EJW
ECON JOURNAL WATCH
Scholarly Comments on Academic Economics

ECON JOURNAL WATCH 18(2)
September 2021: 212–219

Response to "Mortality and Science: A Comment on Two Articles on the Effects of Health Insurance on Mortality"

Jacob Goldin[1], Ithai Lurie[2], and Janet McCubbin[3]

LINK TO ABSTRACT

In Jacob Goldin, Ithai Lurie, and Janet McCubbin (2021—hereafter referred to as GLM), we studied the effect of a randomized outreach experiment to uninsured taxpayers by the Internal Revenue Service (IRS). We found that the intervention increased health insurance coverage and reduced middle-age mortality in subsequent years. Robert Kaestner (2021) criticizes our findings as well as a recent study by Sarah Miller, Norman Johnson, and Laura R. Wherry (2021), which provided evidence that state Medicaid expansions caused reductions in mortality among low-income adults. In neither case does Kaestner provide new empirical evidence or allege an error in analysis—his concerns relate solely to the interpretation and convincingness of the presented results. Here, we briefly respond to Kaestner's criticisms of GLM. We conclude that his concerns do not undermine our main empirical finding, which is that the coverage induced by the IRS outreach reduced mortality.

1. Stanford University, Stanford, CA 94305; National Bureau of Economic Research, Cambridge, MA 02138.
2. Office of Tax Analysis, U.S. Department of the Treasury, Washington, DC 20220.
3. Office of Tax Analysis, U.S. Department of the Treasury, Washington, DC 20220. For excellent research assistance, we are grateful to Vedant Vohra. The views presented here are those of the authors and do not necessarily reflect the Treasury Department or any government agency.

Statistical power

Kaestner's main argument is that the study design in GLM was underpowered and therefore that the results of the analysis, notwithstanding their statistical significance, are likely to be random noise. We find Kaestner's critique unpersuasive for three reasons. First, at reasonable parameter values, the study has more power than he alleges. Second, at plausible power levels, our results continue to support our main qualitative conclusion, that the intervention reduced mortality. Third, several features of our results are consistent with the hypothesis that the intervention reduced mortality but would be quite unlikely to arise if the difference in mortality between the treatment and control groups was due solely to statistical chance.

Like Kaestner, we agree that ex-post power analyses along the lines suggested by Andrew Gelman and John Carlin (2014) can be useful for interpreting a study's results. However, to be informative, such analyses require as inputs accurate parameters to describe the study's expected true effect. In our setting, there are two main challenges to obtaining these parameters (we describe both challenges more fully in GLM when interpreting our results, in section V.C and V.D). First, we do not know the baseline mortality rate for the compliers, the group who obtained additional coverage in response to the intervention. As we described in the paper, and as others have suggested, there is reason to expect that those who respond to interventions like the one we study may be particularly likely to benefit from coverage, such that their mortality rate absent coverage is higher than for the overall population. In fact, we provided suggestive evidence that the baseline mortality rate of the compliers was twice as high as those who remained uninsured after receiving the letter (see Appendix Table A.XXVI).[4] The larger the baseline mortality rate among the compliers, the larger the absolute difference in mortality that our observed coverage effect would imply, and hence, the higher the power of the study design.

The second important unknown input into the power analysis in our setting is that we don't know whether the additional months of coverage induced by the intervention were concentrated among the same individuals or dispersed across a larger number of individuals. For example, we find that the intervention increased coverage by 0.36 months among the middle-age uninsured, which is consistent with anywhere between 2 percent and 29 percent of that population enrolling in additional coverage because of the intervention (Appendix Table A.VI).

4. Unless otherwise noted, all tables and figures refer to GLM or to its appendix (found **here**).

Understanding the number of compliers over which the treatment effect is spread is important for assessing reasonable magnitudes for the effect sizes; for example, we probably wouldn't expect someone who added 12 months of coverage to experience a 12*7%=84% reduction in mortality, but it strikes us as quite reasonable that someone who added 3 additional months of coverage would experience a 3*7%=21% reduction in mortality from it, especially since they could have concentrated their utilization of heath care services during that period (Diamond et al. 2018). This is why it is incorrect for Kaestner to assert that only assumed mortality effects of 3 percent or less are reasonable.

Because we do not feel confident about the true expected mortality difference between the treatment and control groups, the analysis presented in Appendix Figure A.V considered a range of parameter value combinations.[5] Kaestner takes a different, hybrid approach. He assumes the intervention entirely reduced mortality (pushing towards high power) but also assumes (1) the compliers' baseline mortality rate is the same as the overall sample population (pushing towards lower power relative to our expectation) and (2) that the first-stage effect was concentrated among minimal individuals; in other words, he is assuming that health insurance makes people immortal, but there would be relatively few people made immortal and they wouldn't be people at particularly high risk of dying.

Gelman and Carlin (2014) highlight two concerns that can arise in studies that lack sufficient statistical power: obtaining a point estimate that exceeds that magnitude of the true effect (an "M error") and estimating an effect that has the wrong sign (an "S error"). Because our main qualitative conclusion concerns the direction of the effect—i.e., that the intervention reduced mortality—we focus on the likelihood that our findings embodies an S error.[6] To assess the likelihood of an S error, we follow Gelman and Carlin and calculate the probability of obtaining a statistically significant result of incorrect sign, conditional on obtaining a statistically significant difference in the treatment and control group means, using the simulation approach described in Appendix Table V. To reflect our uncertainty

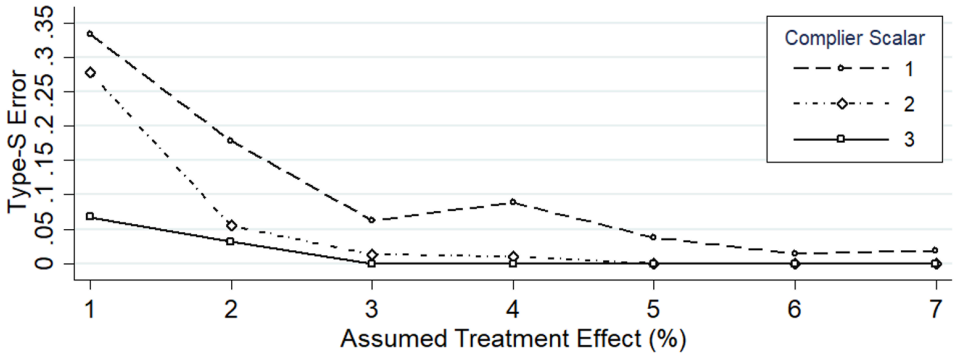
5. Notably, our purpose in conducting this analysis was not to assess the level of power for the study (the intervention had already been conducted). Rather, it was to select the age range for the analysis in a principled way. Note also that the baseline mortality scalar for the compliers was designed to reflect not only the difference between the compliers and the rest of the sample, but also the fact that the sample (selected to be partially or fully uninsured during prior years, and therefore on average lower income and facing greater instability) was likely to have higher mortality than the overall 45–64 population for the United States (the mortality rate used as the input into the analysis). See the note to Appendix Figure A.V for additional detail.

6. In contrast, we think it is quite possible that due to chance, our point estimate exceeds the magnitude of the true effect of the intervention on mortality. As we wrote in the introduction to GLM, "we view the effects at the lower magnitude end of our confidence interval as most plausible" (p. 3).

about the true effect size, we again consider a range of parameter values for the effect of an additional coverage-month on mortality (1 percent through 7 percent) as well as the baseline mortality of the compliers (1, 2, or 3 times the 45–64-year-old U.S. population average).

The results of this exercise are displayed in Figure 1 here. Across parameter combinations, the probability of an S error does not rise above one-third; hence, even in that scenario, it is incorrect to conclude, as Kaestner does, that our results do not provide informative evidence about the direction of the effect of health insurance on mortality. And for the parameter combinations we consider more realistic, the probability of an S error is quite low. For example, with a complier baseline mortality multiplier of 2 (our best guess, based on Appendix Table A.XXVI), the probability of an S error is approximately 5 percent or below for assumed treatment effects of 2 percent or more.

Figure 1. Probability of sign error by assumed treatment effect



Notes: The figure displays the probability of an S error at various assumed effect sizes of health insurance coverage on mortality (x-axis) and various baseline mortality rates for the set of individuals who enroll in one month or more of additional coverage because of the intervention (the compliers). An S error is defined as detecting a difference in the treatment and control group mean mortality rates that is statistically significant at the 5-percent confidence level with sign that is the opposite of the true (assumed) effect, conditional on detecting any difference between the treatment and control group mean mortality rates that is statistically significant at the 5-percent confidence level. Results are based on simulations with N=1000 random draws of treatment and control populations. The baseline mortality rate for compliers (i.e., the mortality rate absent any additional coverage from the intervention) is alternatively assumed equal to one, two, or three times the overall mortality rate of the overall U.S. population of 45–64 year-olds. General population mortality is estimated from population-level mortality rates for 2016 from the Social Security Death Index among individuals alive at the end of 2015. The code used to produce this figure can be downloaded from the journal website (**link**).

We see a similar story when we adopt Kaestner's approach and assume that there are relatively few compliers and that all of those compliers become immortal. Assuming the minimum share of compliers consistent with the data (1.9 percent, reported in Appendix Table A.VI) and a mortality reduction for that group equal to their baseline mortality rate, we also observe low probabilities of an S error. For example, using the baseline mortality rate for this group estimated in Appendix Table A.XXVI implies an S error occurs with probability less than 1 percent.

The third reason we do not think Kaestner's arguments about statistical power undermine our main qualitative finding is that several features of our results are more consistent with the difference in treatment and control group means arising from the intervention as opposed to arising from statistical noise. Most important is the timing of the observed treatment effect. As shown in Figure III, cumulative mortality rates for the treatment and control groups are very similar prior to the intervention, begin to diverge around the time of the intervention, and continue to diverge as the time since the intervention elapses. This is the pattern that one would expect if the intervention induced a persistent difference in coverage between the groups that was effective at extending longevity. In contrast, there would be no reason to expect that the difference in mortality rates would emerge only after the intervention if the difference was simply due to random chance. Similarly, the fact that the difference in mortality rates is concentrated among the subset of households in the treatment group that increased coverage in response to the intervention and not among those that did not (compare Figure III with Figure A.VIII) is what one would expect if the observed difference in means was due to the intervention but not if the difference was due to statistical noise.

In summary, although we of course cannot definitively rule out the possibility that the statistically significant difference in mortality rates we observe is entirely due to chance, accounting for the study's statistical power does not suggest that possibility is particularly likely and several features of the observed results lead us to conclude that this possibility is fairly remote.

External validity

Apart from his concerns about statistical power, Kaestner (2021, 209) writes that the results in GLM have "virtually no external validity" because the results are identified from approximately 21,000 individuals induced to obtain insurance because of the intervention, and this group may differ from the rest of the uninsured population in relevant ways.[7] As Kaestner acknowledges, we emphasize

7. In fact, as discussed in GLM, the estimated mortality effect is identified from the set of individuals who

this point in the paper, noting that "the effect of coverage on mortality may be particularly large among the individuals induced into coverage from the intervention we study, as compared with other policies that reduce insurance" (GLM, 41). Our results, therefore, are perhaps most informative for individuals like the ones we study—i.e., those who already had access to coverage but for whom misperceptions or other frictions prevented take-up. As we noted in the paper, the effect of coverage for this group "is particularly policy relevant, since such individuals' coverage decisions can be shaped through outreach efforts of the type commonly employed by governments and non-profits" (GLM, 4). Collectively, these outreach efforts cost hundreds of millions of dollars each year;[8] we think it is important to understand the effects of the coverage they induce. In contrast, if one was interested in predicting the effect of coverage induced by a state expanding Medicaid, the estimates presented in Miller et al. (2021) may be a better place to start.

Variability in point estimates for subgroups

Finally, Kaestner (2021, 200) casts doubt on the validity of the findings in GLM because the reported point estimates for various subgroups are "all over the place." Again, we find this interpretation of results unpersuasive—in our view, he is reading too much into statistically insignificant differences among subgroups. Indeed, in GLM we characterized these analyses as "exploratory" (GLM, 36).[9] In addition, when these subgroups are analyzed collectively, we find that the overall pattern is that groups that increased their coverage by higher amounts experienced higher reductions in mortality—as one would predict if the observed mortality reduction was due to the intervention—although whether this trend is statistically significant varies depending on which subgroups are considered (as discussed on pages 28–29 of GLM). Notably, some of the estimates presented by Kaestner as evidence against the study's validity actually strike us as supporting the interpretation that the intervention reduced mortality. For example, in considering

would increase coverage in response to the intervention in the event that they receive an outreach letter. As reported in Appendix Table A.VI, this group represents between 1.9 percent and 29 percent of the middle-age sample population, or between approximately 26,000 and 394,000 individuals.

8. Federal funding for ACA-related outreach has fluctuated in recent years, with the budgeted amount for ACA advertising declining from $100 million in 2017 to $10 million in 2018. However, states and non-profits have continued to fund coverage-related outreach efforts at high levels—e.g., California's 2018 budget for Navigator programs was $111 million (Seervai 2017).

9. One reason we chose to report these analyses was that though imprecise, they could usefully be combined with results from other studies to paint a clearer picture.

the analysis, we were initially concerned that the observed mortality reduction may have been driven by a single, outlier treatment arm; we were reassured by the fact (highlighted in Kaestner's Table 3) that *each* treatment arm was associated with a reduction in mortality. Similarly, Kaestner (2021, 203) describes the growing divergence in cumulative mortality rates between the treatment and control groups as an "anomaly," but this pattern is exactly what one would expect to observe if the intervention-induced coverage was driving the mortality difference, since the treatment group continued to enroll in coverage at higher rates than the control throughout the post-treatment period.

Conclusion

Kaestner (2021) criticizes the findings in GLM. As discussed above, we find these criticisms unpersuasive. We continue to believe that the best interpretation of the evidence reported in GLM is that the coverage induced by the IRS intervention reduced mortality.

References

Diamond, Rebecca, Michael J. Dickstein, Timothy McQuade, and Petra Persson. 2018. Take-Up, Drop-Out, and Spending in ACA Marketplaces. *NBER Working Paper* 24668. National Bureau of Economic Research (Cambridge, Mass.). **Link**

Gelman, Andrew, and John Carlin. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* 9(6): 641–651.

Goldin, Jacob, Ithai Z. Lurie, and Janet McCubbin. 2021 (GLM). Health Insurance and Mortality: Experimental Evidence from Taxpayer Outreach. *Quarterly Journal of Economics* 136(1): 1–49. **Link**

Kaestner, Robert. 2021. Mortality and Science: A Comment on Two Articles on the Effects of Health Insurance on Mortality. *Econ Journal Watch* 18(2): 192–211. **Link**

Miller, Sarah, Norman Johnson, and Laura R. Wherry. 2021. Medicaid and Mortality: New Evidence from Linked Survey and Administrative Data. *Quarterly Journal of Economics* 136(3): 1783–1829. **Link**

Seervai, Shanoor. 2017. Cuts to the ACA's Outreach Budget Will Make It Harder for People to Enroll. October 11. Commonwealth Fund (New York). **Link**

About the Authors

Jacob Goldin is an associate professor of law at Stanford Law School. Trained as a lawyer and economist, his research focuses on the taxation of low-income households, health policy, and the application of behavioral economics to program design. Prior to joining Stanford, Goldin worked in the Office of Tax Policy at the U.S. Department of the Treasury and clerked for Judge Richard Posner of the Seventh Circuit Court of Appeals. He holds a J.D. from Yale Law School and a Ph.D. in economics from Princeton University. His email address is jsgoldin@law.stanford.edu.

Ithai Lurie is employed as a financial economist at the Office of Tax Analysis of the U.S. Department of the Treasury. He received his Ph.D. from Northwestern in 2006. His current research focuses on the effects of public intervention through taxes or regulation on consumers' behavior. His work has been published in the *Journal of Health Economics*, *Journal of Public Economics*, *Quarterly Journal of Economics*, and *Review of Economic Studies*. His recent work includes: the long-run effect of Medicaid expansions, a look at the effect of the individual penalty on coverage, and an analysis of Income Response to the Affordable Care Act Premium Tax Credit. His email address is Ithai.Lurie@treasury.gov.

Janet McCubbin specializes in tax administration and tax policy affecting workers and families. She served as Chief, Special Studies Branch, Internal Revenue Service; Director of Economic Security, AARP Public Policy Institute; and Director of the Office of Tax Analysis, U.S. Department of the Treasury. She earned her Ph.D. in Economics at the University of Maryland. Her email address is Janet.McCubbin@treasury.gov.

Go to archive of Comments section
Go to September 2021 issue