Influence of COVID-19 on disability and mortality: a multiverse analysis with post-selection inference

Influence of COVID-19 on Disability and Mortality: a Multiverse Analysis with Post-Selection Inference

Anna Vesely and Rossella Miglio

University of Bologna, Department of Statistical Sciences, Bologna, Italy
anna.vesely2@unibo.it

Abstract. Analytical flexibility in research can lead to variability in results and contribute to the replication crisis. Multiverse analysis addresses this issue by evaluating the robustness of findings across multiple plausible models. The Post-selection Inference approach to Multiverse Analysis (PIMA) further enhances this framework by providing valid statistical inference. In this study, we apply PIMA to assess whether SARS-CoV-2 infection increases the likelihood of disability onset, measured through the Global Activity Limitation Index, or mortality among initially non-disabled individuals. We construct a multiverse of models using longitudinal data from the Survey of Health, Ageing, and Retirement in Europe. Results highlight how significance strongly depends on specific modeling choices, emphasizing the value of PIMA in ensuring robust conclusions.

Keywords: COVID-19, GALI, multiverse analysis, multiple testing

1 Introduction

Analytical decisions in research often involve arbitrary choices, subjective judgments about the data-generating process, or multiple equally justifiable alternatives. This flexibility can be abused and has been one of the underlying causes of the replication crisis across various fields. In this context, multiverse analysis [1] provides a method to assess the robustness of findings across a range of plausible data-analytic options. Furthermore, the recent introduction of the Post-selection Inference approach to Multiverse Analysis (PIMA) [2] adds a general inferential framework to test whether one or more predictors of interest are associated with an outcome. The method combines information from multiple reasonable models and provides strong control of the family-wise error rate (FWER), allowing researchers to claim that a null hypothesis under study can be rejected for any model specification that shows a significant effect. This approach, based on a conditional resampling procedure, is general and flexible, as it applies to any generalized linear model (GLM) and a wide range of data-processing choices. Therefore, users can specify different models by choosing how to transform variables, which nuisance covariates and interactions to include, how to manage outliers and leverage points, etc.

In this study, we propose an application of the PIMA method in the analysis of Coronavirus Disease 2019 (COVID-19) outcomes. Specifically, we investigate whether SARS-CoV-2 infection increases the likelihood of disability onset – measured via a self-reported Global Activity Limitation Index (GALI) [3] – or mortality among initially non-disabled individuals. Our analysis uses longitudinal data from multiple waves of the Survey of Health, Ageing, and Retirement in Europe (SHARE) [4]. We account for a set of potential confounders, including geographical area, socio-demographic variables, and medical history; the multiverse arises from different specifications of these variables and the inclusion or exclusion of an interaction term with gender. We report multiplicity-adjusted p-values for each model of the multiverse, then discuss the key choices that lead to the greatest variability in results.

2 PIMA

The PIMA method [2] builds on the sign-flip score test, a resampling-based test that can be applied to any GLM [5, 6]. Let n independent observations of a response be denoted by $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, along with a predictor of interest $\mathbf{x} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n$ and m nuisance covariates $Z = (\mathbf{z}_1, \dots, \mathbf{z}_n)^\top \in \mathbb{R}^{n \times m}$. Consider a single GLM, where the mean $\mu_i = \mathbb{E}[y_i]$ depends on the covariates through the link function $g(\mu_i) = \eta_i = x_i\beta + \mathbf{z}_i\boldsymbol{\gamma}$. To test the null hypothesis $H_0 : \beta = 0$ against a two-sided alternative, the sign-flip score test relies on B random sign-flipping transformations. Each transformation $b = 1, \dots, B$ defines a diagonal sign-flipping matrix $F^b \in \mathbb{R}^{n \times n}$, where $F^1 = I$ is the identity, while the diagonal elements of the others are independently and uniformly drawn from $\{-1, 1\}$. The b-th permutation test statistic is

$$T^b = \frac{S^b}{\mathrm{sd}\left(S^b \mid F^b\right)} \qquad \text{where} \qquad S^b = \frac{1}{\sqrt{n}} \mathbf{x}^\top W^{1/2}(I - Q)V^{-1/2}F^b(\mathbf{y} - \hat{\boldsymbol{\mu}}).$$

Here $D = \mathrm{diag}\{\partial\mu_i/\partial\eta_i\}$, $V = \mathrm{diag}\{\mathrm{var}(y_i)\}$, $W = DV^{-1}D$,

$$Q = W^{1/2}Z(Z^\top WZ)^{-1}Z^\top W^{1/2}$$

is a particular hat matrix, and $\hat{\boldsymbol{\mu}}$ is a consistent estimate of the true mean computed under H_0. In practical applications, also D, V and the conditional standard deviation of S^b are replaced by consistent estimates. Thus, a p-value is

$$p = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left\{|T^b| \geq |T^1|\right\}.$$

Under lenient assumptions, the test is second-moment exact – for practical purposes, it provides exact control of the type I error rate even for small sample sizes. Moreover, it is robust to generic misspecifications of the variance V.

In the framework of multiverse analysis, interest lies in a set of K different models, each with a coefficient β_k and a corresponding null hypothesis H_{0k} :

$\beta_k = 0$ $(k = 1, \ldots, K)$. PIMA provides a robust and asymptotically exact test for the global null hypothesis $H_{\text{multiverse}} = \bigcap H_{0k}$. For each transformation b, the procedure computes the test statistics T_k^b for all models, using the same sign-flipping matrix, then combines them as

$$T^b_{\text{multiverse}} = \psi\left(|T_1^b|, \ldots, |T_K^b|\right)$$

where ψ is any function that is non-decreasing in each argument (e.g., the maximum). If the resulting global p-value is rejected at level α, we can state with confidence $1 - \alpha$ that the predictor has a significant effect in at least one model.

Furthermore, PIMA provides adjusted p-values for each model by applying the maxT method [7]. This ensures strong control of the FWER, allowing researchers to select a model post hoc from the multiverse without inflating the risk of false positives.

The same inference framework is trivially extended to the case of multiple parameters of interest, with a strategy similar to the extension from a single model to the multiverse. The method is implemented in the R package `pima`, available at https://github.com/livioivil/pima.

3 Application to SHARE Data

We investigate the impact of COVID-19 on health deterioration, exploring a multiverse of plausible models applied to the latest waves of the SHARE survey [4] (W8 - 2019/20, W9 - 2021/22). SHARE is a European longitudinal study that collects data on the health, economic, and social well-being of individuals aged 50 and older, as well as from those living with them. As the COVID-19 pandemic unfolded, the study adapted to assess its impact on the older population by launching two Corona-specific supplementary surveys (CS1 - June/August 2020, CS2 - June/August 2021).

The outcome of interest in our study is the onset of disability or death, with disability assessed through the GALI index [3], a self-reported measure of long standing activity limitations ("For the past six months at least, to what extent have you been limited because of a health problem in activities people usually do?"). From this, we define a binary response that indicates any limitation, regardless of severity, or death. As predictor, we consider a second binary variable that registers any COVID-19 symptom, positive test result, or hospitalization.

We select a population of subjects who had no limitations in W8, then measure the predictor in CS1 and the response in W9. The final dataset includes 11,193 subjects, of which 1.8% reported COVID-19-related events and 22.5% experienced a decline in health. The response is modeled using logistic regression with COVID-19 as regressor, along with several potential confounders: gender (55.9% female, 44.1% male), age (from 32 to 98, mean 67.7), geographical location, presence of a partner, education level, self-reported financial stress, and number of chronic illnesses (from 0 to 8, mean 1.1). Different combinations of confounder specifications create a multiverse of models.

First, age can enter the model as linear, quadratic, natural splines with three degrees of freedom, or cut into four classes (limits $30, 60, 70, 80, 100$, percentages 17.2%, 43.0%, 31.0% and 8.9%). The geographical location is specified by grouping the country of residence into either three macroareas or five areas. Macroareas are determined by the nature of the health care system: Bismark model (payroll-based social health insurance, 47.9%), Beveridge model (tax-funded national health service, 33.0%) and hybrid/transitioning models (19.2%). On the contrary, the areas are defined on a regional basis: West (28.3%), South (21.7%), East (19.6%), North (11.3%) and Baltic/Balkan (19.2%).

Having a spouse or partner in a registered relationship is coded as a binary variable, considering either all partners (74.1%) or only cohabiting partners (73.1%). The level of education can be coded in three classes (primary 13.0%, secondary 54.6%, post-secondary 32.4%) or as a binary variable (post-secondary). Similarly, financial stress may be considered in four classes (none 34.8%, minimal 32.2%, moderate 24.2%, severe 8.8%) or as binary (moderate or severe, 33.0%). The number of chronic illnesses is assumed to have a linear effect or coded as binary variable in two ways: at least one chronic illness (63.4%) or at least two (31.7%).

Finally, all models account for an interaction term between gender and each of the other nuisance covariates, while the interaction term between COVID-19 and gender may either be omitted or included. If included, the coefficients of COVID-19 and this interaction represent the effect of COVID-19 for males and the additional effect for females, respectively. In this case, the statistical significance of both terms is evaluated.

We obtain 384 models, half of which also test the interaction effect, resulting in a total of 576 statistical tests. These are evaluated using PIMA with $B = 10{,}000$ transformations and a significance level $\alpha = 0.05$.

4 Results

The computational time is approximately 2 hours, which can be substantially reduced with fewer transformations, at the cost of more variable results and a potential loss in power.

The primary data-analytic choice influencing the statistical significance of the effect of interest is the inclusion of the interaction term between COVID-19 and gender, which generally leads to less precise estimates and higher standard errors. The coefficient of COVID-19 is significant in $64/384 = 16.7\%$ models (global p-value 0.0404), all of which do not consider the interaction term. When present, the interaction effect is never significant, even before correction for multiplicity (global p-value 0.7307).

Since models that include the interaction term never exhibit significant effects, we will focus on those that exclude it. In these models, the mean estimate of the effect of COVID-19 is 0.3858 (see Fig. 1), corresponding to an almost 50% increase in the odds of health decline for subjects with COVID-19-related events.

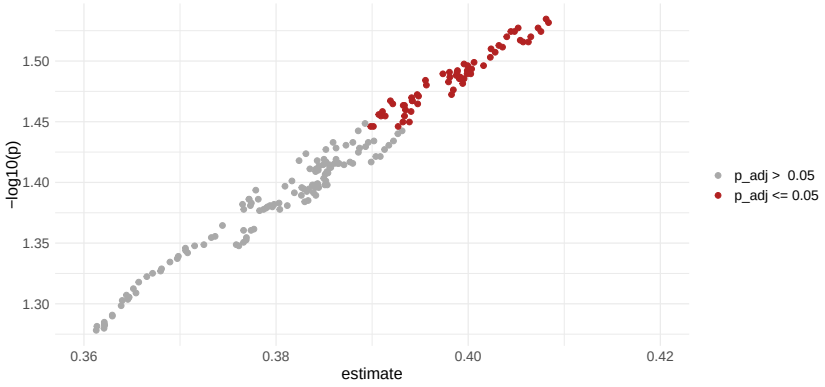

Fig. 1. Effect of COVID-19 in models without the interaction between COVID-19 and gender: estimate of the coefficient and raw p-value ($-\log_{10}$). Red points correspond to significant results after correction for multiplicity (adjusted p-value ≤ 0.05).

The specifications of age and the number of chronic illnesses have a substantial impact on the effect's significance (Fig. 2). In particular, models with age grouped into classes are never significant, while the other three specifications appear less influencial. Grouping likely leads to a loss of information and assumes a step-wise relationship between age and the outcome, whereas non-linear effects are better captured by more flexible transformations. Besides, models with the number of chronic illnesses are never significant, while those accounting for at least two chronic illnesses are most often significant. These differences may reflect variations in the types of chronic illnesses, with some being more common and milder (e.g., hypertension and high cholesterol).

Finally, the geographical level has a moderate influence, with models including areas being less often significant than models including macroareas based on the health care system.

These patterns suggest that significant effects arise only from specific modeling choices. In general, researchers should be aware that some results may be due to inadequate modeling, which can induce spurious correlation between the predictor and the response. Therefore, every significant test must be evaluated with care. This underscores the importance of the multiverse approach and the inferential framework provided by PIMA in gaining deeper insights from the analysis.

Acknowledgements

This research was co-funded by the Italian Complementary National Plan PNC-I.1 "Research initiatives for innovative technologies and pathways in the health and welfare sector" D.D. 931 of 06/06/2022, "DARE - DigitAl lifelong pRevEntion" initiative, code PNC0000002, CUP: B53C22006450001.

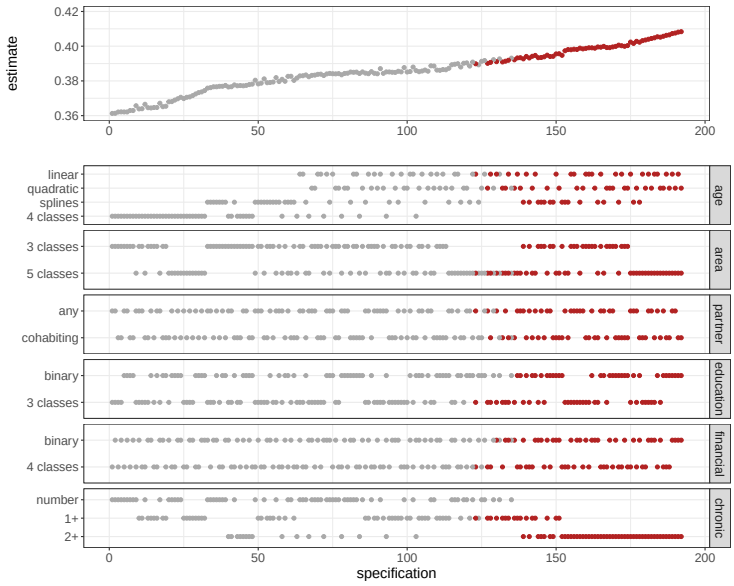

Fig. 2. Effect of COVID-19 in models without the interaction between COVID-19 and gender: model index and estimate of the coefficient (top) and specification of the covariates (bottom). Red points correspond to significant results after correction for multiplicity (adjusted p-value ≤ 0.05).

References

1. Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W.: Increasing transparency through a multiverse analysis. Perspect. Psychol. Sci. 11(5), 702–712 (2016)
2. Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagnì, A., Finos, L.: Post-selection Inference in Multiverse Analysis (PIMA): An inferential framework based on the sign flipping score test. Psychometrika 89(2), 542-568 (2024)
3. Robine, J.M., Jagger, C.: Creating a coherent set of indicators to monitor health across Europe. Eur. J. Public Health 13(1), 6–14 (2003)
4. Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., Schaan, B., Stuck, S., Zuber, S.: Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). Int. J. Epidemiol. 42(4), 992–1001 (2013)
5. Hemerik, J., Goeman, J.J., Finos, L.: Robust testing in generalized linear models by sign flipping score contributions. J. R. Stat. Soc. Ser. B Stat. Method. 82(3), 841–864 (2020)
6. De Santis, R., Goeman, J.J, Hemerik, J., Davenport, S., Finos, L.: Inference in generalized linear models with robustness to misspecified variances. Pre-print (2024)
7. Westfall, P.H., Young, S.S.: Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley, New York (1993)