MSR192

Detecting Potentially Fraudulent Data in Online Discrete Choice Experiments (DCE)

A New Method Using Behaviourally 'Irrelevant' Respondent Variables

C. ISKIWITCH[1], B. WHITE[1], L. v BUTLER[1], L. PANATTONI[2], J. COULTER[3], N. PROOD[2], G. GAHLON[2], N. LAND[2], and M. MARAVIC[2]

[1] SurveyEngine GmbH, Berlin, Germany [2] Precision AQ, New York, NY, USA [3] Pfizer, New York, NY, USA

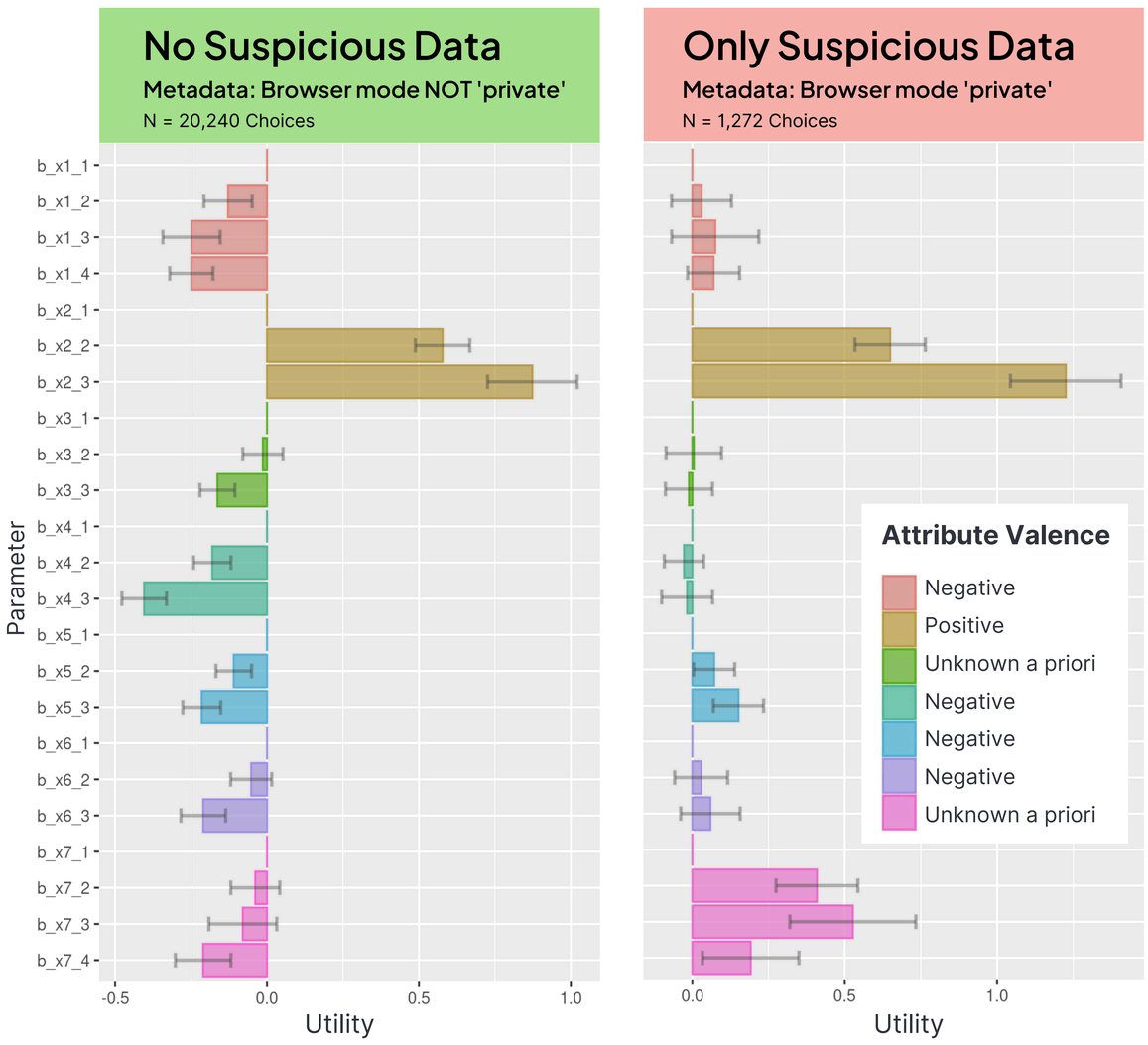
While analyzing data from a multi-country health-focused DCE, we noticed that one country's data had abnormally high fail rates on check questions (dominance and repeat tests). We began an in-depth investigation of our data, including metadata markers. Our aims were to identify problematic data sources, to develop a defensible data exclusion plan, and, ultimately, **to develop a systematic and repeatable approach to identifying potentially fraudulent data in online DCEs.**

Method

We assume that the preferences of legitimate respondents should not vary by a set of behaviourally 'irrelevant' respondent variables such as network characteristics, browser mode, or membership to an arbitrary survey start time period.

Systematic generation of joint multinomial logit (MNL) preference models segmented by these 'irrelevant variables' and their Likelihood Ratio (LR) calculation was performed with the assumption that no segmented models should differ statistically from the aggregate. We hypothesised that significance in the LR test, where there was no legitimate cause, may identify potential fraud in the segmented sample. This method allowed a wide, generic and systematic search for anomalous behaviour and the possibility to trace that behaviour back to its source. Practical implementation of this method occurred during data collection in early 2024, which targeted a final sample size of 450 patients.

Divergent preference models suggest fraud

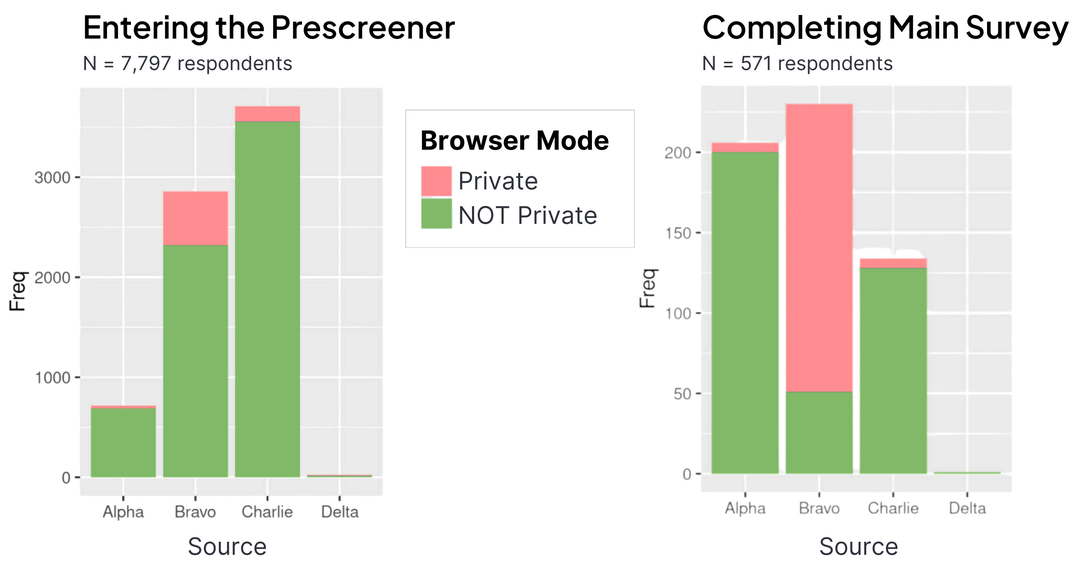


Results

Among the variables examined was a metadata marker indicating whether the respondent's browser was set to private browsing mode. Results of a likelihood ratio (LR) test of DCE data segmenting on this 'irrelevant' variable indicated that those using private browsing mode were behaviourally different from those not using it, $LR(16) = 65.2$, $p < .001$). Visual inspection further suggested the preference weights among the potentially fraudulent data were disordered (see the preference model figures to the left).

With these results, we compared our three recruitment sources on this suspicious variable. We found that source "Bravo" had significantly higher use of private browsing mode beginning the screener, as well as in rates of passing the screener and completing the main survey (see the "Supporting Analysis" figures). Further supporting suspicion of source Bravo, we found an increase over time in the screener passing rates, suggesting respondents from source Bravo used previous knowledge of the screening criteria.

Supporting Analysis



Discussion

We propose that fraudulent responders often leave at least one data trace behind, and, like a crime scene forensic investigation, these can be identified through a systematic approach. Identifying a list of behaviourally 'irrelevant' variables, and segmenting MNL preference models on that list, allowed us to identify the provenance and method of suspected fraudulent survey participation. We repeated this approach with data from a second DCE and found evidence suggesting that other variables may be able to identify fraud. This approach can be part of a clear and justifiable a priori data exclusion plan. The results of this kind of analysis can be an early warning for potential fraud. They can also be used as corroborating evidence alongside abnormalities in, for example, incidence rates or diurnal activity patterns.

Future research should validate this method and examine how it compares to latent class analysis and other proposed approaches for detecting fraudulent data. More generally, preference study protocols and reports should describe the use and success of fraud detection methods.

Contact and Acknowledgement

Carol Iskiwitch, SurveyEngine: Carol@surveyengine.com
Laura Panattoni, Precision AQ: Laura.Panattoni@precisionaq.com
Josh Coulter, Pfizer: Joshua.Coulter@pfizer.com

Thanks to Stephane Hess for contributing to the method

survey engine PRECISION AQ™