .05, and nine teams (31%) found a nonsignificant relationship. No team reported a significant negative relationship."
[silberzahn2017, fig2, ref2]

silberzahn2017 - fig3



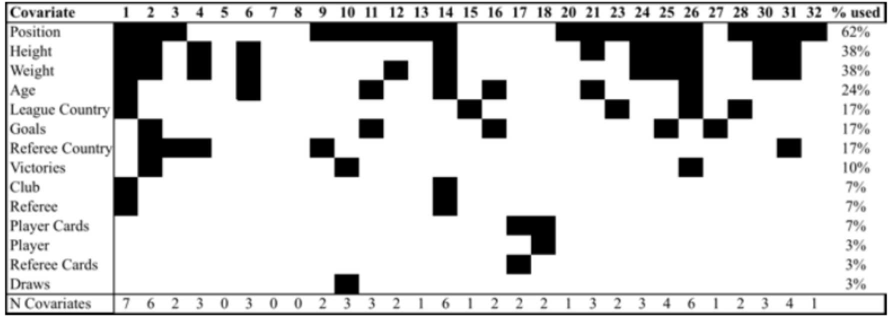
Team	Analytic Approach	Distribution	Odds Ratio
10	Multilevel Regression and Logistic Regression	Linear	1.03
1	OLS Regression With Robust Standard Errors, Logistic Regression	Linear	1.18
4	Spearman Correlation	Linear	1.21
14	WLS Regression With Clustered Standard Errors	Linear	1.21
11	Multiple Linear Regression	Linear	1.25
6	Linear Probability Model	Linear	1.28
17	Bayesian Logistic Regression	Logistic	0.96
15	Hierarchical Log-Linear Modeling	Logistic	1.02
18	Hierarchical Bayes Model	Logistic	1.10
31	Logistic Regression	Logistic	1.12
30	Clustered Robust Binomial Logistic Regression	Logistic	1.28
3	Multilevel Logistic Regression Using Bayesian Inference	Logistic	1.31
23	Mixed-Model Logistic Regression	Logistic	1.31
2	Linear Probability Model, Logistic Regression	Logistic	1.34
5	Generalized Linear Mixed Models	Logistic	1.38
24	Multilevel Logistic Regression	Logistic	1.38
28	Mixed-Effects Logistic Regression	Logistic	1.38
32	Generalized Linear Models for Binary Data	Logistic	1.39
8	Negative Binomial Regression With a Log Link	Logistic	1.39
25	Multilevel Logistic Binomial Regression	Logistic	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	Logistic	1.48
7	Dirichlet-Process Bayesian Clustering	Misc	1.71
21	Tobit Regression	Misc	2.88
12	Zero-Inflated Poisson Regression	Poisson	0.89
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	Poisson	1.30
16	Hierarchical Poisson Regression	Poisson	1.32
20	Cross-Classified Multilevel Negative Binomial Model	Poisson	1.40
13	Poisson Multilevel Modeling	Poisson	1.41
27	Poisson Regression	Poisson	2.93

"Fig. 3. Point estimates (clustered by analytic approach) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are clustered according to the distribution used in their analyses; within each cluster, the teams are listed in order of the magnitude of the reported effect size, from smallest at the top to largest at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot (see Fig. 2). OLS = ordinary least squares; WLS = weighted least squares; Misc = miscellaneous. "
[silberzahn2017, fig3, ref1]

"The teams also varied in their approaches to handling the nonindependence of players and referees, and this variability also influenced both median estimates of the effect size and the rates of significant results. In total, 15 teams estimated a fixed effect or variance component for players, referees, or both; 12 of these teams reported significant effects (median OR = 1.32, MAD = 0.12). Eight teams used clustered standard errors, and 4 of these teams found significant effects (median OR = 1.28, MAD = 0.13). An additional 5 teams did not account for this artifact, and 4 of these teams reported significant effects (median OR = 1.39, MAD = 0.28). The remaining team used fixed effects for the referee variable and reported a nonsignificant result (OR = 0.89)."
[silberzahn2017, fig3, ref3]

"What were the results obtained with the different types of analytic approaches used? Teams that employed logistic or Poisson models tended to report estimates that were larger than those of teams that used linear models (see the effect sizes in Fig. 3, in which the teams are clustered according to the distribution used for analyses). Fifteen teams used logistic models, and 11 of these teams found a significant effect (median OR = 1.34; median absolution deviation, or MAD = 0.07). Six teams used Poisson models, and 4 of these teams found a significant effect (median OR = 1.36, MAD = 0.08). Of the 6 teams that used linear models, 3 found a significant effect (median OR = 1.21, MAD = 0.05). The final 2 teams used models classified as miscellaneous, and both of these teams reported significant effects (ORs = 1.71 and 2.88, respectively)."
[silberzahn2017, fig3, ref2]

silberzahn2017 - tab4



Covariate	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	20	21	23	24	25	26	27	28	30	31	32	% used
Position	■	■	■						■	■	■	■	■	■					■	■	■	■	■	■		■	■	■	■	62%
Height	■	■		■		■								■						■		■	■	■			■	■		38%
Weight	■	■		■		■						■		■								■	■	■			■	■		38%
Age	■					■					■			■		■				■				■						24%
League Country	■														■						■			■		■				17%
Goals		■									■					■							■		■					17%
Referee Country		■	■	■					■																			■		17%
Victories		■								■														■						10%
Club	■													■																7%
Referee	■													■																7%
Player Cards																	■	■												7%
Player																		■												3%
Referee Cards																	■													3%
Draws										■																				3%
N Covariates	7	6	2	3	0	3	0	0	2	3	3	2	1	6	1	2	2	2	1	3	2	3	4	6	1	2	3	4	1	

Table 4. Covariates used by each team. Team numbers are listed on the top and covariates on the left. A shaded box indicates that the corresponding team used the covariate in their final model. The table is ordered by the frequency by which each covariate was used.

Table 4. Covariates used by each team. Team numbers are listed on the top and covariates on the left. A shaded box indicates that the corresponding team used the covariate in their final model. The table is ordered by the frequency by which each covariate was used.
[silberzahn2017, tab4, ref1]

Twenty-nine independent teams of researchers submitted analytical approaches and refined these throughout the crowdsourcing project. Table 2 shows each team's final analytic technique, model specifications and reported effect size.3 Analytic techniques ranged from simple linear regression to complex multilevel regression and Bayesian approaches. Teams also varied highly in their decisions regarding which covariates to include (see R7.1). Table 4 shows that the 29 teams used 21 unique combinations of covariates. Apart from the variable 'games', which was used by all teams, just one covariate (player position, 62%) was used in more than half of the analytic strategies and three were used in just one analysis. Two sets of covariates were used by three teams each, and four sets of covariates were used by two teams each. The remaining 15 teams used a unique combination of covariates.
[silberzahn2017, tab4, ref2]

dragicevic2019 - fig2

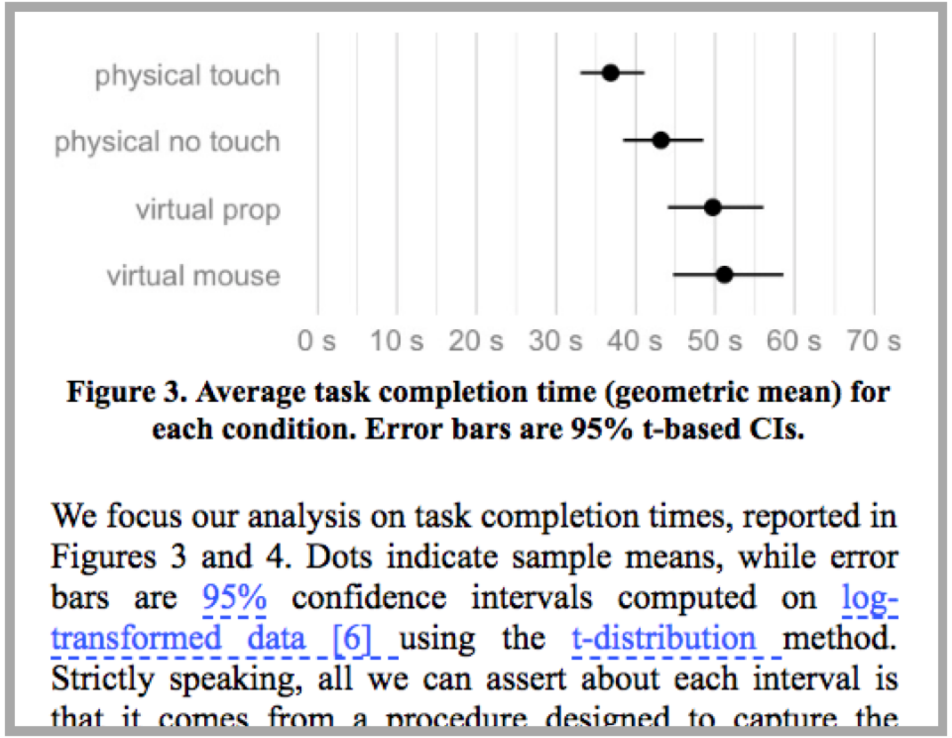


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data [6] using the t-distribution method. Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the

"Figure 2: Excerpt from the mini-paper Freqentist, showing widgets embedded in the text in Bret Victor's [94] style. Operating a widget changes one aspect of the analysis and immediately updates the figure."
[dragicevic2019, fig2, ref1]

"The Freqentist example [36] is a reanalysis of a CHI study evaluating physical visualizations [51]. It is meant to illustrate a few basic multiverse analysis ideas for a typical frequentist analysis with confidence intervals (CIs). The results of the analysis are initially identical to the original paper, including the two figures reporting mean task completion time per technique and pairwise comparisons, with 95% CIs. Four aspects of the analysis can be changed by the reader, which has the effect of immediately updating the two plots and some text elements such as explanations and figure captions. Changes are made by clicking or dragging the elements of the text in blue as in Bret Victor's explorable explanations [94] (see Figure 2)."
[dragicevic2019, fig2, ref2]

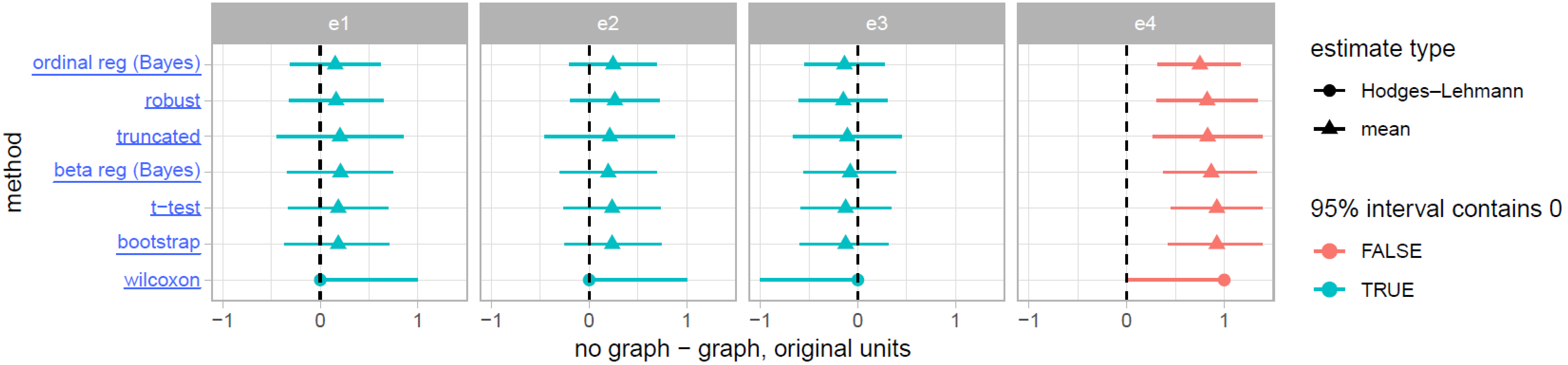
"First, horizontally dragging the "95%" text has the effect of changing the confidence level (7 levels are provided from 50% to 99.9%) and updating the length of error bars in the two figures. This allows the reader to appreciate that the 95% level is arbitrary [66] and thus that CIs should not be interpreted in a strictly dichotomous manner [29]. Meanwhile, readers who insist on interpreting effects as significant or non-significant have the option of changing the customary cutoff of α=.05 (95% CIs), for example to the α=.005 (99.5% CIs) criterion now advocated by some methodologists [15]."
[dragicevic2019, fig2, ref3]

"Clicking the "transformed data" text toggles the text to "untransformed data" and updates the two figures with results from the corresponding analysis. Although some researchers recommend that completion times be log-transformed [79], other researchers may be suspicious of, or unfamiliar with data transformations—this option reassures them that the results hold for untransformed data. Similarly, clicking on "tdistribution" switches the text to "BCa bootstrap" and shows the results of the analysis using non-parametric bootstrap CIs, which tend to be liberal (i.e., too narrow) with small samples but do not require distributional assumptions [59]."
[dragicevic2019, fig2, ref4]

"Finally, the plot with the three planned pairwise comparisons (not shown in Figure 2) shows uncorrected CIs, but the reader can apply a Bonferroni correction by clicking on the text "not corrected for multiplicity". Correction for multiplicity is strongly recommended by many but it is not without drawbacks: there is a controversial and complex literature on the topic [31]. To help the reader interpret the CIs correctly, the mini-paper contains a paragraph that gives the individual and the family-wise CI coverage and false positive rates, which are updated whenever Bonferroni correction is turned on or off, or whenever the confidence level is changed. More details can be found in the mini-paper itself [36]."
[dragicevic2019, fig2, ref5]

"The Freqentist mini-paper covers a total of 7×2×2×2 = 56 unique analyses. The paper concludes that the findings from the original study (i.e., good evidence of a difference for the first two comparisons, inconclusive results for the third one) are reasonably robust, as they hold across the sub-multiverse where the confidence level is at 95% or less."
[dragicevic2019, fig2, ref6]

dragicevic2019 - fig3

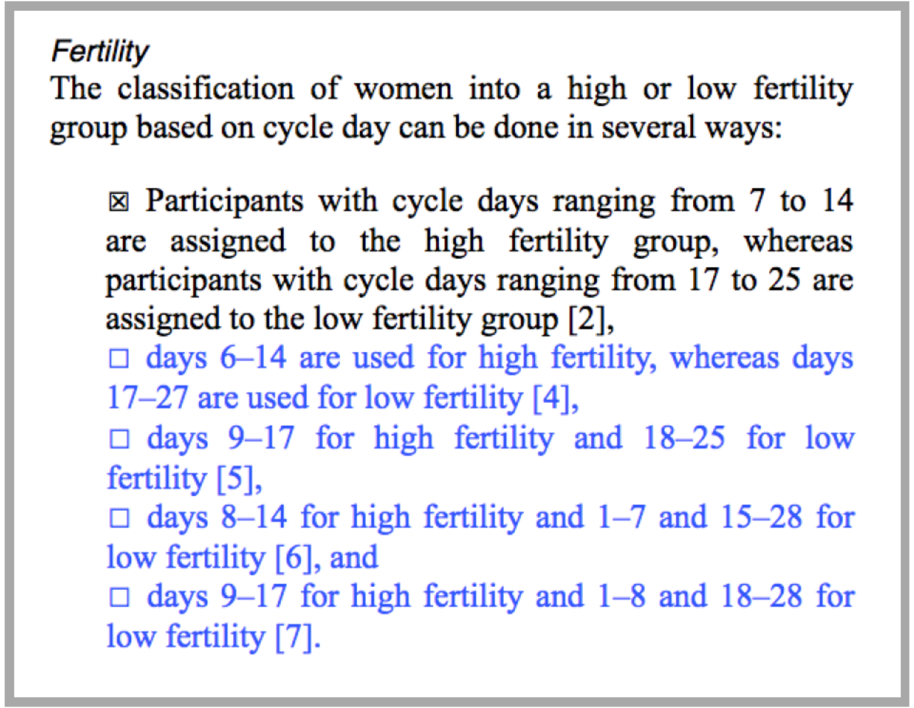


"Figure 3: Plot from the mini-paper Likert, summarizing point estimates and 95% CIs for an effect measured across 4 different experiments (columns) and analyzed using 9 different methods (rows). Clicking on a row label updates the method section. Here no matter how the data are analyzed, no conclusive effect is found for the first three experiments (blue intervals), while there is convincing evidence for an effect in the fourth (red intervals)."
[dragicevic2019, fig3, ref1]

"The Likert mini-paper reanalyzes the four experiments in the original InfoVis study [35] using nine different methods covering a broad range of approaches, including parametric vs. non-parametric and frequentist vs. Bayesian. In contrast with the previous mini-paper, all analysis outcomes are summarized in a static overview figure to facilitate comparison. Seven of the nine methods yield simple effect sizes (e.g., mean differences) which are summarized in the plot shown in Figure 3, while the remaining two methods yield log-odds ratios, reported in a different plot (not shown here). By default, the method section in the mini-paper only details the bootstrap method, which was used in the original study. However, clicking on a row label in the figure changes the method section to provide a description and justification of the selected method, an interpretation of its results, and the p-value for the fourth experiment (when available)."
[dragicevic2019, fig3, ref2]

"The Likert mini-paper covers a total of 9 unique analyses. It concludes that the results are consistent across analyses: no matter how the Likert data are analyzed, no conclusive effect is found for the first three experiments (blue intervals in Figure 3), while there is convincing evidence for an effect in the fourth (red intervals). The results differ slightly nevertheless, and the reader can observe which types of analysis are more conservative and which ones are more liberal."
[dragicevic2019, fig3, ref3]

dragicevic2019 - fig4



Fertility

The classification of women into a high or low fertility group based on cycle day can be done in several ways:

☒ Participants with cycle days ranging from 7 to 14 are assigned to the high fertility group, whereas participants with cycle days ranging from 17 to 25 are assigned to the low fertility group [2],

□ days 6–14 are used for high fertility, whereas days 17–27 are used for low fertility [4],

□ days 9–17 for high fertility and 18–25 for low fertility [5],

□ days 8–14 for high fertility and 1–7 and 15–28 for low fertility [6], and

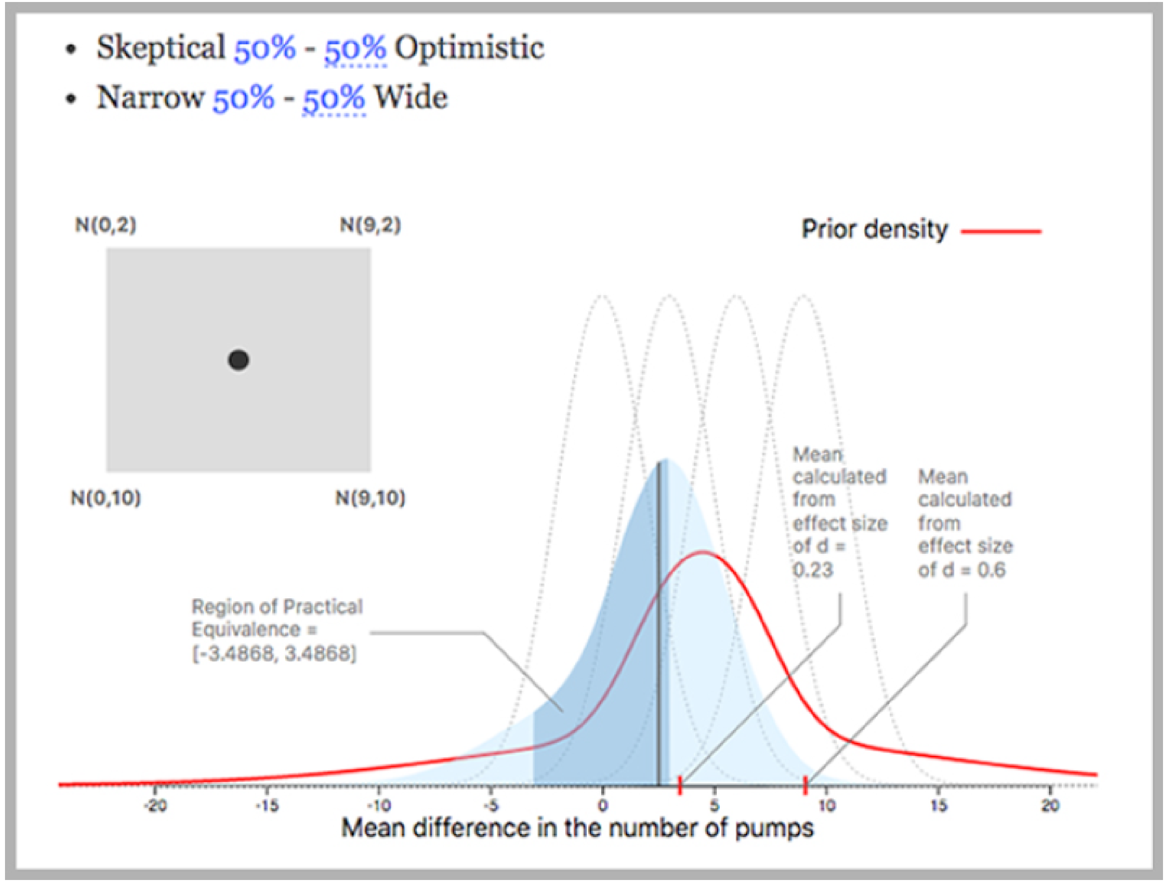
□ days 9–17 for high fertility and 1–8 and 18–28 for low fertility [7].

"Figure 4: Excerpt from the mini-paper Dataverse, listing five different ways of dichotomizing a dependent variable. Elsewhere in the mini-paper, an interaction plot gets updated each time an option is chosen." [dragicevic2019, fig4, ref1]

"The “Constructing the data multiverse” section in Steegen et al. [87] goes through each data processing choice made in the original study [38] and describes alternative choices that could have been reasonably made. The Dataverse minipaper essentially reproduces this section with the difference that the reader can select particular choices. The mini-paper first lists five ways of dichotomizing a particular dependent variable, and lets the reader choose one of them (Figure 4). Four other data processing operations are described afterwards, each with two to three options to choose from. The mini-paper ends with a figure showing the result of the selected analysis in the form of an interaction plot, which is updated each time a different option is chosen in the text." [dragicevic2019, fig4, ref2]

"The Dataverse mini-paper covers 5×2×3×3×2 = 180 unique analyses. Steegen et al. [87] summarizes the multiverse by plotting the 180 corresponding p-values. While this summary provides an extremely useful overview clearly showing that the original findings are not robust, it does not allow the reader to examine detailed outcomes of specific analyses of interest. By making it possible to select any particular analysis and see the resulting effect sizes, the Dataverse mini-paper conveys more complete results than a simple summary of p-values. As in the Freqentist mini-paper the multiverse can be animated, giving a striking demonstration of the variability of effect sizes across the multiverse that can usefully complement the p-value summary." [dragicevic2019, fig4, ref3]

dragicevic2019 - fig5



- Skeptical 50% - 50% Optimistic
- Narrow 50% - 50% Wide

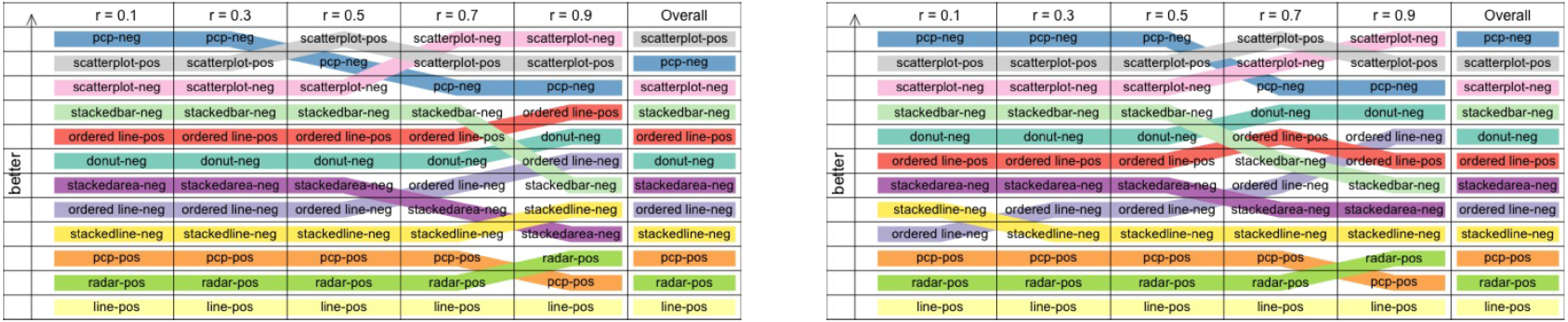
N(0,2)
N(9,2)
N(0,10)
N(9,10)
Prior density
Mean calculated from effect size of d = 0.23
Mean calculated from effect size of d = 0.6
Region of Practical Equivalence = [-3.4868, 3.4868]
-20
-15
-10
-5
0
5
10
15
20
Mean difference in the number of pumps

"Figure 5: Excerpt from the mini-paper Prior depicting the prior and posterior densities. Readers can use the 2D selection widget (left inset gray box) or drag the highlighted percentages to change the prior." [dragicevic2019, fig5, ref1]

"Unlike other examples, these two axes are continuous. The reader can change their prior either by clicking and dragging on a point in a 2-dimensional space (see Figure 5), or by clicking and dragging on text sliders (like how confidence level can be adjusted in the Freqentist mini-paper)." [dragicevic2019, fig5, ref2]

"In the browser, as users interact with Tangle widgets or our 2D widget (Figure 5) to move along the two dimensions (location and scale), we calculate the weights for the prior distributions and the corresponding weights for the posteriors using the above formula. We then calculate the mixture posterior density and visualize it using D3.js in real time." [dragicevic2019, fig5, ref3]

dragicevic2019 - fig6

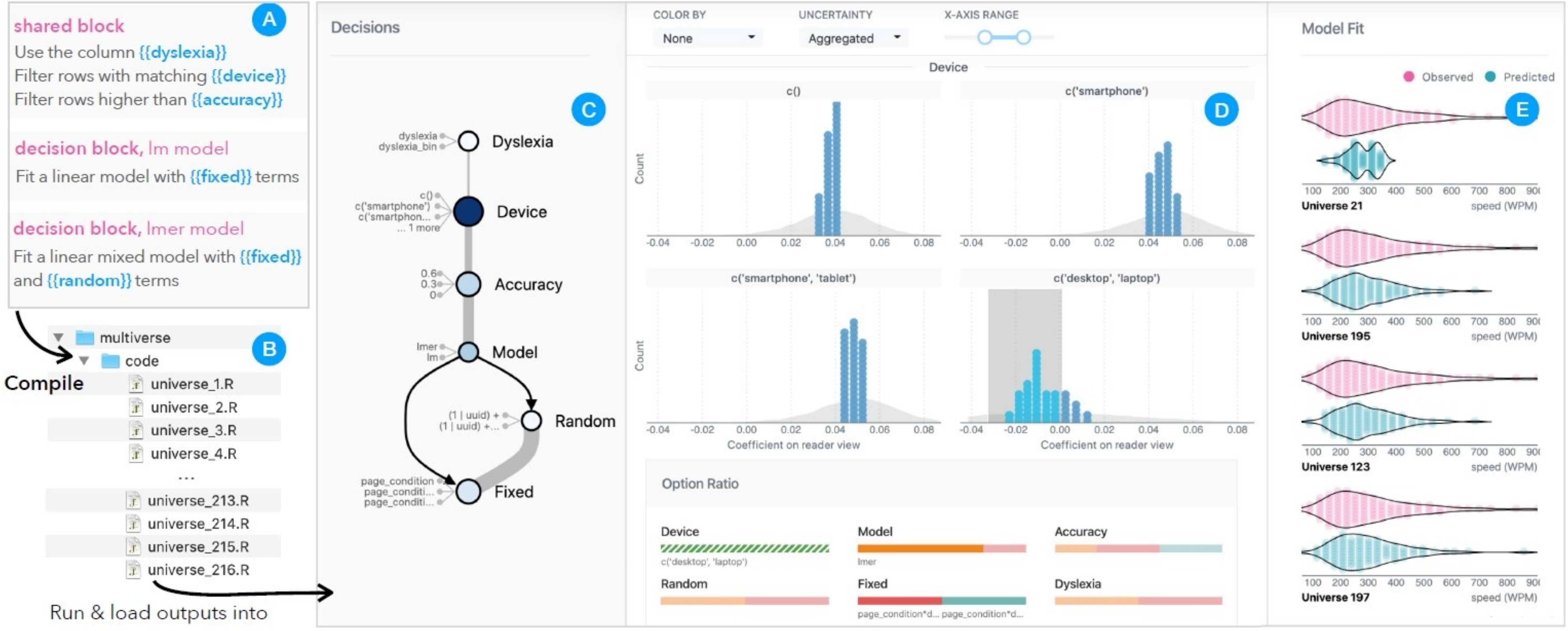


"Figure 6: Left: plot showing a ranking of visualizations in their ability to convey correlation [48]. Right: an alternative plot that could have reasonably come up in an exact replication, created by bootstrapping the experimental dataset. Some results hold (e.g., the bottom of the ranking) while some do not (e.g., the top and middle of the ranking). The mini-paper Dance allows to animate between 100 of those plots."
[dragicevic2019, fig6, ref1]

"The mini-paper reproduces the analysis from the original study, with its four plots. It also lets readers replace the original dataset with any of the 100 bootstrap datasets. When the dataset changes, each of the 4 plots changes slightly. More interestingly, animating the multiverse yields a “dance of plots” similar to Cumming’s dance of p-values [28] and other statistical dances [32], with the difference that the sampling distribution is estimated from data rather than simulated."
[dragicevic2019, fig6, ref2]

"Animating the multiverse of bootstrap datasets allows the reader to appreciate the reliability of the different quantities, trends and patterns depicted by each plot and to carry out “inference by eye” [30]: a pattern that is stable across the multiverse is a good indication that it is reliable. This is an example of the use of hypothetical outcome plots (HOPs) for conveying uncertainty [50, 53]. Compared to static representations of inferential information such as error bars, this technique has the advantage of being applicable to any plot. It is especially useful for revealing statistical uncertainty that is hidden in some plots, such as the ranking plot reproduced in Figure 6. More examples can be found in the mini-paper."
[dragicevic2019, fig6, ref3]

liu2020 - fig1



shared block
Use the column {{dyslexia}}
Filter rows with matching {{device}}
Filter rows higher than {{accuracy}}

decision block, lm model
Fit a linear model with {{fixed}} terms

decision block, lmer model
Fit a linear mixed model with {{fixed}} and {{random}} terms

Compile

Run & load outputs into

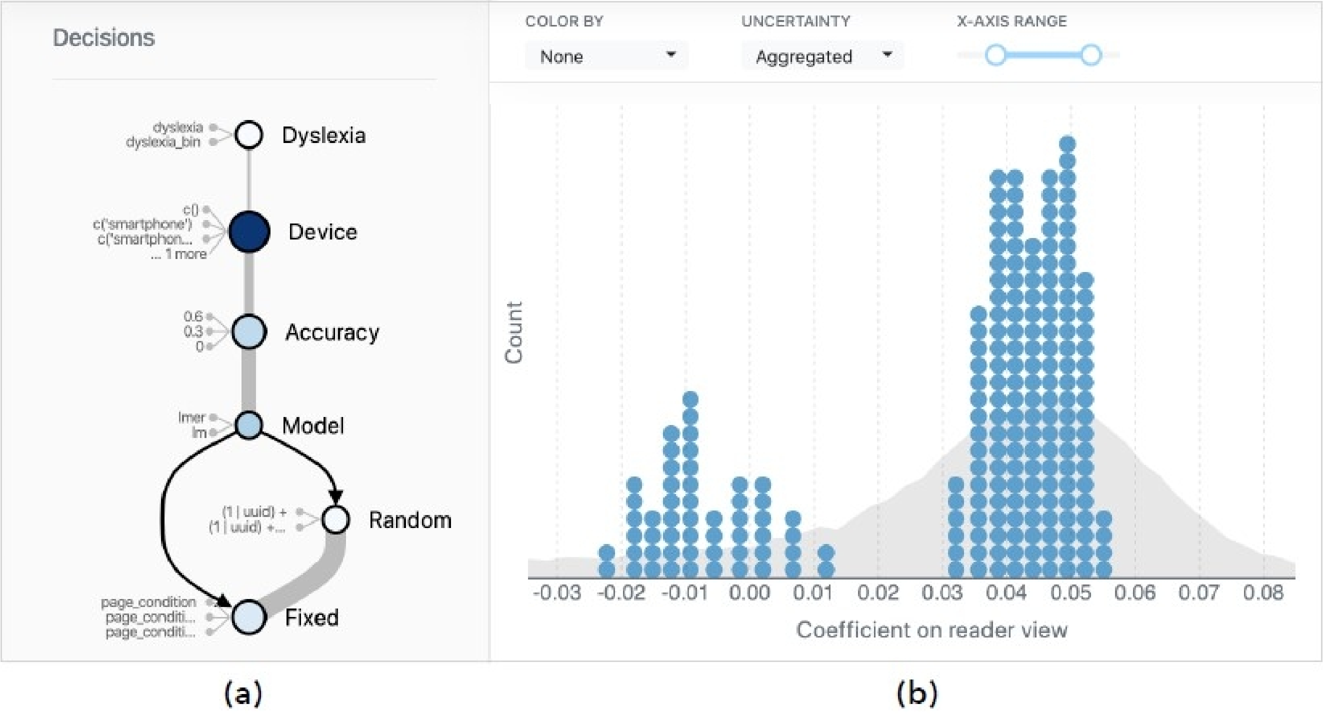
COLOR BY: None
UNCERTAINTY: Aggregated

Fig. 1. Authoring and visualizing multiverse analyses with Boba. Users start by annotating a script with analytic decisions (a), from which Boba synthesizes a multiplex of possible analysis variants (b). To interpret the results from all analyses, users start with a graph of analytic decisions (c), where sensitive decisions are highlighted in darker blues. Clicking a decision node allows users to compare point estimates (d, blue dots) and uncertainty distributions (d, gray area) between different alternatives. Users may further drill down to assess the fit quality of individual models (e) by comparing observed data (pink) with model predictions (teal).
[liu2020, fig1, ref1]

To further investigate model quality, Emma drills down to individual universes by clicking a dot in the outcome view. She sees in the model fit view (Fig. 1e) that a model gives largely mismatched predictions.
[liu2020, fig1, ref2]

Clicking a result in the outcome view populates the model fit view with visual predictive checks, which show how well predictions from a given model replicate the empirical distribution of observed data [14], allowing users to further assess model quality (T5). The model fit visualization juxtaposes violin plots of the observed data and model predictions to facilitate comparison of the two distributions (see Fig. 1e). Within the violin plots, we overlay observed and predicted data points as centered density dot plots to help reveal discrepancies in approximation due to kernel density estimation. When the number of observations is large (S1), we plot a representative subset of data, sampled at evenly spaced percentiles, as centered quantile dotplots [25]. As clicking individual universes can be tedious, the model fit view suggests additional universes that have similar point estimates to the selected universe.
[liu2020, fig1, ref3]

liu2020 - fig5



COLOR BY: None

UNCERTAINTY: Aggregated

X-AXIS RANGE

Fig. 5. Decision view and outcome view. (a) The decision view shows analytic decisions as a graph with order and dependencies between them, and highlights more sensitive decisions in darker colors. (b) The outcome view visualizes outputs from all analyses, including individual point estimates and aggregated uncertainty. [liu2020, fig5, ref1]

[all other references are to existing categories and tasks] [liu2020, fig5, [ref2]]

liu2020 - fig6

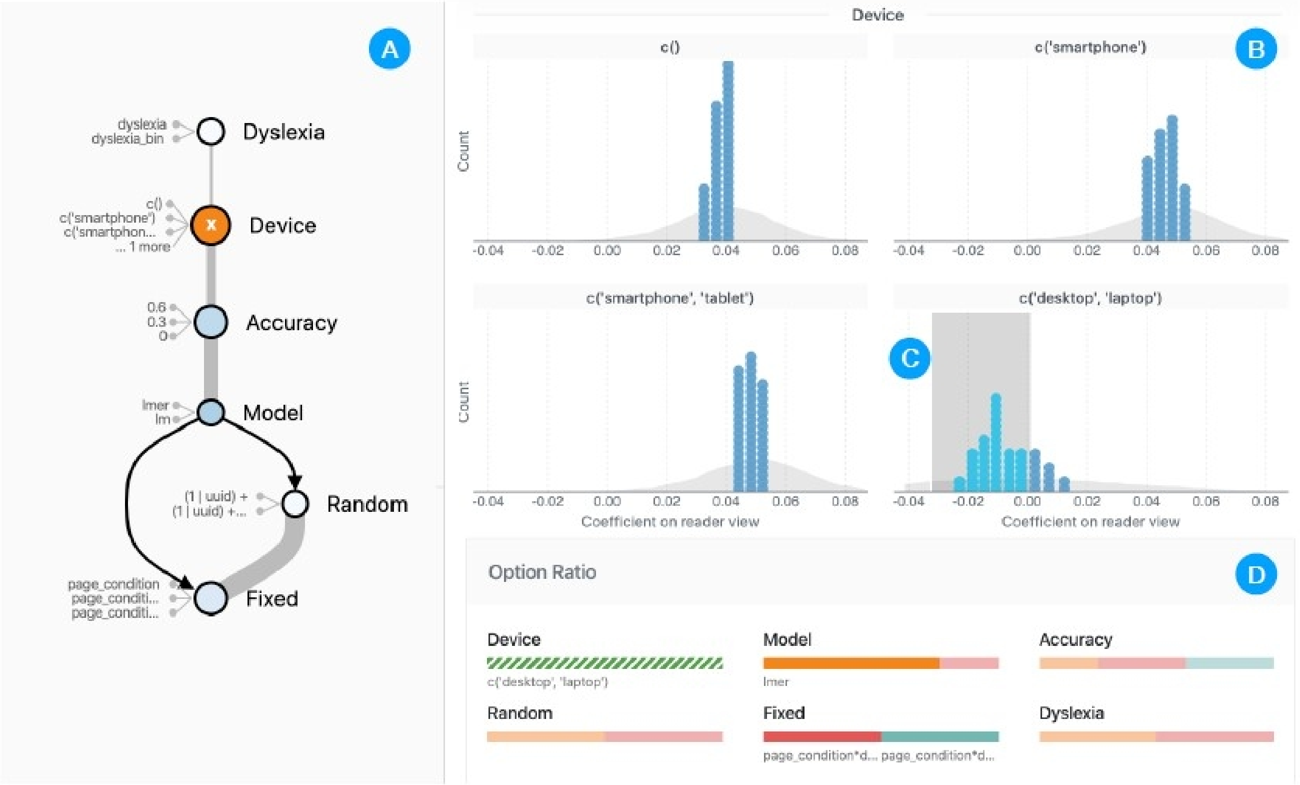


Fig. 6. Facet and Brushing. Clicking a node in the decision view (a) divides the outcome view into a trellis plot (b), answering questions like “does the decision lead to large variations in effect size?” Brushing a region in the outcome view (c) reveals dominant alternatives in the option ratio view (d), answering questions like “what causes negative results?”
[liu2020, fig6, ref1]

[all other references are to existing categories and tasks]
[liu2020, fig6, [ref2]]

liu2020 - fig7

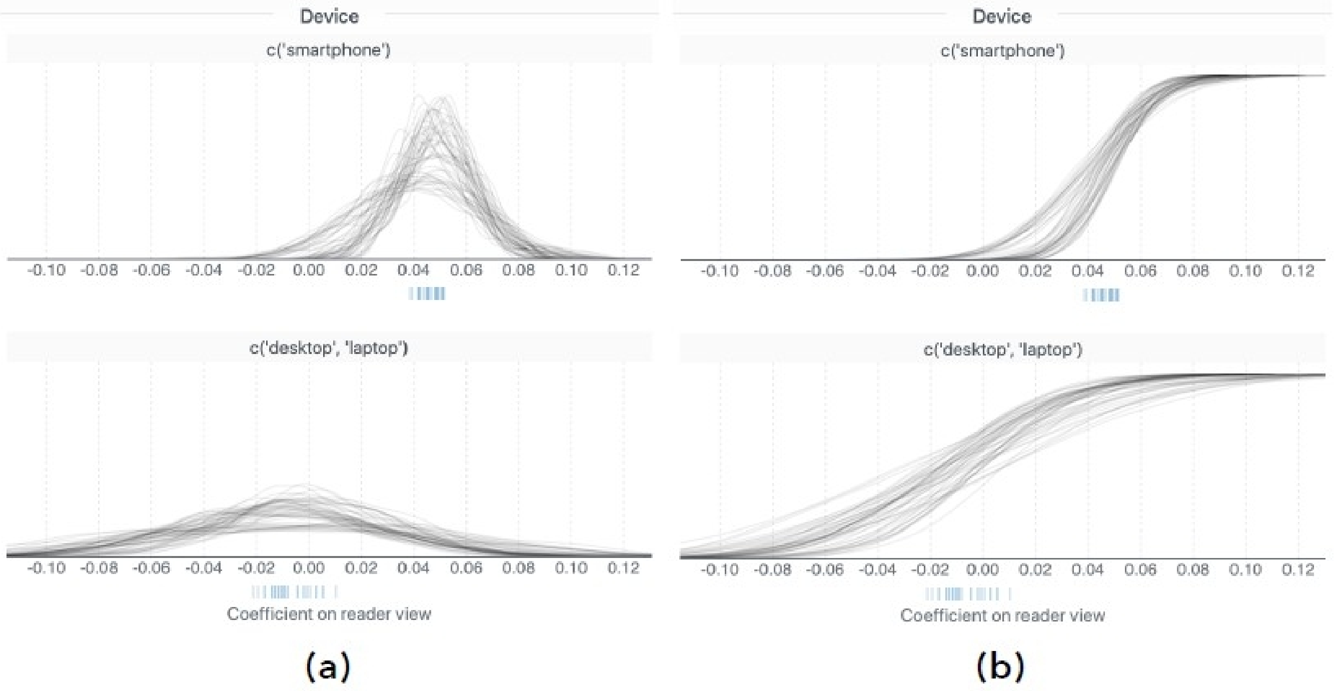


Fig. 7. PDFs (a) and CDFs (b) views visualize sampling distributions from individual universes. Toggling these views in a trellis plot allows users to compare the variance between conditions.
[liu2020, fig7, ref1]

Besides aggregated uncertainty, Boba allows users to examine uncertainty from individual universes (Fig. 7). In a dropdown menu, users can switch to view the probability density functions (PDFs) or cumulative distribution functions (CDFs) of all universes. A PDF is a function that maps the value of a random variable to its likelihood, whereas a CDF gives the area under the PDF. In both views, we draw a cubic basis spline for the PDF or CDF per universe, and reduce the opacity of the curves to visually “merge” the curves within the same space. There is again a one-to-one mapping between a visual element and a universe to afford interactions. To help connect point estimates and uncertainty, we draw a strip plot of point estimates beneath each PDFs/CDFs chart (Fig. 7, blue dashes), and show the corresponding sampling distribution PDF when users mouse over a universe in the dot plot.
[liu2020, fig7, ref2]

liu2020 - fig8

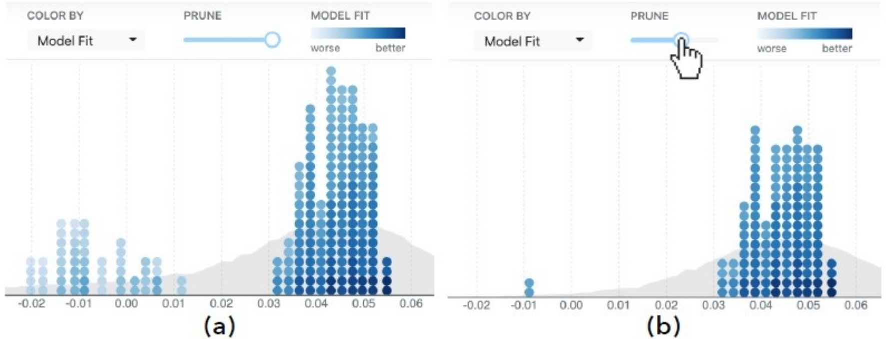


Fig. 8. (a) Coloring the universes according to their model fit quality. (b) Removing universes that fail to meet a model quality threshold. [liu2020, fig8, ref1]

Boba enables an overview of model fit quality across all universes (T5) by coloring the outcome view with a model quality metric (Fig. 8a). We use normalized root mean squared error (NRMSE) to measure model quality and map NRMSE to a single-hue colormap of blue shades where a darker blue indicates a better fit. [liu2020, fig8, ref3]

Now that Emma understands what decisions lead to null effects, she wonders if these results are from trustworthy models. She changes the color-by field to get an overview of model fit quality (Fig. 8a) and sees that the universes around zero have a poorer fit. She then uses a slider to remove universes that fail to meet a quality threshold (Fig. 8b).
[liu2020, fig8, ref2]

liu2020 - fig9

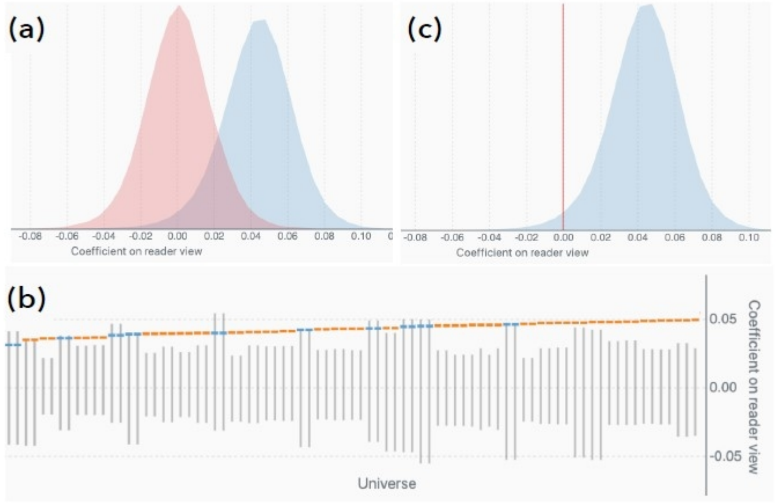


Fig. 9. Inference views. (a) Aggregate plot comparing the possible outcomes of the actual multiverse (blue) and the null distribution (red). (b) Detailed plot showing the individual point estimates and the range between the 2.5th and 97.5th percentile in the null distribution (gray line). Point estimates outside the range are colored in orange. (c) Alternative aggregate plot where a red line marks the expected null effect.
[liu2020, fig9, ref1]

To support users in making inference and judging how reliable the hypothesized effect is (T6), Boba provides an inference view at the end of the analysis workflow, after users have engaged in exploration. Once in the inference view, all earlier views and interactions are inaccessible to avoid multiple comparison problems [60] arising from repeated inference. The inference view contains different plots depending on the outputs from the authoring step, so that users can choose between robust yet computationally-expensive methods and simpler ones.
[liu2020, fig9, ref3]

In addition, Boba enables users to propagate concerns in model fit quality to the inference view in two possible ways. The first way employs a model averaging technique called stacking [58] to take a weighted combination of the universes according to their model fit quality. The technique learns a simplex of weights, one for each universe model, via optimization that maximizes the log-posteriordensity of the held-out data points in a k-fold cross validation. Boba then takes a weighted combination of the universe distributions to create the aggregate plot. While stacking provides a principled way to approach model quality, it can be computationally expensive. As an alternative, Boba excludes the universes below the model quality cutoff users provide in Sect. 5.4. The decisions of the cutoff and whether to omit the universes are made before a user enters the inference view.
[liu2020, fig9, ref5]

After an in-depth exploration, Emma proceeds to the final step, asking “given the multiverse, how reliable is the effect?” She confirms a warning dialog to arrive at the inference view (Fig. 9).
[liu2020, fig9, ref2]

A more robust inference utilizes the null distribution – the expected distribution of outcomes when the null hypothesis of no effect is true. In this case, the inference view shows an aggregate plot followed by a detailed plot (Fig. 9ab). The aggregate plot (Fig. 9a) compares the null distribution (red) to possible outcomes of the actual multiverse (blue) across sampling and decision variations. The detailed plot (Fig. 9b) shows point estimates (colored dots) against 95% confidence intervals representing null distributions (gray lines) for each universe. Each point estimate is orange if it is outside the range, or blue otherwise. Underneath both plots, we provide descriptions (supplemental Fig. 1) to guide users in interpretation: For the aggregate plot, we prompt users to compare the distance between the averages of the two densities to the spread. For the detailed plot, we count the number of universes with the point estimate outside its corresponding range. If the null distribution is unavailable, Boba shows a simpler aggregate plot (Fig. 9c) where the expected effect size under the null hypothesis is marked with a red line.
[liu2020, fig9, ref4]

liu2020 - fig10

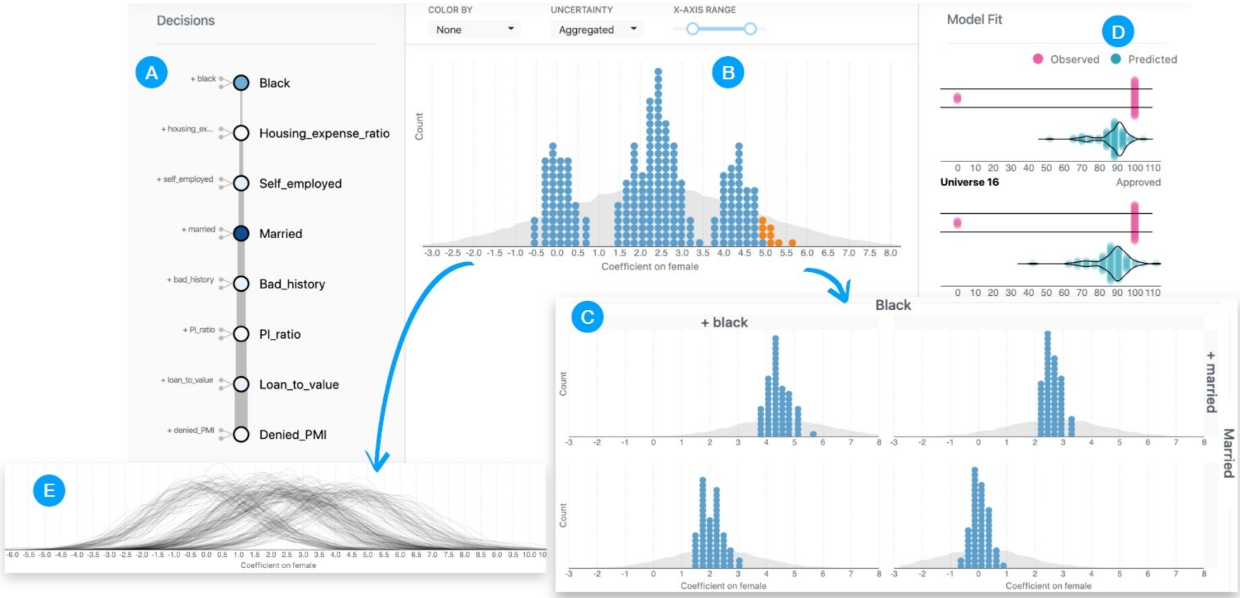


Fig. 10. A case study on how model estimates are robust to control variables in a mortgage lending dataset. (a) Decision view shows that black and married are two consequential decisions. (b) Overall outcome distribution follows a multimodal distribution with three peaks. (c) Trellis plot of black and married indicates the source of the peaks. (d) Model fit plots show that models produce numeric predictions while observed data is categorical. (e) PDFs of individual sampling distributions show significant overlap of the three peaks.
[liu2020, fig10, ref1]

We first demonstrate that the default views in the Boba Visualizer afford similar insights on uncertainty, robustness, and decision sensitivity. Upon launching the visualizer, we see a decision graph and an overall outcome distribution (Fig. 10). The decision view (Fig. 10a) highlights two sensitive decisions, black and married. The outcome view (Fig. 10b) shows that the point estimates are highly varied with conflicting implications. The aggregated uncertainty in the outcome view (Fig. 10b, background gray area) has a wide spread, suggesting that the possible outcomes are even more varied when taking both sampling and decision variability into account. These observations agree with the summary metrics in previous work, though Boba uses a different, non-parametric method to quantify decision sensitivity, as well as a different method to aggregate end-to-end uncertainty.
[liu2020, fig10, ref2]

The patterns revealed by ad-hoc visualizations in previous work are also readily available in the Boba Visualizer, either in the default views or with two clicks guided by prominent visual cues. The default outcome view (Fig. 10b) shows that the point estimates follow a multimodal distribution with three separate peaks. Clicking the two highlighted (most sensitive) nodes in the decision view (Fig. 10a) produces a trellis plot (Fig. 10c), where each subplot contains only one cluster. From the trellis plot, it is evident that the leftmost and rightmost peaks in the overall distribution come from two particular combinations of the influential variables. Alternatively, users might arrive at similar insights by brushing individual clusters in the default outcome view.
[liu2020, fig10, ref3]

Finally, the uncertainty and model fit visualizations in Boba surface potential issues that previous work might have overlooked. First, though the point estimates in Fig. 10b fall into three distinct clusters, the aggregated uncertainty distribution appears unimodal despite a wider spread. The PDF plot (Fig. 10e) shows that sampling distribution from one analysis typically spans the range of multiple peaks, thus explaining why the aggregated uncertainty is unimodal. These observations suggest that the multimodal patterns exhibited by point estimates are not robust when we take sampling variations into account. Second, we assess model fit quality by clicking a dot in the outcome view and examining the model fit view (Fig. 10d). As shown in Fig. 10d, while the observed data only takes two possible values, the linear regression model produces a continuous range of predictions. It is clear from this visual check that an alternative model, for example logistic regression, is more appropriate than the original linear regression models, and we should probably interpret the results with skepticism given the model fit issues. These observations support our arguments in Sect. 3.2 that uncertainty and model fit are potential blind spots in prior literature.
[liu2020, fig10, ref4]

liu2020 - fig11

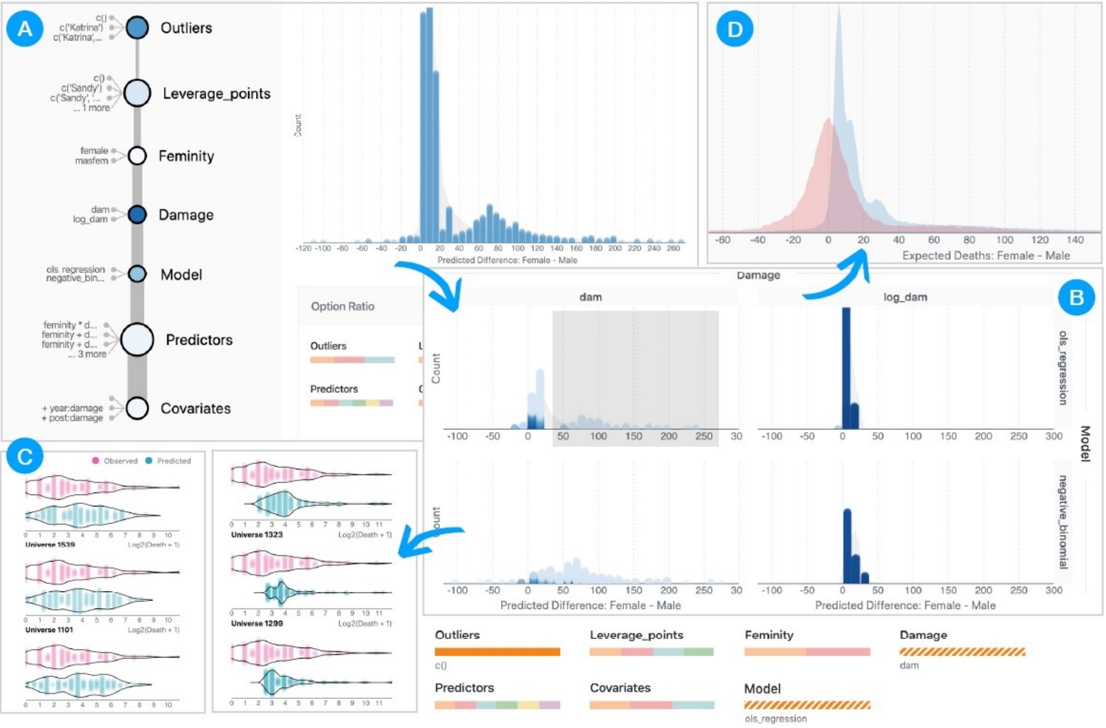


Fig. 11. A case study on whether hurricanes with more feminine names have caused more deaths. (a) The majority of point estimates suggest a small, positive effect, but there are considerable variations. (b) Faceting and brushing reveal decision combinations that produce large estimates. Coloring by model quality shows that large estimates are from questionable models, and predictive checks (c) confirms model fit issues. (d) Inference view shows that the observed and null distributions are different in terms of mode and shape, yet with highly overlapping estimates.
[liu2020, fig11, ref1]

Now that we have reasons to be skeptical of the large estimates, the remaining universes still support a small, positive effect. How reliable is the effect? We proceed to the inference view to compare the possible outcomes in the observed multiverse and the expected distribution under the null hypothesis (Fig. 11d). The two distributions are different in terms of mode and shape, yet they are highly overlapping, which suggests the effect is not reliable. The detail plot depicting individual universes (supplemental Fig. 1) further confirms this observation. Out of the entire multiverse, only 3 universes have point estimates outside the 2.5th and 97.5th percentile of the corresponding null distribution.
[liu2020, fig11, ref3]

But do we have evidence that certain outcomes are less trustworthy? We toggle the color-by drop-down menu so that each universe is colored by its model quality metric (Fig. 11b). The large estimates are almost exclusively coming from models with a poor fit. We further verify the model fit quality by picking example universes and examining the model fit view (Fig. 11c). The visual predictive checks confirm issues in model fit, for example the models fail to generate predictions smaller than 3 deaths, while the observed data contains plenty such cases.
[liu2020, fig11, ref2]

liu2020 - tasks

Explicit tasks supported by the Boba visualization system

T1: Decision Overview – gain an overview of the decision space to understand the multiverse and contextualize subsequent tasks.
[liu2020, tasks, T1]

T2: Robustness Overview – gauge the overall robustness of findings obtained through all reasonable specifications.
[liu2020, tasks, T2]

T3: Decision Impacts – identify what combinations of decisions lead to large variations in outcomes, and what combinations of decisions are critical in obtaining specific outcomes.
[liu2020, tasks, T3]

T4: Uncertainty – assess the end-to-end uncertainty as well as uncertainty associated with individual universes.
[liu2020, tasks, T4]

T5: Model Fit – assess the model fit quality of individual universes to distinguish trustworthy models from questionable ones.
[liu2020, tasks, T5]

T6: Inference – perform statistical inference to judge how reliable the hypothesized effect is, while accounting for model quality.
[liu2020, tasks, T6]

delgiudice2020 - fig4

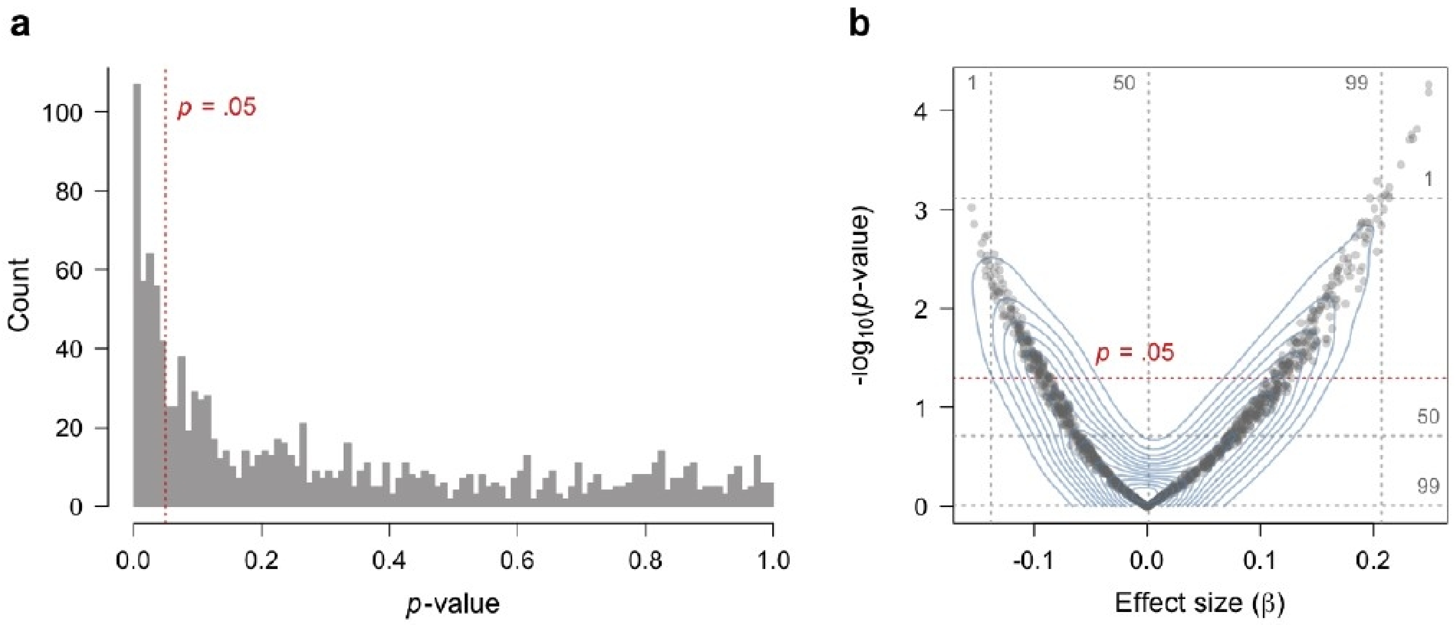


Figure 4. Results of the full multiverse-style analysis of the simulated dataset. (a) Distribution of p-values across 1,216 specifications. (b) Vibration of effects (VoE) plot showing the joint distribution of p-values and effect sizes for the same specifications. [delgiudice2020, fig4, ref1]

The distribution of p-values and vibration of effects in the full multiverse are shown in Figure 4. The median p was .194. Just 27% of the effects reached the conventional threshold of a = .05. Effect sizes ranged from b = -.16 to .25, with a median of b = .01. The VoE plot shows a clear “Janus effect” (see Patel et al., 2015), as the regression coefficients at the 1st and 99th percentiles of the effect size distribution have opposite signs (-.14 and .21, respectively). These results could be easily interpreted as indications of poor robustness and replicability. The median effect size across specifications was very close to zero and far from conventional significance thresholds, even though the true effect size in the population was b = .20 (before accounting for measurement validity). Investigators using the mean of the multiverse as a “robust” estimate would wrongly conclude that the effect of inflammation on depression is about zero. [delgiudice2020, fig4, ref2]

delgiudice2020 - fig5

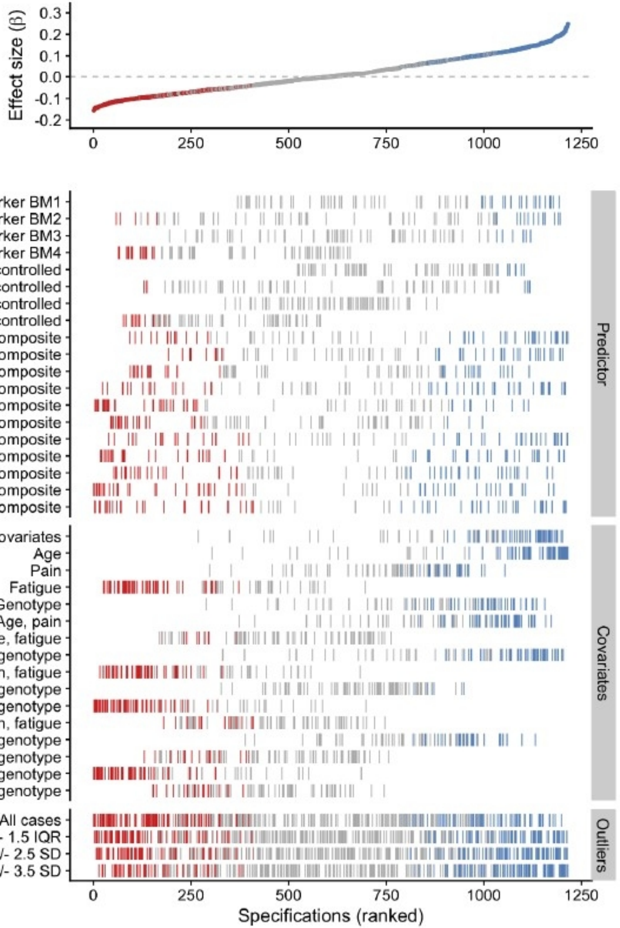


Figure 5. Specification curve for the simulated dataset (full multiverse of 1,216 specifications). Blue = positive effect sizes significant at a = .05. Red = positive effect sizes significant at a = .05. [delgiudice2020, fig5, ref1]

Clearly, the specification curve offers more opportunities to inspect the results for systematic patterns than the summary plots of Figure 4. Most investigators would probably recognize that the direction of effects depends strongly on whether fatigue is included as a covariate. Without explicit consideration of measurement validity, the results for alternative predictors may appear to suggest a lack of consistency, or at least marked sensitivity to the precise operationalization of inflammation. Overall, these results could readily be interpreted as a mixture of chance variation and high dependence on the details of the analysis.
[delgiudice2020, fig5, ref3]

Figure 5 displays a specification curve for the full multiverse. The significant effects are split between positive and negative. The pattern for alternative predictors reflects the impact of measurement validity, which is lower for individual biomarkers (especially with simultaneous entry) and higher for composites. But the central tendency of effects is similar across predictors. As for covariates, inspection of Figure 5 indicates that combinations that include fatigue tend to yield negative effects, whereas the direction tends to be positive when fatigue is excluded. Regardless of the general direction of effects, every combination produces a fair amount of nonsignificant findings. Alternative cutoffs for outliers do not seem to have a systematic impact, except that including all cases shifts the distribution toward somewhat more negative effects.
[delgiudice2020, fig5, ref1]

delgiudice2020 - fig6

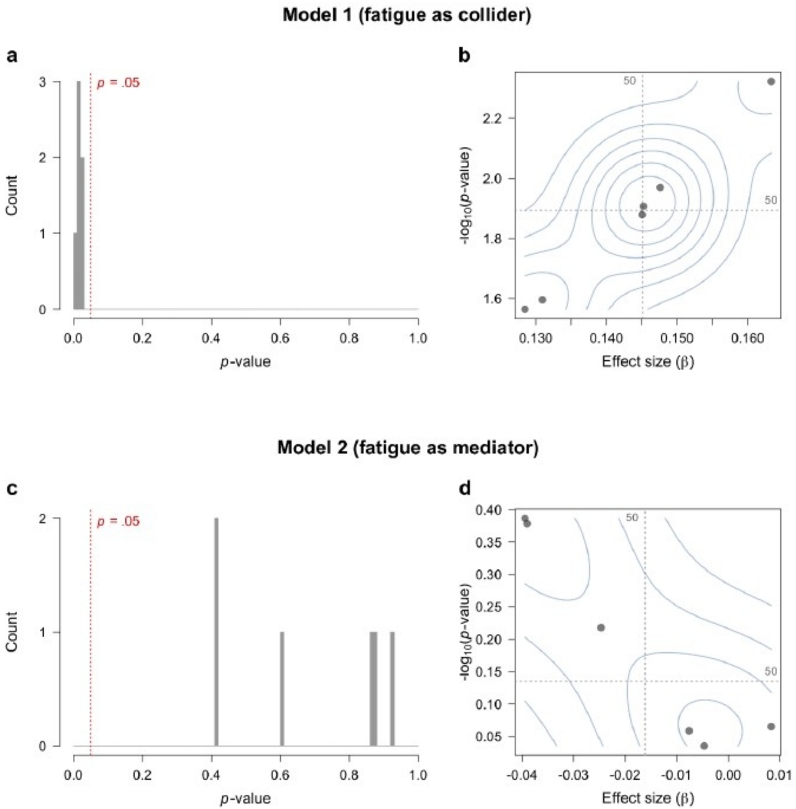
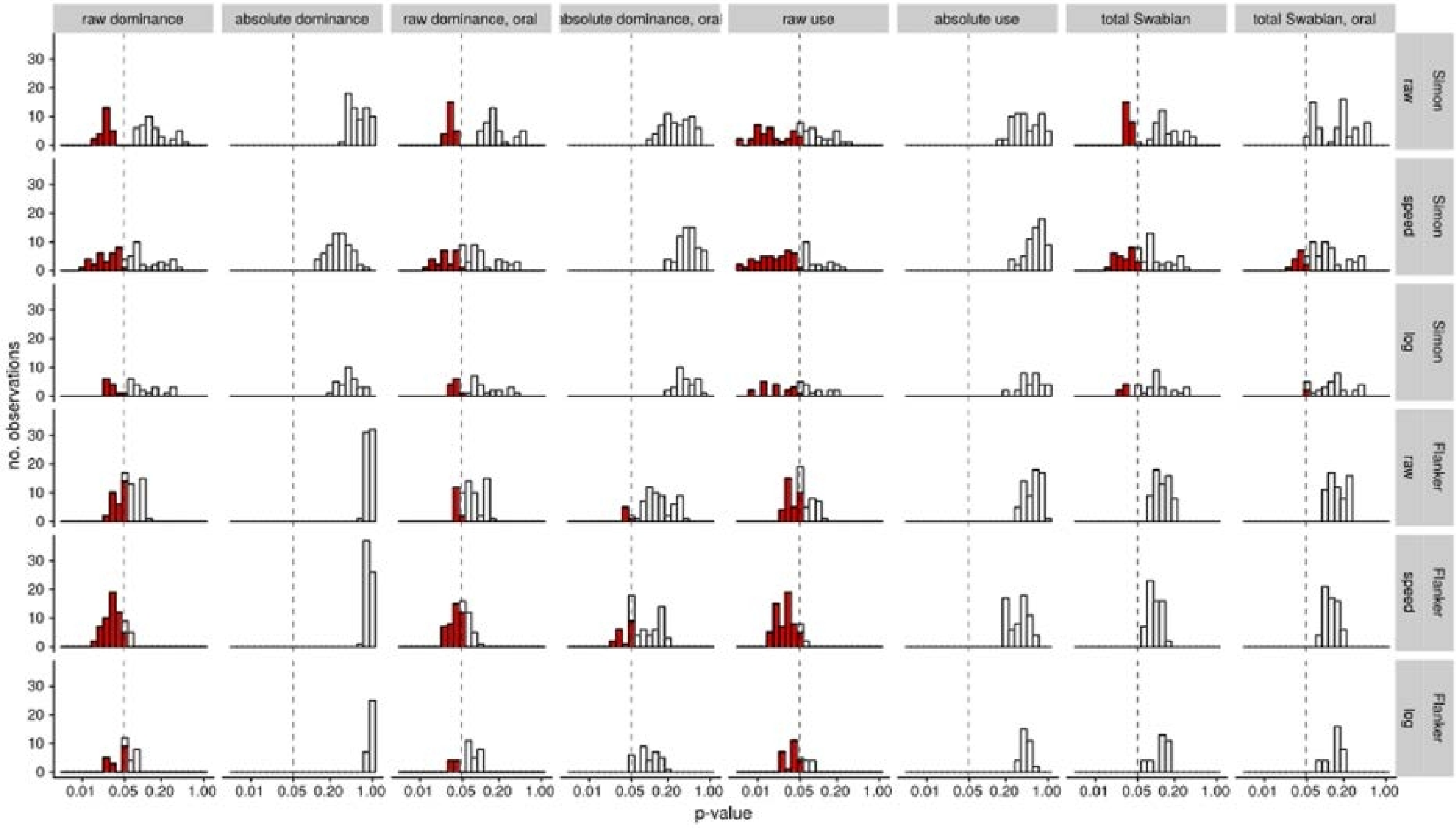


Figure 6. Results of the principled multiverse-style analyses of the simulated dataset. (a, c) Distribution of p-values across 6 specifications. (b, d) Vibration of effects (VoE) plots showing the joint distribution of p-values and effect sizes for the same specifications.
[delgiudice2020, fig6, ref1]

In sum, analyses of the principled multiverses revealed two homogeneous clusters of effects, indicating that the exact biomarker composite employed as a predictor and the choice of cutoff for outliers do not substantially change the conclusions of the study. What does make a difference is whether fatigue is treated as a collider and excluded as a covariate (Model 1), or treated as a mediator and controlled for in the analysis (Model 2). Making an informed decision between these models would require additional empirical evidence (e.g., experimental or quasiexperimental studies), theoretical developments, or both.
[delgiudice2020, fig6, ref3]

Figure 6 shows the distribution of p-values and VoE in the two principled multiverses. In the multiverse based on Model 1 (i.e., the true model that generated the data), all six effects were positive and statistically significant at a = .05, with median p = .012. Effect sizes clustered in a narrow range between b = .14 and .16; the median was b = .15. The consistency of effects within this multiverse is reflected in the VoE plot of Figure 6b. In the multiverse based on Model 2 (which incorrectly assumes that fatigue is a mediator), the effects ranged from b = -.04 to .01, with a median (and mean) of b = -.02. These small negative effects failed to meet the threshold for significance; the median p-value was .733.
[delgiudice2020, fig6, ref2]

This is an example of using faceting to encode options and parameters



This is an example of using faceting to encode different universes according to two parameters (each with two options)

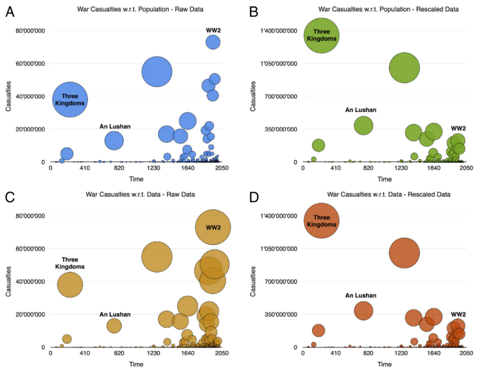


Fig. 2. War casualties over time, using raw (A, C) and rescaled (B, D) data. The size of each bubble represents the size of each event with respect to today's world population (A, B) and with respect to the total casualties (raw: C, rescaled: D) in the data set.

This is an example of something we don't consider a Composite plot. The axes aren't aligned, and don't appear to have a "super-additive" effect of supporting a task the plots individually can't

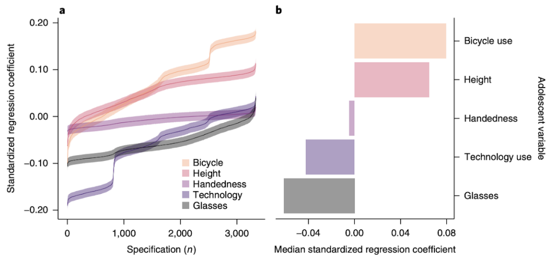


Fig. 5 | Comparison specifications for MCS. Visualization of the comparison specifications hypothesized to have little or no influence on well-being: bicycle use, height, handedness and wearing glasses. This graph shows SCA for both the variable of interest (mean technology use) and the comparison variables; it highlights the range of possible results of a simple cross-sectional regression of the variables of interest on adolescent well-being. Wearing glasses has the most negative association with adolescent well-being (black, median $\beta = -0.061$, median $n = 7{,}963$, partial $\eta^2 = 0.005$, median standard error $= 0.010$); and more negative than the association of technology use with well-being (purple, median $\beta = -0.042$, median $n = 7{,}964$, partial $\eta^2 = 0.002$, median standard error $= 0.010$). Handedness (red/purple, median $\beta = -0.004$, median $n = 7{,}972$, partial $\eta^2 < 0.001$, median standard error $= 0.010$), height of the adolescent (red, median $\beta = 0.065$, median $n = 7{,}910$, partial $\eta^2 = 0.005$, median standard error $= 0.010$) and whether the adolescent often rides a bicycle (yellow, median $\beta = 0.080$, median $n = 7{,}974$, partial $\eta^2 = 0.007$, median standard error $= 0.010$) have more positive associations with adolescent well-being than does technology use. **a,** How different analytical decisions (specifications, shown on the x axis) lead to different statistical outcomes (standardized regression coefficient, shown on the y axis). Each line represents a different variable of interest while the error bars represent the standard error. **b**, The resulting median standardized regression coefficients for those SCAs linking the variables of interest with adolescent well-being.

This is where the visualization can be handy: a means to address questions (above) to an end (conclusions below). We survey what people have done, but anything can be designed :)

Inspection
(visual, analytical, or other)

How do outcomes vary?

outcome variation across parameters/options

"Among other differences with all of these approaches, Specification Curve Analysis: (i) helps identify the source of variation in results across specifications via a descriptive specification curve (see Figure 2), and (ii) ..."
[simonsohn2015, fig2, ref3]

"This enables readers to visually identify both the variation in effect size across specifications, and its covariation with operationalization decisions."
[simonsohn2015, fig2, ref2]

"Finally, for voting and donation preferences (Panels C and D, respectively), it is hard to extract a consistent pattern of fluctuation across the different choice combinations."
[steegen2016, fig2, ref5]

"It seems that all arbitrary choices for data processing can have an impact on whether the obtained data set will lead to a significant or a nonsignificant outcome."
[steegen2016, fig2, ref5]

"For two variables—religiosity in Study 1 (Panel A) and fiscal political attitudes (Panel C)—the multiverse analysis reveals a near-uniform distribution, indicating that the p value for the interaction effect between fertility and relationship varies widely across the multiverse."
[steegen2016, fig1, ref3]

"For religiosity, 7 out of the 120 choice combinations lead to a significant interaction effect, whereas the remaining 94% lead to p values ranging from .05 to 1.0. For fiscal political attitudes, 8% of the 210 choice combinations lead to a significant interaction (p < .05), whereas the remaining choice combinations lead to p values across the entire range from .05 to 1.0."
[steegen2016, fig1, ref3]

"In each panel, the cells show the different p values that can be obtained across all choice combinations for data processing."
[steegen2016, fig2, ref2]

shape of outcome distribution

"For two variables—religiosity in Study 1 (Panel A) and fiscal political attitudes (Panel C)—the multiverse analysis reveals a near-uniform distribution, indicating that the p value for the interaction effect between fertility and relationship varies widely across the multiverse."
[steegen2016, fig1, ref3]

"Figure 1 shows the modeling distribution as a density graph of all the estimates calculated; the vertical line marks the 11 percent wage premium estimate from Table 2. Estimates as low as 9 percent and as high as over 20 percent are possible in the model space."
[young2017, fig1, ref3]

"The modeling distribution is multimodal with clusters of estimates around zero, 2.3, and 4.5 percent. . It seems hard to draw substantive conclusions from the evidence without knowing more about the modeling distribution. Why do these estimates vary so much? Why is the distribution so non-normal? "
[young2017, fig2, ref5]

"As shown in Table 8, the tax coefficient is statistically significant in only 1.5 percent of all models. The mean estimate is almost exactly zero, and estimates are evenly split between positive tax flight estimates (48.9 percent) and wrong-signed negative estimates (51.1 percent). "
[young2017, fig4, ref5]

"The modeling distribution is relatively normal: There are no critically important modeling decisions that generate bimodality in the estimates. As shown in Figure 4, the significant estimates reported in Table 7 above are extreme outliers in the modeling distribution."
[young2017, fig4, ref5]

The ones below are outcomes under particular assumptions, used to contrast sets of options

"Under these assumptions, the estimates of the gender effect are tightly centered around zero, with an almost even split between positive (52 percent) and negative (48 percent) estimates, none of which are statistically significant. Here, there is no evidence at all for a gender effect. "
[young2017, fig3, ref3]

"In contrast, the second distribution is defined by the opposite assumption: race and marital status must be in the model, but all combinations of the other controls are possible. Under these assumptions, the estimates cluster around a 4.5 percent higher mortgage acceptance rate for women. Both the significance rate and the sign stability are 100 percent— complete robustness."
[young2017, fig3, ref3]

outcome variation across parameters/options | shape of outcome distribution

visualization tasks?

?

"Next, we describe the VoE methodology for the association between serum vitamin D and all-cause mortality (Fig. 1B). ... The HR and the respective P-value for the association of that variable with all-cause mortality are estimated for all 8,193 models with different combinations of 13 adjusting variables using Cox proportional hazards time-to-event regression (Fig. 1D). We visualized the VoE for a given variable by plotting the HR vs. -log10(P-value) as two-dimensional histogram and a contour plot (Fig. 1E)."
[patel2015, fig1, ref4]

?

"To identify the key variable(s) that separated these different distributions, we visualized the VoE by coloring each point on whether it included (or did not include) each one of the 13 adjustment variables in the model, leading to 13 separate visualizations. ... "
[patel2015, fig3, ref3]

?

"To identify the key variable(s) that separated these different distributions, we visualized the VoE by coloring each point on whether it included (or did not include) each one of the 13 adjustment variables in the model ... Specifically, models that contained the smoking adjustment variable (Fig. 3B, yellow points) had HR lower than the models without the smoking adjustment and lower -log10(P-values) (Fig. 3B, black points). One source of cadmium exposure includes smoking ... "
[patel2015, fig3, ref3]

?

"To identify the key variable(s) that separated these different distributions, we visualized the VoE by coloring each point on whether it included (or did not include) each one of the 13 adjustment variables in the model ... In serum cadmium, we observed the two distinct modes were indicative of models that contained or did not contain current or past smoking (Fig. 3B). Specifically, models that contained the smoking adjustment variable (Fig. 3B, yellow points) had HR lower than the models without the smoking adjustment and lower -log10(P-values) (Fig. 3B, black points). ... "
[patel2015, fig3, ref3]

visual/inspect tasks? counting, frequencies, etc.

commonality/rareness of multiverse elements: counting of occurrence, relative rates

commonality of conclusions? *(are these just outcomes, or something else?)*

"Figure 2 shows that PNAS could have published nearly 1,700 letters showing individual specifications that make the effect go away (without deviating from the original red circle). It also could have published 37 responses with individual specifications showing the robustness of the findings. It would be better to publish a single specification curve in the original paper."
[simonsohn2015, fig2, ref7]

"What were the consequences of this variability in analytic approaches? ... Twenty teams (69%) found a significant positive relationship, p < .05, and nine teams (31%) found a nonsignificant relationship. No team reported a significant negative relationship."
[silberzahn2017, fig2, ref2]

"In total, 15 teams estimated a fixed effect or variance component for players, referees, or both; 12 of these teams reported significant effects (median OR = 1.32, MAD = 0.12). Eight teams used clustered standard errors, and 4 of these teams found significant effects (median OR = 1.28, MAD = 0.13). An additional 5 teams did not account for this artifact, and 4 of these teams reported significant effects (median OR = 1.39, MAD = 0.28). The remaining team used fixed effects for the referee variable and reported a nonsignificant result (OR = 0.89)."
[silberzahn2017, fig3, ref3]

commonality of outcomes

"The first pattern is exemplified by the association between serum levels of vitamin D and mortality (Fig. 2A). ... Most of the results are centered on HR ~ 0.72 and P ~ 10^-4 (two-dimensional mode)."
[patel2015, fig2, ref2]

"For a-tocopherol, most of the HR and P-values were concentrated around 1 and nonsignificance, respectively. However, 1% of the models had an HR < 0.875 (12.5% decreased risk of death for 1SD increase in exposure) with a nominally significant P-value (P < 0.05), whereas another 1% of the models had HR > 1.05 (5% increased risk for death for 1SD increase of exposure), albeit without reaching nominal significance."
[patel2015, fig2, ref5]

"The figure shows that the majority of specifications lead to estimates of the sign predicted by the original authors (feminine hurricanes produce more deaths), though a very small minority of all estimates are statistically significant (p<.05). The point estimates range from -1 to +12 additional deaths."
[simonsohn2015, fig2, ref4]

"For religiosity in Study 2 (Panel A), most data sets constructed under the second option for relationship assessment (R2) yield a nonsignificant interaction effect."
[steegen2016, fig2, ref3]

"The multiverse analysis revealed that almost all choice combinations for data processing lead to large p values."
[steegen2016, fig2, ref6]

"We also visualized trends corresponding to the number of adjusting variables (k), plotting the median effect size and P-value for each k from 0 to 13. We recorded the proportion of estimates that achieved different levels of nominal statistical significance (P < 0.05, 0.0001)."
[patel2015, fig1, ref6]

(outliers/rare outcomes)

"For example, we can see that obtaining a negative point estimate requires a fairly idiosyncratic combination of operationalizations: (i) not taking into account the year of the storm, (ii) operationalizing severity of the storm by the log of damages, (iii) conducting an OLS regression, etc. A researcher motivated to show a negative point estimate would be able to report twenty different specifications that do so, but the specification curve shows that a negative point estimate is atypical."
[simonsohn2015, fig2, ref5]

commonality of associations between outcomes and parameters/options

"For the remaining four variables, roughly half of the choice combinations lead to a significant interaction effect. In particular, for religiosity in Study 2 (Panel B), 88 out of the 210 choice combinations (42%) lead to a p value smaller than .05. Regarding social political attitudes (Panel D), 49% of the p values is smaller than .05. Finally, 46% and 57% of the p values are smaller than .05 for voting (Panel E) and donation (Panel F) preferences, respectively."
[steegen2016, fig1, ref4]

"Of the 417 variables, 91 (22%) variables had similar behavior to thyroxine in which HR were attenuated and the P-values rose above the nominal level of significance (P > 0.05) as the number of adjusting variables, k, increased (see Fig. S1 and Table S3/ Appendix at www.jclinepi.com)."
[patel2015, fig2, ref3]

"Of the 417 variables, 53 (13%) exhibited similar behavior to vitamin D, where all associations were beyond the level of nominal statistical significance, but the association was attenuated with a greater number of adjustment variables (see Fig. S1/Appendix at www.jclinepi.com)."
[patel2015, fig2, ref2]

"Twenty-six (6%) of the 417 variables exhibited similar behavior to urinary creatinine where the effect sizes increased and P-values decreased for larger k."
[patel2015, fig2, ref4]

"Examples such as those in Fig. 2A-D represented the VoE patterns for 72% of the 417 associations."
[patel2015, fig2, ref6]

"The Janus effect is common: 131 (31%) of the 417 variables had their 99th percentile HR > 1 and their 1st percentile HR < 1."
[patel2015, fig2, ref5]

" ... Multimodality of effects was clearly seen in 71 of the 417 (17%) assessed variables. ... "
[patel2015, fig3, ref2]

commonality of parameters/options *(when parameters/options are not all equally likely/common)*

"Fifteen teams used logistic models, and 11 of these teams found a significant effect (median OR = 1.34; median absolution deviation, or MAD = 0.07). Six teams used Poisson models, and 4 of these teams found a significant effect (median OR = 1.36, MAD = 0.08). Of the 6 teams that used linear models, 3 found a significant effect (median OR = 1.21, MAD = 0.05). The final 2 teams used models classified as miscellaneous, and both of these teams reported significant effects (ORs = 1.71 and 2.88, respectively)."
[silberzahn2017, fig3, ref2]

"In total, 15 teams estimated a fixed effect or variance component for players, referees, or both; 12 of these teams reported significant effects (median OR = 1.32, MAD = 0.12). Eight teams used clustered standard errors, and 4 of these teams found significant effects (median OR = 1.28, MAD = 0.13). An additional 5 teams did not account for this artifact, and 4 of these teams reported significant effects (median OR = 1.39, MAD = 0.28). The remaining team used fixed effects for the referee variable and reported a nonsignificant result (OR = 0.89)."
[silberzahn2017, fig3, ref3]

visualization tasks?

quantitative associations?

"We created metrics to express the distributions of VoE for a variable (Fig. 1F). ... We also assessed whether associations appeared on both sides of the null (HR <1 and HR >1): depending on what adjustments are chosen, the results may suggest that the variable of interest is associated with either increased or decreased mortality."
[patel2015, fig1, ref5]

"We also visualized trends corresponding to the number of adjusting variables (k), plotting the median effect size and P-value for each k from 0 to 13. We recorded the proportion of estimates that achieved different levels of nominal statistical significance (P < 0.05, 0.0001)."
[patel2015, fig1, ref6]

distinctive association pattern?

"4.3. Identifying "multimodality of effects" with VoE: ... For example, the overall VoE for serum cadmium on mortality indicates strong association with mortality (Fig. 3A); all of the HRs are >1.2 per 1 SD change in serum cadmium levels, and P-values in all analytical scenarios are <0.05. However, two modes are visually evident (Fig. 3A)."
[patel2015, fig3, ref2]

"To identify the key variable(s) that separated these different distributions, we visualized the VoE by coloring each point on whether it included (or did not include) each one of the 13 adjustment variables in the model ... In serum cadmium, we observed the two distinct modes were indicative of models that contained or did not contain current or past smoking (Fig. 3B). ... Furthermore, we observed that models that included (or did not include) alcohol drinking also resulted in separate modes in P-values (Fig. 3C)."
[patel2015, fig3, ref3]

Inspection

A few examples of: **outcome types**

effect sizes (sign/direction, magnitude)

"Across the 256 possible combinations of controls, the effect of gender is typically positive but only 25 percent of the estimates are statistically significant. And 12 percent of the estimates have the opposite sign (though none of those estimates are significant)."
[young2017, fig2, ref3]

"The specification "curve" shows the estimated effect size across all specifications, sorted by magnitude, accompanied below by a "dashboard chart" indicating the operationalizations behind each result (see e.g., Figure 2)."
[simonsohn2015, fig2, ref2]

"Figure 2 reports the descriptive specification curve for the hurricanes example. The top panel depicts estimated effect size, in additional fatalities, of a hurricane having a feminine rather than masculine name. The figure shows that the majority of specifications lead to estimates of the sign predicted by the original authors (feminine hurricanes produce more deaths), though a very small minority of all estimates are statistically significant (p<.05). The point estimates range from -1 to +12 additional deaths."
[simonsohn2015, fig2, ref4]

"As this figure and Table 3 show, the estimated effect sizes ranged from 0.89 (slightly negative) to 2.93 (moderately positive) in odds-ratio (OR) units; the median estimate was 1.31. The confidence intervals for many of the estimates overlap, which is expected because they are based on the same data."
[silberzahn2017, fig2, ref2]

"Fifteen teams used logistic models, and 11 of these teams found a significant effect (median OR = 1.34; median absolution deviation, or MAD = 0.07). Six teams used Poisson models, and 4 of these teams found a significant effect (median OR = 1.36, MAD = 0.08). Of the 6 teams that used linear models, 3 found a significant effect (median OR = 1.21, MAD = 0.05). The final 2 teams used models classified as miscellaneous, and both of these teams reported significant effects (ORs = 1.71 and 2.88, respectively)."
[silberzahn2017, fig3, ref2]

"Figure 1 shows the modeling distribution as a density graph of all the estimates calculated; the vertical line marks the 11 percent wage premium estimate from Table 2. Estimates as low as 9 percent and as high as over 20 percent are possible in the model space."
[young2017, fig1, ref4]

"In contrast, the second distribution is defined by the opposite assumption: race and marital status must be in the model, but all combinations of the other controls are possible. Under these assumptions, the estimates cluster around a 4.5 percent higher mortgage acceptance rate for women. Both the significance rate and the sign stability are 100 percent— complete robustness."
[young2017, fig3, ref3]

"As shown in Table 8, the tax coefficient is statistically significant in only 1.5 percent of all models. The mean estimate is almost exactly zero, and estimates are evenly split between positive tax flight estimates (48.9 percent) and wrong-signed negative estimates (51.1 percent). "
[young2017, fig4, ref5]

statistical significance (binary or continuous)

"As this figure and Table 3 show, the estimated effect sizes ranged from 0.89 (slightly negative) to 2.93 (moderately positive) in odds-ratio (OR) units; the median estimate was 1.31. The confidence intervals for many of the estimates overlap, which is expected because they are based on the same data."
[silberzahn2017, fig2, ref2]

"The figure shows that the majority of specifications lead to estimates of the sign predicted by the original authors (feminine hurricanes produce more deaths), though a very small minority of all estimates are statistically significant (p<.05). The point estimates range from -1 to +12 additional deaths."
[simonsohn2015, fig2, ref4]

"What were the consequences of this variability in analytic approaches? ... Twenty teams (69%) found a significant positive relationship, p < .05, and nine teams (31%) found a nonsignificant relationship. No team reported a significant negative relationship."
[silberzahn2017, fig2, ref2]

"Fifteen teams used logistic models, and 11 of these teams found a significant effect (median OR = 1.34; median absolution deviation, or MAD = 0.07). Six teams used Poisson models, and 4 of these teams found a significant effect (median OR = 1.36, MAD = 0.08). Of the 6 teams that used linear models, 3 found a significant effect (median OR = 1.21, MAD = 0.05). The final 2 teams used models classified as miscellaneous, and both of these teams reported significant effects (ORs = 1.71 and 2.88, respectively)."
[silberzahn2017, fig3, ref2]

"In each panel, the cells show the different p values that can be obtained across all choice combinations for data processing."
[steegen2016, fig2, ref2]

"In particular, for religiosity in Study 2 (Panel B), 88 out of the 210 choice combinations (42%) lead to a p value smaller than .05. Regarding social political attitudes (Panel D), 49% of the p values is smaller than .05. Finally, 46% and 57% of the p values are smaller than .05 for voting (Panel E) and donation (Panel F) preferences, respectively."
[steegen2016, fig1, ref4]

"Across the 256 possible combinations of controls, the effect of gender is typically positive but only 25 percent of the estimates are statistically significant. And 12 percent of the estimates have the opposite sign (though none of those estimates are significant)."
[young2017, fig2, ref3]

"As shown in Table 8, the tax coefficient is statistically significant in only 1.5 percent of all models. The mean estimate is almost exactly zero, and estimates are evenly split between positive tax flight estimates (48.9 percent) and wrong-signed negative estimates (51.1 percent). "
[young2017, fig4, ref5]

"For religiosity, 7 out of the 120 choice combinations lead to a significant interaction effect, whereas the remaining 94% lead to p values ranging from .05 to 1.0."
[steegen2016, fig1, ref3]

"For two variables—religiosity in Study 1 (Panel A) and fiscal political attitudes (Panel C)—the multiverse analysis reveals a near-uniform distribution, indicating that the p value for the interaction effect between fertility and relationship varies widely across the multiverse."
[steegen2016, fig1, ref3]

effect sizes | statistical significance

broader interpretation, conclusions about scientific uncertainty, research questions, etc.

May or may not be vis related

Interpretation and Conclusions?

?

?	?
[column width placeholder - delete when something is here]	[column width placeholder - delete when something is here]
?	?

uncertainty/ambiguity

"For religiosity [...] The multiverse analysis revealed that almost all choice combinations for data processing lead to large p values. As such nonsignificant findings in general represent nothing more than uncertainty, this pattern of results clearly raises serious questions regarding the finding on the effect of fertility found in the single data set analysis, and should make a researcher hesitant to trust the single data set finding."
[steegen2016, fig2, ref6]

"If no strong arguments can be made for certain choices, we are left with many branches of the multiverse that have large p values. In these cases, the only reasonable conclusion on the effect of fertility is that there is considerable scientific uncertainty. One should reserve judgment and acknowledge that the data are not strong enough to draw a conclusion on the effect of fertility."
[steegen2016, fig2, ref8]

"For most other variables, there was considerable ambiguity: The interaction seemed to be significant across about half of the arbitrary choice combinations. In these cases, the conclusion on the effect of fertility strongly depends on the evaluation of the different processing options."
[steegen2016, fig2, ref7]

"Figure 2 shows that PNAS could have published nearly 1,700 letters showing individual specifications that make the effect go away (without deviating from the original red circle). It also could have published 37 responses with individual specifications showing the robustness of the findings. It would be better to publish a single specification curve in the original paper."
[simonsohn2015, fig2, ref7]

"We created metrics to express the distributions of VoE for a variable (Fig. 1F). ... We also assessed whether associations appeared on both sides of the null (HR <1 and HR >1): depending on what adjustments are chosen, the results may suggest that the variable of interest is associated with either increased or decreased mortality."
[patel2015, fig1, ref5]

"However, when we relax the assumption that any one of these control variables must be in the model—allowing us to consider all possible combinations of the controls—there is much uncertainty about the estimate."
[young2017, fig2, ref3]

"It seems hard to draw substantive conclusions from the evidence without knowing more about the modeling distribution. Why do these estimates vary so much? Why is the distribution so non-normal? What combinations of control variables are critical to finding a positive and significant result? These questions lead us to the next stage in our analysis: understanding model influence."
[young2017, fig2, ref5]

non-acceptance, dismissal, or disbelief?

"The effect of fertility on religion seems too sensitive to arbitrary choices and thus too fragile to be taken seriously."
[steegen2016, fig2, ref6]

"Readers need not take a position on whether it does or does not make sense to include a damages x pressure interaction in the model to determine if the original findings are robust."
[simonsohn2015, fig2, ref6]

"The figure shows that the majority of specifications lead to estimates of the sign predicted by the original authors (feminine hurricanes produce more deaths), though a very small minority of all estimates are statistically significant (p<.05). The point estimates range from -1 to +12 additional deaths."
[simonsohn2015, fig2, ref4]

"Returning to Figure 1, this appears to be a Panel C situation. Original authors and critics disagree on the set of valid specifications to run. The specification curve results from Figure 2 show that, while such disagreements may be legitimate and profound, we do not need to address them to determine what to make of the hurricanes data. In particular, the figure shows that even keeping the same set of observations as the original study and treating damages in the same way as treated in the original, modifying virtually any arbitrary analytical decision renders the original effect nonsignificant."
[simonsohn2015, fig2, ref6]

"In this case, when the robustness analysis is so overwhelmingly nonsupportive, the influence analysis has less to work with. However, there are a few informative points. Compared to Poisson, the negative binomial and OLS log-linear models give less positive estimates. Estimates from themodels using IRS rather than ACS data are more positive. This suggests that the most supportive evidence will come from using Poisson with the IRS data (reported as model 3 above), and the least supportive evidence will come from using OLS log-linear models with ACS data. Yet, even when we narrow our robustness testing to the most supportive estimator (Poisson) and data set (IRS), there is weak support: while the sign stability is 100 percent, the income tax effect is significant in only 1 percent of those models."
[young2017, fig4, ref6]

"While it is possible to support the tax flight hypothesis with a few knife-edge model specifications, there is remarkably little support even in a more narrow and supportive robustness analysis."
[young2017, fig4, ref8]

outcome robustness

"In this first pattern, one concludes that although adjustment weakens the magnitude relationship between vitamin D levels and mortality, inferences regarding the relationship are similar throughout all scenarios of adjustment."
[patel2015, fig2, ref2]

"4.3. Identifying "multimodality of effects" with VoE: ... For example, the overall VoE for serum cadmium on mortality indicates strong association with mortality (Fig. 3A); all of the HRs are >1.2 per 1 SD change in serum cadmium levels, and P-values in all analytical scenarios are <0.05. However, two modes are visually evident (Fig. 3A)."
[patel2015, fig3, ref2]

" Our conclusion is that, within the scope of these model ingredients, the positive union wage premium is a clear and strongly robust result. This suggests that the decline of unionization in America may well have contributed to middle-class wage stagnation—and not just for male workers (Rosenfeld 2014)."
[young2017, fig1, ref4]

"Figure 4. Robustness checks for predictors. Coefficient plot showing a consistent effect of the fertility predictor among naturally cycling women (red) but not hormonal contraception users (black) across several predictor and model specifications (explained in further detail in the text)."
[arslan2018, fig4, ref1]

"Under these assumptions, the estimates cluster around a 4.5 percent higher mortgage acceptance rate for women. Both the significance rate and the sign stability are 100 percent—complete robustness."
[young2017, fig3, ref3]

evaluate/rank validity of options and impact on conclusions (weighting)?

"Both the authors performing the multiverse analysis and the readers of the research can construct arguments in favor or against certain choices, and the validity of these arguments will help drawing the conclusion. For example, if additional information suggests that the fifth option of assessing fertility is clearly superior, then Panel A of Figure 2 indicates that there is little evidence for an effect of fertility on religiosity in Study 2. On the other hand, if additional information suggests that the second option of assessing fertility is clearly superior, then most choice combinations lead to a significant interaction effect."
[steegen2016, fig2, ref7]

"Combinatorially including or excluding these variables produces models that we regard as nonsense, so we impose the assumption that they must be in all models. However, we leave as debatable the controls for distance, contiguity, other tax rates, economic performance of the states, and a rich set of natural amenities which have been previously shown to influence migration (McGranahan 1999)."
[young2017, fig4, ref4]

"It remains technically possible that the one-in-a-thousand specifications of Table 7 present the best, most theoretically compelling estimates. If so, authorswould need to carefully explain to readers why such painstakingly exact model assumptions are required, and why virtually any departure from model 2 or 3 fails to support the conclusions."
[young2017, fig4, ref8]

"A related point is that not all options are necessarily exactly interchangeable. Some options might seem better than others, at least for some researchers. If such is the case, this knowledge can be used to construct arguments for interpreting results such as those shown in Figure 2. However, a multiverse analysis should involve all plausible construction alternatives, not just the most plausible ones. When only one choice is clearly and unambiguously the most appropriate one, variation across this choice is uninformative."
[steegen2016, fig2, ref12]

explain reason for pattern/feature?

"As this figure and Table 3 show, the estimated effect sizes ranged from 0.89 (slightly negative) to 2.93 (moderately positive) in odds-ratio (OR) units; the median estimate was 1.31. The confidence intervals for many of the estimates overlap, which is expected because they are based on the same data."
[silberzahn2017, fig2, ref2]

?

"To identify the key variable(s) that separated these different distributions, we visualized the VoE by coloring each point on whether it included (or did not include) each one of the 13 adjustment variables in the model ... Specifically, models that contained the smoking adjustment variable (Fig. 3B, yellow points) had HR lower than the models without the smoking adjustment and lower -log10(P-values) (Fig. 3B, black points). One source of cadmium exposure includes smoking, and we concluded that the correlation between smoking and exposure to cadmium might be driving the multimodal behavior of VoE."
[patel2015, fig3, ref3]

"We observed three modes in the association between triglyceride levels and mortality (Fig. 3D-F). The multimodal plots indicated that total cholesterol and diabetes were driving these modes. For example, in models that did not contain these two adjustments, the associations had smaller P-values and a smaller range of HR. Furthermore, in models containing diabetes, HR were attenuated. The multimodal pattern seems reasonable in light of the high correlation between triglyceride levels and total cholesterol levels/risk for diabetes. ... "
[patel2015, fig3, ref4]

subjective validity of particular options / combination of options

"For most other variables, there was considerable ambiguity: The interaction seemed to be significant across about half of the arbitrary choice combinations. In these cases, the conclusion on the effect of fertility strongly depends on the evaluation of the different processing options. Both the authors performing the multiverse analysis and the readers of the research can construct arguments in favor or against certain choices, and the validity of these arguments will help drawing the conclusion. For example, if additional information suggests that the fifth option of assessing fertility is clearly superior, then Panel A of Figure 2 indicates that there is little evidence for an effect of fertility on religiosity in Study 2. On the other hand, if additional information suggests that the second option of assessing fertility is clearly superior, then most choice combinations lead to a significant interaction effect."
[steegen2016, fig2, ref7]

conditional robustness/interpretation

"In order to draw robust conclusions from these data, one must make a substantive judgment about two key modeling assumptions: the inclusion of race and marital status. None of the other model ingredients affect the basic conclusion. These two model assumptions determine the results."
[young2017, fig3, ref3]

reducing/changing the multiverse

"In general, deflating the multiverse involves developing a better and more complete theorizing of the constructs of interest and improving their measurement."
[steegen2016, fig2, ref9]

"Looking at the alternative options for assessing relationship status, it seems that the ambiguous Option 2 in the relationship status question could be formulated more precisely, so that relationship status assessment is no longer an arbitrary choice. This would have narrowed down the multiverses to 40 and 70 choice combinations in Study 1 and 2, respectively."
[steegen2016, fig2, ref9]

"A second approach for deflating the multiverse involves developing more complete and more precise theory in such a way that some options are theoretically superior than others, and it should be preferred when constructing data sets. ... The development of elaborated theories concerning these issues would narrow down the number of alternative options and deflate fluctuation."
[steegen2016, fig2, ref10]

summarize/reduce outcomes

Direct metrics (contrasts, summaries and robustness ratio)

"The multiverse analysis does not produce a single value summarizing the evidential value of the data, nor does it imply a threshold for an effect to reach to be declared robustly significant. Nevertheless, one might try to summarize the multiverse analysis more formally. One reasonable first step is to simply average the p values in the multiverse, in this case averaging all the numbers displayed in Figure 1 or 2. This mean value can be considered as the p value of a hypothetical preregistered study with conditions chosen at random among the possibilities in the multiverse and seems like a fair measurement in a setting where all of the possible data processing choices seem plausible (as in the example presented here, where the different options are drawn from other papers in the relevant literature)."
[steegen2016, fig1, ref5]

"As shown in Table 3, the average estimate across all of these models is 14.0. This simply represents the average coefficient across all models and is not necessarily the most theoretically defensible."
[young2017, fig1, ref4]

"The mean estimate from all models is 2.29 and the average sampling standard error is 1.61—indicating that the mean estimate is not statistically significant."
[young2017, fig2, ref4]

"Influence Analysis of the Gender Effect in Mortgage Lending: For the mortgage lending analysis, Table 6 shows the influence of control variables on the coefficient of interest (female). The Delta-Beta effect of controls is reported in order of absolute magnitude influence. To aid interpretation, we also report Delta-Beta as a percent change in the estimate from the mean of the modeling distribution (2.29 as in Table 7)."
[young2017, fig3, ref2]

"The average sampling standard error is 2.4, and the modeling standard error is 2.5—uncertainty about the estimate derives equally from the data and from the model. The combined total (sampling and modeling) standard error is 3.5.6 The robustness ratio—the mean estimate divided by the total standard error—is 4.05. By the standard of a t-test, this would be considered a strongly robust result, which agrees with the 100 percent sign stability and significance rates."
[young2017, fig1, ref4]

"The mean estimate from all models is 2.29 and the average sampling standard error is 1.61—indicating that the mean estimate is not statistically significant. In addition, the modeling standard error is 1.60—the estimates vary across models just as much as would be expected from drawing new samples. The total standard error—incorporating both sampling and modeling variance is 2.27, roughly the same size as the estimate itself, yielding a robustness ratio of 1.01."
[young2017, fig2, ref4]

" To aid interpretation, we also report Delta-Beta as a percent change in the estimate from the mean of the modeling distribution (2.29 as in Table 7)."
[young2017, fig3, ref2]

"The robustness ratio—the mean estimate divided by the total standard error—is 0.01. The modeling distribution is relatively normal: There are no critically important modeling decisions that generate bimodality in the estimates. As shown in Figure 4, the significant estimates reported in Table 7 above are extreme outliers in the modeling distribution."
[young2017, fig4, ref5]

comparing metrics across phenomena/variables of interest

"4.4. Summary of common patterns of the VoE: Figure 4 shows the distribution of the fold deviation of HR from the null (HR = 1.00), the -log10(P-value), RHR, and RP for all 417 variables considered."
[patel2015, fig4, ref2]

"Returning to the prototypical examples that we discussed previously, the RHR for vitamin D and thyroxine was moderate 1.14 (44th percentile) and 1.15 (51st percentile; Figs. 4C, 2A, and B). However, their RPs were among the largest and equal to 4.7 (93rd percentile) and 2.90 (84th percentile), respectively (Figs. 4D, 2A, and B). For urinary creatinine, the scenarios of adjustment had less prominent VoE. The RHR and RP for urinary creatinine was 1.07 (5th percentile) and 0.98 (47th percentile; Fig. 4C and D)."
[patel2015, fig4, ref3]

"The RHR for a-tocopherol (with the Janus effect) was higher (1.21, 71st percentile, Fig. 4C). Variables that demonstrated multimodality, such as serum cadmium and triglycerides, tended to have larger VoE."
[patel2015, fig4, ref4]

"We observed a weak correlation between RHR and RP (see Fig. S2/ Appendix at www.jclinepi.com, p = 0.09, P = 0.06)."
[patel2015, fig4, ref2]

Interpretation and Conclusions?