Replication of a Research Claim from Kim & Radoias (2016), from Social Science & Medicine

Replication Team: Anna Abatayo, Andrew Tyner, and Esteban Méndez-Chacón

Center for Open Science, Charlottesville, VA & Central Bank of Costa Rica

Research Scientist: Andrew Tyner

Action Editor: Rich Lucas

Final Report

September 30, 2020

Replication of a Research Claim from Kim & Radoias (2016), from Social Science & Medicine

Claim Summary

The claim selected for replication from Kim & Radoias (2016) is that, for the specific case of asymptomatic disease detection, education should have a clear positive effect for individuals in poor health status (and at the aggregate level), and a smaller (possibly zero or negative) effect for individuals in good health status; the positive effect for individuals in poor health status is the portion of this claim selected for the SCORE program. This reflects the following statement from the paper's abstract: "In terms of disease detection, more educated respondents have a higher probability of being diagnosed, but only conditional on being in poor general health." Evidence in support of the claim is found in Table 2 (Kim & Radoias, 2016, p. 19), which contains the probit regression results for the determinants of hypertension under-diagnosis. The dependent variable is a dummy equal to one for those respondents who were found to be hypertensive during the IFLS [Indonesian Family Life Survey] screenings but were not previously diagnosed by a doctor. The three separate columns represent, in order, the results for the entire sample, the results for the subsample consisting of respondents in good general health, and the results for the subsample consisting of respondents in poor general health. For the SCORE program, the analysis of the subsample consisting of respondents in poor general health is selected. The predictor of interest is Years of Education (see the right column of Table 2 for details of the model). Education matters for these people, as more educated persons generally have higher opportunity costs of feeling sick and hence value their health higher, which pushes them harder to look for a cure in a doctor’s office.

Focal hypothesis H*: Among the sample of respondents in poor general health who were found to be hypertensive during a screening, the probability of being undiagnosed decreases with education.

Replication Criteria

Criteria for a successful replication attempt for the SCORE project is a statistically significant effect (alpha = .05, two tailed) in the same pattern as the original study on the focal hypothesis test (H*).

Replication Result

Table R.1 contains the marginal effects of the probit regression for the determinants of hypertension under-diagnosis for the subsample consisting of respondents in poor general health. Column (1) of Table R.1 shows the results without the distance to health care variable as the final test for the focal hypothesis. **The result suggests that an additional year of schooling reduces the probability of being under-diagnosed by 0.00733, and the effect is statistically significant at the 5% level (p = 0.021).** The replication study prioritizes the probit regression without the DISTANCE variable as the final test for the focal hypothesis. **Thus, this replication of the claim was successful according to the SCORE criteria.** The analytic sample included 673 observations, which did not meet the minimum threshold of 970 observations defined by the power analysis.

Similarly, Column (2) of Table R.1 reports the marginal effects of the probit regression including distance as a control variable. The result suggests that an additional year of schooling reduces the probability of being under-diagnosed by 0.01783, and the effect is statistically significant at the 1% level (p = 0.002).

Table R.1. Hypertension under-diagnosis probit regression for respondents in poor health (Dependent variable: The probability of being under-diagnosed).

	(1)	(2)
Years of Education	**-0.00733** **(0.00317) ****	**-0.01783** **(0.00563)*****
Log PCE	0.01423 (0.01704)	0.02896 (0.03006)
Time Preference	0.00545 (0.01681)	0.03174 (0.03055)
Risk Preference	-0.02020 (0.01670)	0.00947 (0.03272)
Female	-0.02926 (0.02618)	-0.07179 (0.06023)
Distance to Health Center		0.00125 (0.00152)
Sample Size	**673**	**164**

The table reports marginal effects with standard errors in parentheses. Respondent's age and age squared are included in all regressions. Significance levels: *-significant at 10% level **-significant at 5% level ***-significant at 1% level.

Methods & Materials

The following materials are publicly available on the OSF site:

- The **preregistration** file: Kim_SocSciMed_2016_AqDO_7945 (COS_Méndez-Chacón) Preregistration.pdf
- The **IFLS5** documentation that includes codebooks, user guides, and information on the survey. Filenames:
    - IFLS5_User_Guide_Vol_1.pdf
    - IFLS5_User_Guide_Vol_2.pdf
    - RAND_WR1143z2.pdf
    - RAND_WR1143z3.pdf
    - risk_time_preferences_logic.pdf
    - ifls2014_hhd_B1.txt
    - ifls2014_hhd_B2.txt
    - ifls2014_hhd_B3A.txt
    - ifls2014_hhd_B3B.txt
    - ifls2014_hhd_B4.txt
    - ifls2014_hhd_B5.txt
    - ifls2014_hhd_BEK.txt
    - ifls2014_hhd_BK.txt
    - ifls2014_hhd_BT.txt
    - ifls2014_hhd_BUS.txt
    - ifls2014_hhd_EK.txt
    - ifls2014_hhd_FE.txt

    - ifls2014_hhd_HTRACK.txt
    - ifls2014_hhd_PTRACK.txt
- The **R code** to produce the replication sample, and details on its use. Filenames:
    - README.txt
    - kim_7945_data_prep.R
- The R code also includes a function that summarizes the variables in each dataset with a set of key descriptives, to verify that the code is producing the intended dataset. Filenames:
    - rep_data_test.tsv
    - weights_data_test.tsv
    - alt_ed_data_test.tsv
- The **raw data** file and the **analytic data** cannot be uploaded directly to OSF, because accessing IFLS5 data requires registration on the RAND website [https://www.rand.org/well-being/social-and-behavioral-policy/data/FLS/IFLS/access.html]. Two of the relevant terms included with this registration are as follows: "You will not distribute the IFLS Public Use Data files to others. If you plan to work with other people using these data, you will ask them to register or register them yourself. If you are a data librarian, you will ask users to register if they obtain a copy of the data from you...You will acknowledge the IFLS as the source of the data for analysis in all reports and publications based on these data. Desired citations for each wave are found on the IFLS data download page." Approval of the registration is almost instantaneous. A link to access the data is sent soon after registration.
- Three **data dictionaries** for every dataset, provided as tsv files. Filenames:
    - main_dataset_dictionary.tsv
    - weights_dictionary.tsv
    - alternative_education_dictionary.tsv

- The **code for replication.** Along with the analytical sample, this is the only file required to replicate the original study. To replicate, just change the working directory to where the data is in your computer and run this file using Stata (the code was written using Stata 15.1). Filename:
    - Kim & Radoias 2016 - Replication Analysis.do
- The **output** from the Stata analyses, available in two formats: smcl (Stata output) and a pdf file. Filenames:
    - Kim-Radoias_Replication.smcl
    - Kim-Radoias_Replication.pdf

Deviations from the Original Study

1. The original study relied on data from the fourth wave of the Indonesian Family Life Survey (IFLS), while the replication study uses the fifth wave.
2. The distance to the health center was used as an explanatory variable by Kim & Radoias in the main specification. However, in the replication sample, the variable contains serious limitations. For example, this was only asked of respondents who had visited a medical provider in the last four weeks. Further, only respondents who knew the distance have a value for this variable. Due to the issues with the distance variable, two probit regressions are estimated. The replication study prioritizes the probit regression without the distance variable as the final test for the focal hypothesis. The reason is to reduce the limitations arising from the measurement of the distance variable. However, as Table R.1 shows, the distance variable does not affect the main results of the replication analysis: among the sample of respondents in poor general health who were found to be hypertensive during a screening, the probability of being undiagnosed decreases with education.

Citation

Kim, Y., & Radoias, V. (2016). Education, individual time preferences, and asymptomatic disease detection. *Social Science & Medicine*, 150, 15–22. https://doi.org/10.1016/j.socscimed.2015.11.051

Strauss, J., F. Witoelar, and B. Sikoki. "The Fifth Wave of the Indonesia Family Life Survey (IFLS5): Overview and Field Report". March 2016. WR-1143/1-NIA/NICHD.