RESEARCH ARTICLE

REVISED The relative incidence of COVID-19 in healthcare workers versus non-healthcare workers: evidence from a web-based survey of Facebook users in the United States [version 2; peer review: 1 approved, 1 approved with reservations, 1 not approved]

Abraham D. Flaxman[1], Daniel J. Henning[2], Herbert C. Duber[1,2]

[1]Institute for Health Metrics and Evaluation, University of Washigton, Seattle, WA, 98195, USA
[2]Department of Emergency Medicine, University of Washigton, Seattle, WA, 98195, USA

v2 **First published:** 27 Nov 2020, **4**:174
https://doi.org/10.12688/gatesopenres.13202.1
Latest published: 20 Jul 2021, **4**:174
https://doi.org/10.12688/gatesopenres.13202.2

Abstract

Background: Healthcare workers are at the forefront of the COVID-19 pandemic and it is essential to monitor the relative incidence rate of this group, as compared to workers in other occupations. This study aimed to produce estimates of the relative incidence ratio between healthcare workers and workers in non-healthcare occupations.
Methods: Analysis of cross-sectional data from a daily, web-based survey of 1,822,662 Facebook users from September 8, 2020 to October 20, 2020. Participants were Facebook users in the United States aged 18 and above who were tested for COVID-19 because of an employer or school requirement in the past 14 days. The exposure variable was a self-reported history of working in healthcare in the past four weeks and the main outcome was a self-reported positive test for COVID-19.
Results: On October 20, 2020, in the United States, there was a relative COVID-19 incidence ratio of 0.73 (95% UI 0.68 to 0.80) between healthcare workers and workers in non-healthcare occupations.
Conclusions: In fall of 2020, in the United States, healthcare workers likely had a lower COVID-19 incidence rate than workers in non-healthcare occupations.

Keywords

COVID-19, healthcare workers

Open Peer Review

Reviewer Status ✔ ✘ ?

	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 20 Jul 2021	✔ report		
version 1 27 Nov 2020	? report	✘ report	? report

1. **Alex Reinhart**, Carnegie Mellon University, Pittsburgh, USA

2. **Devan Hawkins**, MCPHS University, Boston, USA
Marcy Goldstein-Gelb, National Council for Occupational Safety and Health, Somerville, USA

3. **Tim Driscoll**, The University of Sydney, Sydney, Australia

Any reports and responses or comments on the

article can be found at the end of the article.

This article is included in the Coronavirus (COVID-19) collection.

Corresponding author: Abraham D. Flaxman (abie@uw.edu)

Author roles: Flaxman AD: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Henning DJ**: Methodology, Writing – Review & Editing; **Duber HC**: Methodology, Writing – Review & Editing

Competing interests: ADF has consulted recently for Janssen; SwissRe; Sanofi; Merck for Mothers; and Agathos, Ltd. DJH has received research funding from Baxter and performed consulting services for Cytovale. HCD has no competing interests to disclose.

Grant information: This work was supported by the Bill and Melinda Gates Foundation [OPP1170133] and the National Science Foundation [DMS-1839116].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Flaxman AD, Henning DJ and Duber HC. **The relative incidence of COVID-19 in healthcare workers versus non-healthcare workers: evidence from a web-based survey of Facebook users in the United States [version 2; peer review: 1 approved, 1 approved with reservations, 1 not approved]** Gates Open Research 2021, **4**:174
https://doi.org/10.12688/gatesopenres.13202.2

First published: 27 Nov 2020, **4**:174 https://doi.org/10.12688/gatesopenres.13202.1

REVISED **Amendments from Version 1**

In this update, we have corrected two issues in our data analysis, resulting in a substantial change to one sensitivity analysis and minor changes to other results. We have also substantially moderated the discussion to ensure we keep readers aware of the limitations of our approach and do not over-state the implications our findings.

Any further responses from the reviewers can be found at the end of the article

Introduction

In August, the Peterson-KFF Health System Tracker published a collection of charts showing how healthcare utilization has declined during the COVID-19 pandemic in the United States[1], showing that facility discharge volume dropped by over 25% and cancer screening volumes dropped by over 85% from levels in 2019. This decrease is consistent with evidence from other sources[2,3], and could be driven by a perceived risk of interacting with workers at health facilities. It is yet to be seen how much this delayed and foregone care will reduce population health. Meanwhile, a Wall Street Journal analysis of Centers for Disease Control and Prevention (CDC) data found that at least 7,400 COVID-19 infections were transmitted in US hospitals in 2020[4]. Access to adequate resources for infection prevention among health care workers (HCWs) remains a topic of urgent importance[5].

The existing evidence quantifying the relative COVID-19 incidence rate among HCWs as compared to workers in non-healthcare occupations (non-HCWs) has focused on the first wave of the pandemic, and found that HCWs are at higher risk of COVID[6–9]. We hypothesized that by fall of 2020 there was not a substantially elevated rate of COVID-19 infection among HCWs and that HCWs might even have lower incidence rate than non-HCWs, and we analyzed data from a large survey of Facebook users to investigate.

Methods

Study design

We analyzed individual participant data from a large, web-based survey of Facebook users aged 18 and above in the United States (around 300,000 respondents per week). Every day Facebook offered a random sample of US-based users a Qualtrics survey run by the Delphi lab at Carnegie Mellon University who made it rapidly available to other academic researchers[10,11]. Facebook also provided survey weights to adjust for non-response probability and to match the age and sex distribution at the national level[12,13]. This sort of survey data has been used previously to perform population based analyses related to COVID-19, though never before at such large scale[14,15]. Our analysis relied on the responses to two lines of questions: (1) questions about recent work history, worded as, "In the past 4 weeks, did you do any kind of work for pay?" and if so, "[p]lease select the occupational group that best fits the main kind of work you were doing in the last four weeks"; and (2) questions about COVID-19 testing history, worded as, "Have you **ever** been tested for coronavirus (COVID-19)?", "[h]ave you been tested for coronavirus (COVID-19) in the **last 14 days**?", "[d]id this test find that you had coronavirus (COVID-19)", and "[d]o any of the following reasons describe why you were tested for coronavirus (COVID-19) in **the last 14 days**? Please select all that apply."

We analyzed the six weeks of data from September 8, 2020 to October 20, 2020, which provided more than 80% power to detect a 30% difference between COVID-19 incidence in HCWs and non-HCWs.

Variables

To quantify the relative risk of COVID-19 among healthcare workers (HCWs) versus workers in non-healthcare occupations (non-HCWs), we used the response to the occupational group question as our exposure variable (we coded respondents who selected option "Healthcare practitioners and technicians" or "Healthcare support" as HCWs, and all others, including those with a missing value, as non-HCWs). We identified individuals with COVID-19 as those who reported that they had tested positive for COVID-19 in the last 14 days.

Statistical methods

We calculated the endorsement rate of positive COVID-19 test (ER) for the HCW and non-HCW population as the survey-weighted percent of respondents in either group who reported COVID-19, and calculated the relative COVID-19 incidence ratio (RR) with the equation

$$RR = (ER \text{ among HCWs}) / (ER \text{ among non-HCWs}).$$

We quantified the uncertainty in this ratio using non-parametric bootstrap resampling to obtain a 95% uncertainty interval[16]. To control for confounding due to differential access to COVID-19 testing, we restricted our analysis to only HCWs and non-HCWs who were tested in the last 14 days because their employer or school required it.

As sensitivity analyses, we considered also alternative inclusion criteria and more restrictive subsets of HCWs. The survey provided survey weights that adjust for non-response bias, which we used in our main analysis. However, these weights were designed to represent the national population, and therefore might not represent the HCW population as accurately. As a sensitivity analysis, we repeated our calculation using the unweighted data. To investigate the possibility that workplace testing practices differ between HCW and non-HCW occupational settings, we also repeated our analysis with additional filtering based on the "why you were tested" question. In the main result we used the subset of individuals who responded that they were tested in the last 14 days because of employer/educational requirements, and this question has a "select all that apply" answer type, and also includes "I felt sick" as an option. As a sensitivity analysis, we used only those individuals who were tested because of a workplace requirement and did not feel sick.

Ethical statement
These research activities used no identifiable private information and were therefore exempt from institutional board review.

Results
The survey data contained 43,430 respondents who were tested due to workplace requirements in the time period we focused on, 14,660 HCWs and 28,770 non-HCWs (see Table 1 for demographic details). There were 2,145 respondents who reported a positive test for COVID-19 in the last 14 days (588 among HCWs and 1,557 among non-HCWs).

Among HCWs with a required test, 588 of 14,660 (4.0%) reported a positive test in the last 14 days, while among non-HCWs with a required test, 1,557 of 28,770 (5.4%) reported a positive test, for a relative COVID-19 incidence ratio of 0.73 (95% UI 0.68 to 0.80) (Table 2).

Our power calculation simulation results showed that 7,000 simulants provide 80% power to reject a null hypothesis that HCWs and non-HCWs have the same RR if, in truth, the RR is 0.7. Since the survey currently collects a weekly volume of around 7,000 individuals who report taking a required COVID-19 test, the simulation results imply that six weeks of data will provide more than sufficient power.

Sensitivity analyses
When we repeated our calculation using the unweighted survey responses to calculate the COVID-19 incidence ratio, we found nearly identical relative incidence ratio of 0.74 (95% UI 0.69 to 0.79).

When we repeated our analysis restricted to only specific subtypes of HCWs, as afforded by the questionnaire, we found a range of risks, usually less than 1.0, with substantially less certainty due to small sample sizes (Table 3).

When we used only those individuals who were tested because of a workplace requirement *and* did not feel sick, we obtained a relative risk closer to 1.0. Using only those tested because of a workplace requirement who also *did* feel sick

Table 1. Characteristics of survey respondents.

	Non- healthcare workers		Healthcare workers	
	n	(%)	n	(%)
Total	1,699,214	100.0	123,448	100.0
Tested in last 14 days	133,533	7.9	22,594	18.3
Test required by work or school	28,770	1.7	14,660	11.9
Among those with required test				
Male gender	9,303	32.3	2,106	14.4
Age in years				
18 to 24	3,595	12.5	818	5.6
25 to 34	4,994	17.3	2,544	17.4
35 to 44	5,146	17.9	3,255	22.2
45 to 54	5,179	18.0	3,587	24.5
55 to 64	4,227	14.7	3,345	22.8
65 to 74	1,307	4.5	976	6.7
75 and older	503	1.7	121	0.8

Table 2. Relative COVID-19 incidence rate (RR) and counts of healthcare workers and non-healthcare workers and their crude counts and rates.

Healthcare workers			Non-healthcare workers				
Tested	Positive	%	Tested	Positive	%	RR	95% UI
14,660	588	4.0	28,770	1,557	5.4	0.73	0.68 to 0.80

Table 3. Relative COVID-19 incidence rate (RR) and counts of healthcare workers (HCWs) and non-healthcare workers stratified by worker subtype.

HCW subtype	Number of non-subtype HCWs	Number of subtype HCWs	Relative risk	Lower bound	Upper bound
All HCWs	28,770	14,660	0.73	0.69	0.80
Physician or surgeon	43,139	291	2.71	1.86	3.60
Registered nurse (including nurse practitioner)	40,262	3,168	0.66	0.62	0.82
Licensed practical or licensed vocational nurse	41,318	2,112	0.73	0.60	0.86
Physician assistant	43,274	156	0.63	0.33	1.13
Dentist	43,392	38	0.85	0.24	2.22
Any other treating practitioner	43,046	384	0.56	0.31	0.81
Pharmacist	43,345	85	0.28	0.08	0.72
Any therapist	42,165	1,265	0.51	0.37	0.63
Any health technologist or technician	41,841	1,589	1.01	0.79	1.17
Veterinarian	43,395	35	0.29	0.00	1.28
Nursing assistant or psychiatric aide	41,812	1,618	1.02	0.80	1.22
Home health or personal care aide	42,847	583	0.77	0.52	1.00
Occupational or physical therapy assistant or aide	43,350	80	1.47	0.80	2.31
Massage therapist	43,426	4	10.16	0.00	13.21
Dental assistant	43,412	18	0.00	0.00	0.00
Medical assistant	43,280	150	1.25	0.64	1.96
Medical transcriptionist	43,402	28	0.56	0.00	1.38
Pharmacy aide	43,413	17	0.00	0.00	0.00
Phlebotomist	43,397	33	2.75	0.63	4.06
Veterinary assistant	43,422	8	1.74	0.00	6.97
Any other healthcare support worker	41,104	2,326	0.55	0.46	0.66

we still obtained a relative risk substantially smaller than 1.0 (Table 4). Although this finding could suggest that differences in testing patterns between healthcare and other work settings are partially responsible for the different positivity rates among HCWs and non-HCWs, it could also be driven by greater access to COVID-19 testing for confirmation of illness among HCWs experiencing symptoms. The recall period of 14 days provides ample time for an individual to receive a workplace test without symptoms, then develop symptoms, and then receive another test to determine if the symptoms are due to COVID-19, and HCWs might have more opportunity to access such a follow-up test, since they are visiting a healthcare setting for work already.

Discussion

This study utilized a population-based approach to examine the relative risk of COVID-19 infection among HCW compared with non-HCW. We founda relative COVID-19 incidence ratio substantially and significantly less than 1.0, which can be cautiously interpreted as a positive result, indicating that infection control measures being taken by HCWs in Fall of 2020 were effective.

Our findings are consistent with the limited other evidence available on the risk of COVID-19 in healthcare facility settings[17–20], although also contrast with evidence from prior research that has found that HCWs are at higher risk of

Table 4. Relative COVID-19 incidence rate (RR) and counts of healthcare workers and non-healthcare workers stratified by those who reported they felt/did not feel sick as an additional reason for getting tested.

	Number of non-HCWs	Number of HCWs	Relative risk	Lower bound	Upper bound
Test required, did not feel sick	25,236	13,610	1.09	1.01	1.27
Test required, felt sick	3,534	1,050	0.80	0.69	0.92

COVID[6–9]. This outbreak and our understanding of it have both changed rapidly in the past, and may do so again, so we will continue to update this information.

Limitations

This work has at least three limitations. First, our results are based on self-reported data from a sample of Facebook users and therefore subject to both recall bias and social desirability bias, and may not be representative of the general population or the HCW population. The questions we relied on did not seem particularly at risk for these biases, although the question "have you been tested for COVID-19 in the last 14 days?" likely included positive responses from individuals who received seroprevalence testing as well as PCR testing, which could also introduce a small amount of bias; using this 14-day recall period as a proxy for incidence of COVID-19 could also introduce a small amount of bias. The impact of nonresponse bias is harder to gauge, however; our sensitivity analysis shows that the survey weights do influence our results. Second, our approach required a large sample size to obtain a sufficiently precise estimate of RR, but this seems safer than including respondents who did not report receiving a required test, as that could introduce confounding. Third, it is possible that there was still uncontrolled confounding due to differential access to tests between HCWs and non-HCWs. Our sensitivity analysis found substantively similar results when restricted only to individuals who had workplace testing when they did not feel sick, but since we have only considered respondents with tests required by their employer or school, this might focus on non-HCW setting with better-than-average infection control policies (for example, they *are* doing asymptomatic testing) and therefore the relative risk for HCWs might be even lower than our method estimated.

Conclusion

In October, 2020, in the United States the relative infection ratio of HCWs to non-HCWs was lower than 1.0. Infection control remains essential and HCWs must continue to be protected as the COVID-19 pandemic continues, to ensure safety to themselves, their co-workers, and their patients.

Data availability

Underlying data

The underlying data used in this study are available to academic researchers for research purposes from Facebook at: https://www.facebook.com/research-operations/rfp/?title=covid19-symptom-survey-data-access. Conditions of access and instructions for applications can be found at https://dataforgood.fb.com/docs/covid-19-symptom-survey-request-for-data-access/.

Code availability

Reproducibility code available from: https://github.com/aflaxman/covid_hcw_rr

Archived code at time of resubmission: https://doi.org/10.5281/zenodo.4270367[21].

References

1. **How have healthcare utilization and spending changed so far during the coronavirus pandemic?** Peterson-KFF Health System Tracker. [cited 2020 Oct 21].
Reference Source
2. Chernew ME, Fendrick AM, Armbrester K, *et al.*: **COVID-19 Effects On Care Volumes: What They Might Mean And How We Might Respond**. [cited 2020 Oct 21].
Reference Source
3. Alexander GC, Tajanlangit M, Heyward J, *et al.*: **Use and Content of Primary Care Office-Based vs Telemedicine Care Visits During the COVID-19 Pandemic in the US.** *JAMA Netw Open.* 2020; **3**(10): e2021476.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Evans M: **Hospitals Failed to Fully Contain Covid-19 Inside Their Walls.** *Wall Street Journal.* WSJ News Exclusive, 2020; [cited 2020 Oct 21].
Reference Source
5. Jewett RL: **Battle rages inside US hospitals over how Covid-19 strikes and kills.** *Guardian.* 2020; [cited 2020 Oct 21].
Reference Source
6. Baker MG, Peckham TK, Seixas NS: **Estimating the burden of United States workers exposed to infection or disease: A key factor in containing risk of COVID-19 infection.** *PLoS One.* 2020; **15**(4): e0232452.
PubMed Abstract | Publisher Full Text | Free Full Text
7. CDC COVID-19 Response Team: **Characteristics of Health Care Personnel with COVID-19 — United States, February 12–April 9, 2020.** *MMWR Morb Mortal Wkly Rep.* 2020; **69**(15): 477–481.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Hawkins D, Davis L, Kriebel D: **COVID-19 deaths by occupation, Massachusetts, March 1-July 31, 2020.** *Am J Ind Med.* 2021; **64**(4): 238–44.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Ran L, Chen X, Wang Y, *et al.*: **Risk Factors of Healthcare Workers With**

Coronavirus Disease 2019: A Retrospective Cohort Study in a Designated Hospital of Wuhan in China. *Clin Infect Dis.* 2020; **71**(16): 2218–21.
PubMed Abstract | Publisher Full Text | Free Full Text

10. **COVID-19 Symptom Surveys through Facebook.** The Delphi Blog. [cited 2020 Oct 21].
Reference Source

11. **COVID Symptom Survey.** Delphi Epidata API. [cited 2021 May 8].
Reference Source

12. Barkay N, Cobb C, Eilat R, *et al.*: **Weights and Methodology Brief for the COVID-19 Symptom Survey by University of Maryland and Carnegie Mellon University, in Partnership with Facebook.** *ArXiv200914675 Cs.* 2020; [cited 2020 Oct 21].
Reference Source

13. **Data for Good: New Tools to Help Health Researchers Track and Combat COVID-19**. About Facebook. 2020; [cited 2020 Oct 21].
Reference Source

14. Wang PW, Lu WH, Ko NY, *et al.*: **COVID-19-Related Information Sources and the Relationship With Confidence in People Coping with COVID-19: Facebook Survey Study in Taiwan.** *J Med Internet Res.* 2020; **22**(6): e20021.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Srivastav AK, Sharma N, Samuel AJ: **Impact of Coronavirus disease-19 (COVID-19) lockdown on physical activity and energy expenditure among physiotherapy professionals and students using web-based open E-survey sent through WhatsApp, Facebook and Instagram messengers.** *Clin Epidemiol Glob Health.* 2021; **9**: 78–84.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Efron B: **Bootstrap Methods: Another Look at the Jackknife.** *Ann Stat.* 1979; **7**(1): 1–26.
Reference Source

17. Nalleballe K, Siddamreddy S, Kovvuru S, *et al.*: **Risk of coronavirus disease 2019 (COVID-19) from hospital admission during the pandemic.** *Infect Control Hosp Epidemiol.* undefined/ed; 2020; 1–2.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Ridgway JP, Robicsek AA: **Risk of coronavirus disease 2019 (COVID-19) acquisition among emergency department patients: A retrospective case control study.** *Infect Control Hosp Epidemiol.* 2021; **42**(1): 105–107.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Reale SC, Fields KG, Lumbreras-Marquez MI, *et al.*: **Association Between Number of In-Person Health Care Visits and SARS-CoV-2 Infection in Obstetrical Patients.** *JAMA.* 2020; **324**(12): 1210–1212.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Self WH, Tenforde MW, Stubblefield WB, *et al.*: **Seroprevalence of SARS-CoV-2 Among Frontline Health Care Personnel in a Multistate Hospital Network - 13 Academic Medical Centers, April-June 2020.** *MMWR Morb Mortal Wkly Rep.* 2020; **69**(35): 1221–6.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Flaxman A: **aflaxman/covid_hcw_rr: As resubmitted to Gates Open Research. (Version v1.1.0).** *Zenodo.* 2021.
http://www.doi.org/10.5281/zenodo.4270367

Open Peer Review

Current Peer Review Status: ✔ ✘ ?

Version 2

Reviewer Report 29 July 2021

https://doi.org/10.21956/gatesopenres.14528.r30927

✔ **Alex Reinhart**

Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, USA

I thank the authors for their careful revisions. These revisions address all of my comments in the initial round of review. I still found Table 3 slightly confusing (I initially interpreted "Number of non- subtype HCWs" to mean HCWs not of the subtype, rather than including all respondents, including non-HCWs), but otherwise I believe the revisions have improved the manuscript.

Competing Interests: I am a member of the Delphi group at Carnegie Mellon University. Delphi, in collaboration with Facebook and researchers at the University of Maryland, conducts the survey whose data is analyzed in this article, and I manage much of the process on behalf of Delphi (with assistance from Delphi team members). Delphi makes this data available to many researchers, including the authors of this article. I was not involved in the analysis conducted by the authors of this article, and have not corresponded with them about this research, so my review of the scientific merit of the work has been conducted independently. I confirm that this has not affected my ability to write an objective and unbiased review of this article.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 01 April 2021

https://doi.org/10.21956/gatesopenres.14411.r30426

Tim Driscoll

School of Public Health, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia

This paper presents an analysis of data collected from United States' respondents to a Facebook survey and focuses on a comparison of the rate of COVID-19 in health care workers compared to workers in other sectors. The main finding was that infection is less common in health care workers compared to non-health care workers, with the authors concluding that the results suggest it is "safe" (in terms of risk of COVID-19 infection) to be a health care worker. The methodology seems appropriate. The structure of the paper is good and the meaning is generally clear.

In terms of the Methods, there are inconsistencies in the terminology and I can't see any reason for this. Most particularly, there is mention of an "*endorsement rate*", which is the basis of the "*relative COVID-19 incidence ratio*", but this endorsement rate is not mentioned again in the manuscript. In the Results section, there is mention of a "*relative COVID-19 prevalence ratio*" and a "*Relative COVID-19 incidence rate*". In the Discussion, "*relative COVID-19 incidence ratio*" is mentioned again. I presume all three of these terms represent the same quantity. If so, it seems just one term should be used. If not, there needs to be further explanation about what has been calculated and why. It appears that the information presented is prevalence rather than incidence, because although the testing was in the previous 14 days the positive result could reflect past disease, depending on the type of test. If it is assumed the testing was done via PCR and further assumed this PCR test would only be positive for recent (in the previous two weeks or so) infection, then incidence would be an appropriate term to use, but then the implications of this assumption should be considered in the Discussion. Either way, the uncertainty arising from lack of information about the testing seems to be a limitation that could usefully be included at the end of the Discussion.

The conclusion that "*HCWs need not fear contracting or transmitting infections more than other workers do…*" seems too strong given the limitations of the data used for this study and the "*…limited other evidence available…*", as acknowledged by the authors. Similarly, the preceding statement that the result is "*…an unequivocally positive finding…*" is at odds with the limitations considered later in the paper. I agree that if the results are accepted on face value they imply that health care workers are at lower risk than non-health care workers, but the other aspects just mentioned mean that conclusions based on these results should be guarded. Also, health care workers are analysed as a group, or in smaller but still broad groups in Table 3. This group will contain a mixture of people working directly with the public (front-line health workers) in a clinical setting and people working in health care but with minimal contact with patients. It might well be that the front-line health workers do indeed have a higher risk of infection than the general public, but that this is not reflected in the study results because the other health care workers have a much lower risk of infection. The fact that the "*Physician or surgeon*" group appears to have a higher risk (RR=2.6) supports this concern. Having mentioned Table 3, the interpretation of this is not clear. Why are there different numbers of non-health care workers in each row, and why do they appear in any row if each row represents a different type of health care worker? It would be helpful to explain this.

There is quite a bit of space in the paper considering the power of the study. The reason for this is not clear. The power calculations are based on an assumed difference of at least 30% in the

"prevalence" of COVID-19 between health care workers and non-health care workers. This would be important if the difference found was less than 30%. However, since the difference found was 30%, the power calculations don't seem relevant. Also, the program to undertake this power calculation was included in the paper. I am not sure this adds much; I don't mind it being there but it is not further considered and in fact isn't directly referred to – it just appears in the text at the end of, or actually part of, the last sentence in the section describing the power calculation. That seems odd.

The authors rightly identify some limitations in their work. These primarily result from the data used in the analysis rather than from the analysis used. The authors note the potential for some forms of reporting bias and for uncontrolled confounding, both of which I agree may be of concern. They also mention the need for a large sample size, which doesn't seem to be a limitation in terms of interpreting the results of the study; the large sample size is not a source of bias, just something that requires greater statistical resources.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: Dr Flaxman works at the Institute for Health Metrics and Evaluation, which runs the Global Burden of Disease (GBD) study. I am head of the Occupational Risk Factors Expert Working Group working on the GBD study. I have co-authored papers with Dr Flaxman that have arisen from this study but have not worked closely with him on any aspects of the study and the papers that we have co-authored have had a large number of co-authors. I don't have a personal relationship with Dr Flaxman. I believe I can provide an objective review of this paper.

Reviewer Expertise: Epidemiology, occupational medicine

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 May 2021

Abraham Flaxman, University of Washigton, Seattle, USA

In terms of the Methods, there are inconsistencies in the terminology and I can't see any reason for this. Most particularly, there is mention of an "endorsement rate", which is the basis of the "relative COVID-19 incidence ratio", but this endorsement rate is not mentioned again in the manuscript. In the Results section, there is mention of a "relative COVID-19 prevalence ratio" and a "Relative COVID-19 incidence rate". In the Discussion, "relative COVID-19 incidence ratio" is mentioned again. I presume all three of these terms represent the same quantity. If so, it seems just one term should be used. If not, there needs to be further explanation about what has been calculated and why. It appears that the information presented is prevalence rather than incidence, because although the testing was in the previous 14 days the positive result could reflect past disease, depending on the type of test. If it is assumed the testing was done via PCR and further assumed this PCR test would only be positive for recent (in the previous two weeks or so) infection, then incidence would be an appropriate term to use, but then the implications of this assumption should be considered in the Discussion. Either way, the uncertainty arising from lack of information about the testing seems to be a limitation that could usefully be included at the end of the Discussion.

Response: We have standardized our terminology on incidence, which we think is the most precise and accurate of the terms we used originally; thank you for calling attention to this inconsistency. We have also added to the limitations section to highlight the way 14-day recall is not exactly "incidence".

The conclusion that "HCWs need not fear contracting or transmitting infections more than other workers do..." seems too strong given the limitations of the data used for this study and the "...limited other evidence available...", as acknowledged by the authors. Similarly, the preceding statement that the result is "...an unequivocally positive finding..." is at odds with the limitations considered later in the paper. I agree that if the results are accepted on face value they imply that health care workers are at lower risk than non-health care workers, but the other aspects just mentioned mean that conclusions based on these results should be guarded. Also, health care workers are analysed as a group, or in smaller but still broad groups in Table 3. This group will contain a mixture of people working directly with the public (front-line health workers) in a clinical setting and people working in health care but with minimal contact with patients. It might well be that the front-line health workers do indeed have a higher risk of infection than the general public, but that this is not reflected in the study results because the other health care workers have a much lower risk of infection. The fact that the "Physician or surgeon" group appears to have a higher risk (RR=2.6) supports this concern.

Response: We have moderated the discussion in light of this comment, as well as the similar concerns from Reviewer 2.

Having mentioned Table 3, the interpretation of this is not clear. Why are there different numbers of non-health care workers in each row, and why do they appear in any row if each row represents a different type of health care worker? It would be helpful to explain this.

Response: Each row besides the first row compares a subtype of HCWs to everyone who is

not of that subtype. We have edited the column headings to make this clearer.

There is quite a bit of space in the paper considering the power of the study. The reason for this is not clear. The power calculations are based on an assumed difference of at least 30% in the "prevalence" of COVID-19 between health care workers and non-health care workers. This would be important if the difference found was less than 30%. However, since the difference found was 30%, the power calculations don't seem relevant. Also, the program to undertake this power calculation was included in the paper. I am not sure this adds much; I don't mind it being there but it is not further considered and in fact isn't directly referred to – it just appears in the text at the end of, or actually part of, the last sentence in the section describing the power calculation. That seems odd.

Response: We did this power calculation in so much detail because we wanted to get our results out as soon as possible, but not so soon that we were fooled by chance variation in the data. We have taken it out to focus the reader on the most important parts, especially now that there is so much more data available.

The authors rightly identify some limitations in their work. These primarily result from the data used in the analysis rather than from the analysis used. The authors note the potential for some forms of reporting bias and for uncontrolled confounding, both of which I agree may be of concern. They also mention the need for a large sample size, which doesn't seem to be a limitation in terms of interpreting the results of the study; the large sample size is not a source of bias, just something that requires greater statistical resources.

Response: We thank the reviewer for this perspective, and have attempted to edit the limitations section to make it clearer.

Competing Interests: As stated in manuscript.

Reviewer Report 29 March 2021

https://doi.org/10.21956/gatesopenres.14411.r30475

Devan Hawkins
Department of Public Health Program, Schools of Arts and Sciences, MCPHS University, Boston, MA, USA
Marcy Goldstein-Gelb
National Council for Occupational Safety and Health, Somerville, MA, USA

Thank you for the invitation to review this paper. The paper addresses an important topic (the risk of acquiring COVID-19 among healthcare workers). The authors apply unique methods to study

the problem. However, we have some concerns about how the analysis was performed and how the results were interpreted. Below, we provide details about these concerns.

Introduction:

- The authors should provide some information about previous studies that have examined the risk for COVID-19 among healthcare workers and also justify why they hypothesized that healthcare workers would have a lower risk. Some studies have suggested that they have an elevated risk. Below are some studies that have examined the risk/potential risk for COVID-19 among healthcare workers:
  - Baker *et al.* (2020[1]).
  - Burrer *et al.* (2020[2]).
  - Hawkins *et al.* (2020[3]).
  - Ran *et al.* (2020[4]).

Methods:

- The authors should explain the justification for weighting to the overall Facebook population more. If the goal is to ensure that the healthcare workers survey from Facebook are representative of healthcare workers, this type of weighting may not help.
- Was industry information available? There is good reason to suspect that risk will be different across different industry. In some cases, HCWs will even be working from home with telehealth. It may be useful to:
  - 1) Compare healthcare workers employed in the healthcare industry to other health care workers
  - 2) Examine the risk among different industries
- We strongly recommend including all positive tests as a sensitivity analysis not just those required by work. I agree that differential testing may introduce a bias, but it would be better to show all the data so that we can consider the potential magnitude of that bias. There may actually be an even greater differential between HCW and other workers. In fact, probably most non-health care workers don't get tested through employer requirements, and only know that they have COVID after becoming sick.
- Additionally, we strongly recommend having a different reference population than all non-healthcare workers. Other high risk workers are included in the current reference group, which may have the impact of making the risk among healthcare workers appear lower. Potentially consider including major census or SOC occupations for comparison.
- For non-health care workers, did they ask whether they worked outside the home, or was there just an assumption that they did. Naturally if they were tested but work from home, that would be an overrepresentation of work-relatedness, though I would assume it would not be an employer requirement if they work from home.
- Was the survey only conducted in English?

Results:

- The demographics for healthcare workers should be compared to national data about healthcare workers demographics. This data can be obtained from the CPS or census. CPS is linked here: https://www.bls.gov/cps/tables.htm

- Consider separating occupations into major categories for more fair comparisons. You may consider weighting to this data rather than the Facebook demographics.

- Is race/ethnicity data available? If workers of color are under-represented this could introduce bias to the study, because these workers may be more likely to be employed in higher risk healthcare occupations.

- Table 3: How do the distributions of detailed occupations compare to national data about employment in these occupations? The CPS data linked above can be used to assess this. Bias may be introduced if certain occupations are underrepresented.

- Table 3: The authors should discuss the variability in rates according to specific healthcare occupations. They may consider including the groups according to major healthcare occupations (practioners, support, etc.). Some occupations have elevated rates.

Discussion:

- We strongly recommend removing this finding: “an unequivocally positive findings, indicating that infection control measures being taken by HCWs in total are effective.” Based on the limitations of this study, we do not believe that the findings support this conclusion. The findings may be suggestive of effective measures being taken if some of the limitations in the methods/results are addressed.

- Consider other findings linked above which are not consistent with this study’s findings of a lower risk among HCWs.

- We strong discourage concluding that HCWs should not fear contracting or transmitting infections more than other workers. HCWs don't base their fear on how their likelihood of exposure compares to other worker fears - they're afraid, according to other factors, including often not having adequate protection methods.

References

1. Baker MG, Peckham TK, Seixas NS: Estimating the burden of United States workers exposed to infection or disease: A key factor in containing risk of COVID-19 infection.*PLoS One*. 2020; **15** (4): e0232452 PubMed Abstract | Publisher Full Text
2. CDC COVID-19 Response Team, CDC COVID-19 Response Team, Burrer S, de Perio M, et al.: Characteristics of Health Care Personnel with COVID-19 — United States, February 12–April 9, 2020. *MMWR. Morbidity and Mortality Weekly Report*. 2020; **69** (15): 477-481 Publisher Full Text
3. Hawkins D, Davis L, Kriebel D: COVID-19 deaths by occupation, Massachusetts, March 1-July 31, 2020.*Am J Ind Med*. 2021; **64** (4): 238-244 PubMed Abstract | Publisher Full Text
4. Ran L, Chen X, Wang Y, Wu W, et al.: Risk Factors of Healthcare Workers With Coronavirus Disease 2019: A Retrospective Cohort Study in a Designated Hospital of Wuhan in China. *Clinical*

Infectious Diseases. 2020; **71** (16): 2218-2221 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Devan Hawkins: Occupational health epidemiologist

We confirm that we have read this submission and believe that we have an appropriate level of expertise to state that we do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 17 May 2021

Abraham Flaxman, University of Washigton, Seattle, USA

Introduction:

- *The authors should provide some information about previous studies that have examined the risk for COVID-19 among healthcare workers and also justify why they hypothesized that healthcare workers would have a lower risk. Some studies have suggested that they have an elevated risk. Below are some studies that have examined the risk/potential risk for COVID-19 among healthcare workers:*
- *1. Baker MG, Peckham TK, Seixas NS: Estimating the burden of United States workers exposed to infection or disease: A key factor in containing risk of COVID-19 infection.PLoS One. 2020; **15** (4): e0232452 PubMed Abstract | Publisher Full Text*
- *2. CDC COVID-19 Response Team, CDC COVID-19 Response Team, Burrer S, de Perio M, et al.: Characteristics of Health Care Personnel with COVID-19 — United States, February 12–April 9, 2020. MMWR. Morbidity and Mortality Weekly Report. 2020; **69** (15): 477-481 Publisher Full Text*

- *3. Hawkins D, Davis L, Kriebel D: COVID-19 deaths by occupation, Massachusetts, March 1-July 31, 2020.Am J Ind Med. 2021; **64** (4): 238-244 PubMed Abstract | Publisher Full Text*

- *4. Ran L, Chen X, Wang Y, Wu W, et al.: Risk Factors of Healthcare Workers With Coronavirus Disease 2019: A Retrospective Cohort Study in a Designated Hospital of Wuhan in China. Clinical Infectious Diseases. 2020; **71** (16): 2218-2221 Publisher Full Text*

Response: Thank you for calling our attention to this growing body of work. We have added to this introduction to include this prior work and clarify our hypothesis.

Methods:

- *The authors should explain the justification for weighting to the overall Facebook population more. If the goal is to ensure that the healthcare workers survey from Facebook are representative of healthcare workers, this type of weighting may not help.*

Response: Thank you for identifying this risk to the validity of our findings. We have added more detail about the weights in the Study Design section, as well as additional caveats about using the weights for the HCW population in sensitivity analyses in the Statistical Methods section. We have also added to the limitations section to provide more caveats about the risk of non-response bias.

Was industry information available? There is good reason to suspect that risk will be different across different industry. In some cases, HCWs will even be working from home with telehealth. It may be useful to:

- *1) Compare healthcare workers employed in the healthcare industry to other health care workers*
- *2) Examine the risk among different industries*

Response: Unfortunately, the survey instrument does not distinguish between occupation and industry, and therefore we can only examine risk between different occupations, as identified by responses to the question "[p]lease select the occupational group that best fits the main kind of work you were doing in the last four weeks". Respondents selected a single category from a short list, and then a detailed category from a longer list, and all of the detailed categories that of HCW are listed in Table 3.

We strongly recommend including all positive tests as a sensitivity analysis not just those required by work. I agree that differential testing may introduce a bias, but it would be better to show all the data so that we can consider the potential magnitude of that bias. There may actually be an even greater differential between HCW and other workers. In fact, probably most non-health care workers don't get tested through employer requirements, and only know that they have COVID after becoming sick.

Response: The results of this proposed sensitivity analysis might surprise the reviewer: in an analysis of all survey respondents (123,448 HCWs and 1,699,214 non-HCWs) we find that among HCWs (tested and untested), 1,674 of 123,448 (1.4%) reported a positive test in the last 14 days; while among non-HCWs (tested and untested), 11,963 of 1,699,214 (0.70%) reported a positive test. This yields a ratio of 1.8 (95% UI 1.52 to 2.03), but it is confounded by the fact that HCWs have greater access to testing than non-HCWs and cannot be used as an estimate of the relative incidence ratio of COVID-19.

If we restrict our analysis to only individuals who have been tested in the last 14 days, we find 156,127 respondents who were tested (regardless of workplace requirements) in the time period we focused on, 22,594 HCWs and 133,533 non-HCWs; Among HCWs tested (regardless of whether the test was required), 1,674 of 22,594 (7.4%) reported a positive test in the last 14 days, while among non-HCWs tested (regardless of whether the test was required), 11,963 of 133,533 (8.96%) reported a positive test, for an RR of 0.8 (95% UI 0.78 to 0.83).

Response: We prefer to keep this complexity out of the main paper; in some occupations, required testing happens only after symptoms develop, and in light of this, we prefer our sensitivity analysis using only required tests among asymptomatic workers to investigating this potential risk of confounding.

Additionally, we strongly recommend having a different reference population than all non-healthcare workers. Other high risk workers are included in the current reference group, which may have the impact of making the risk among healthcare workers appear lower. Potentially consider including major census or SOC occupations for comparison.

Response: We prefer to focus our discussion on a comparison of HCWs with all non-HCWs, but the reviewer raises an interesting additional question. Although we choose to leave a full investigation of these occupational comparisons for future work, we cannot resist examining them briefly in this response. After HCWs, the occupation with the highest rates of required testing are (16) Other occupation, (2) education, training, and library, (11) office and administration services, and (7) food preparation and serving. Our comparison of HCWs to workers in occupation "Other" found a relative COVID-19 incidence ratio of 0.97 (95% UI 0.82 to 1.12).

This also identifies an important divergence between the “non-HCW” population and the worker population---there are 9,652 respondents without an occupation code included in the non-HCW population. Repeating our analysis with these respondents excluded finds a ratio of 0.60 (95% UI 0.55 to 0.67).

For non-health care workers, did they ask whether they worked outside the home, or was there just an assumption that they did. Naturally if they were tested but work from home, that would be an overrepresentation of work-relatedness, though I would assume it would not be an employer requirement if they work from home.

Response: The survey does include the question “Was any of your work for pay in the last four weeks outside your home?”, and as an additional sensitivity analysis which we excluded from our report we considered the same analysis stratified on work-from-home status. We were surprised to find quantitatively similar results among those who work from home and those who do not.

Was the survey only conducted in English?

The survey was translated into multiple languages (Spanish, French, Portuguese, Chinese,

Vietnamese). We have added a reference to the https://cmu-delphi.github.io/delphi-epidata/symptom-survey/ website with full details on the survey instrument.

Results:

- *The demographics for healthcare workers should be compared to national data about healthcare workers demographics. This data can be obtained from the CPS or census. CPS is linked here: https://www.bls.gov/cps/tables.htm*

Response: We appreciate this suggestion, but prefer to keep the main paper simpler and instead include the comparison in this response only. Among survey respondents, HCWs were 85.7% female, while among employed persons in 2020, "Healthcare practitioners and technical occupations" were 74.4% female. The age distribution was also similar, but not identical.

Consider separating occupations into major categories for more fair comparisons. You may consider weighting to this data rather than the Facebook demographics.

Response: We agree that this would be a valuable extension of the approach we have applied in this paper, but we would like to limit the scope of this work to focus solely on the comparison of HCWs to non-HCWs, and leave further investigation and comparison of other occupations and categories for future work. We agree that additional sensitivity analyses would be warranted in this future work to determine if alternative weighting of the data yields substantively divergent results. We believe, however, that our sensitivity analyses for the HCW versus non-HCW comparison establish that the substantive finding of an RR substantially below 1.0 for HCWs is robust.

Is race/ethnicity data available? If workers of color are under-represented this could introduce bias to the study, because these workers may be more likely to be employed in higher risk healthcare occupations.

Response: The survey instrument did include race and ethnicity information, but we do not currently have access to these columns of the data. Subsequent work investigating racial and ethnic differences in both response rates and test results would be very interesting.

Table 3: How do the distributions of detailed occupations compare to national data about employment in these occupations? The CPS data linked above can be used to assess this. Bias may be introduced if certain occupations are underrepresented.

Response: Some of the age distributions are quite similar, for example for nurses, while others have small sample sizes and are probably biased by differential response patterns, for example physicians. Though we included all subcategories for completeness, we felt it was important to include the sample size as well, to make sure readers were not overly influenced by the calculations based on only a small number of respondents.

We agree that this would be a valuable extension of the approach we have applied in this paper, but we would like to limit the scope of this work to focus solely on the comparison of HCWs to non-HCWs, and leave further investigation and comparison of other occupations

and categories for future work.

Discussion:

- *We strongly recommend removing this finding: "an unequivocally positive findings, indicating that infection control measures being taken by HCWs in total are effective." Based on the limitations of this study, we do not believe that the findings support this conclusion. The findings may be suggestive of effective measures being taken if some of the limitations in the methods/results are addressed.*

Response: We appreciate the reviewers recommendation and we have substantially moderated the discussion to ensure we keep readers aware of the limitations of our approach and do not over-state the implications our findings.

Consider other findings linked above which are not consistent with this study's findings of a lower risk among HCWs.

Response: We have referred to this contrasting evidence base in the discussion now, as well as in the introduction.

- We strong discourage concluding that HCWs should not fear contracting or transmitting infections more than other workers. HCWs don't base their fear on how their likelihood of exposure compares to other worker fears - they're afraid, according to other factors, including often not having adequate protection methods.

Response: We have moderated the language in our conclusion, and thank the reviewer again for helping us avoid over-stating the implications of our findings.

Competing Interests: As stated in manuscript.

Reviewer Report 04 December 2020

https://doi.org/10.21956/gatesopenres.14411.r30079

?

Alex Reinhart
Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, USA

This presents a timely and useful analysis of large-scale survey data. For an analysis like this, it's very important to clearly present the meaning of the data and the caveats in the survey design; the authors do a good job here, and my comments here focus on making the paper even clearer.

The analysis seems reasonable overall, and, subject to the limitations of the survey design, a useful contribution to the area.

I've separated my comments into "Main comments", which I think should be addressed to make

the article more sound, and "Minor comments" that just make minor improvements to the paper.

Main comments:

- The "Sensitivity analyses" section (page 5) explains that "When we repeated our calculation using the unweighted survey responses to calculate the COVID-19 incidence ratio, we found an even smaller relative incidence ratio of 0.4 (95% UI 0.3 to 0.5)." This seems surprising. Do you have any hypotheses that could explain why this is? It suggests that either the age and gender distributions for HCWs and non-HCWs are quite different (since the survey weights correct for age and gender) or that the estimated non-response for the groups are quite different.

- The last paragraph of the Discussion suggests the possibility that "since we have only considered respondents with tests required by their employer or school, this might focus on non-HCW setting with better-than-average infection control policies". This may be a good subject for an additional table of results: A comparison of the distributions of occupation among non-HCW people who were required to be tested and those who were not. Such a table would tell the reader whether those who are required to be tested are from an unusual group of occupations, to help tell whether those occupations might be higher or lower risk than average.

- Table 3 contains a "Number of non-HCWs" column, but I don't know how to interpret this. What does it mean to say that there were 26,805 non-HCWs in the "All HCWs" row?

- In the Limitations (page 6), the authors mention recall bias and social desirability bias as possible problems. But another key bias would be response bias: while Facebook's weights try to adjust for non-response, if they do not completely adjust for every possible factor related to non-response, there can still be bias. For example, if people who are much more concerned about COVID and take more precautions are also more likely to participate in the survey, and if Facebook does not have covariates that can predict this accurately, the survey sample can be biased relative to the population. It would be good to address this and indicate how it could affect the results.

Minor comments:

- The "Study design" subsection mentions that "Facebook also provided survey weights to adjust for the demographics of the active Facebook user population." It would be good to be explicit about what corrections are included in the weights:
  - The weights adjust for non-response, using Facebook's estimate of the probability of each sampled individual participating in the survey.
  - The weights are then post-stratified by age and gender only.

- In the "Study design" subsection, the second paragraph states "We analyzed the most recently available six weeks of data from September 6, 2020 to October 18, 2020", but Wave 4 of the survey (containing the occupation and testing questions) was only deployed on September 8, 2020. If data from September 6 and 7 was included, I assume it was left out of the study, because the respondents would not have answered the relevant questions.

- It may help readers to be explicit about the survey text and its location. The survey documentation site contains the full text of each survey wave, and referring to this could

help readers who want to read the survey text and flow.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: I am a member of the Delphi group at Carnegie Mellon University. Delphi, in collaboration with Facebook and researchers at the University of Maryland, conducts the survey whose data is analyzed in this article, and I manage much of the process on behalf of Delphi (with assistance from Delphi team members). Delphi makes this data available to many researchers, including the authors of this article. I was not involved in the analysis conducted by the authors of this article, and have not corresponded with them about this research, so my review of the scientific merit of the work has been conducted independently. I confirm that this has not affected my ability to write an objective and unbiased review of this article.

Reviewer Expertise: I am a professional statistician and assistant teaching professor of Statistics & Data Science at Carnegie Mellon University. I am also a member of the Delphi group, and manage the collection of the survey data described in this article; see my Competing Interests for further details.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 17 May 2021

Abraham Flaxman, University of Washigton, Seattle, USA

This presents a timely and useful analysis of large-scale survey data. For an analysis like this, it's very important to clearly present the meaning of the data and the caveats in the survey design; the authors do a good job here, and my comments here focus on making the paper even clearer.
Response: We thank the reviewer for this assessment.

The analysis seems reasonable overall, and, subject to the limitations of the survey design, a useful contribution to the area.
I've separated my comments into "Main comments", which I think should be addressed to make the article more sound, and "Minor comments" that just make minor improvements to the paper.

Main comments:

- *The "Sensitivity analyses" section (page 5) explains that "When we repeated our calculation using the unweighted survey responses to calculate the COVID-19 incidence ratio, we found an even smaller relative incidence ratio of 0.4 (95% UI 0.3 to 0.5)." This seems surprising. Do you have any hypotheses that could explain why this is? It suggests that either the age and gender distributions for HCWs and non-HCWs are quite different (since the survey weights correct for age and gender) or that the estimated non-response for the groups are quite different.*

Response: This appears to be an error in our number-plugging! In the archived code corresponding to this submission, we have a relative incidence ratio of 0.70 (95% UI 0.65 to 0.74). We apologize for this and thank the reviewer for their careful reading that helped find and fix this defect!

The last paragraph of the Discussion suggests the possibility that "since we have only considered respondents with tests required by their employer or school, this might focus on non-HCW setting with better-than-average infection control policies". This may be a good subject for an additional table of results: A comparison of the distributions of occupation among non-HCW people who were required to be tested and those who were not. Such a table would tell the reader whether those who are required to be tested are from an unusual group of occupations, to help tell whether those occupations might be higher or lower risk than average.

Response: We appreciate the reviewer's suggestion, but prefer to restrict the scope of this paper to focus only on HCWs, and leave investigation of other occupations for future research.

Table 3 contains a "Number of non-HCWs" column, but I don't know how to interpret this. What does it mean to say that there were 26,805 non-HCWs in the "All HCWs" row?

Response: Thank you for flagging this confusing terminology. By "non-HCWs" we meant the number of respondents who are *not* in the HCW subgroup for which the row reports the relative risk. We have renamed the column headers to make this clearer.

- *In the Limitations (page 6), the authors mention recall bias and social desirability bias as possible problems. But another key bias would be response bias: while Facebook's weights try to adjust for non-response, if they do not completely adjust for every possible factor related to non-response, there can still be bias. For example, if people who are much more concerned about COVID and take more precautions are also more likely to participate in the survey, and if Facebook does not have covariates that can predict this accurately, the survey sample can be biased relative to the population. It would be good to address this and indicate how it could affect the results.*

Response: Thank you for calling attention to this important limitation. We have added a sentence to the limitations section about it.

Minor comments:

- *The "Study design" subsection mentions that "Facebook also provided survey weights to adjust for the demographics of the active Facebook user population." It would be good to be explicit about what corrections are included in the weights:*
  - *The weights adjust for non-response, using Facebook's estimate of the probability of each sampled individual participating in the survey.*
  - *The weights are then post-stratified by age and gender only.*

Response: We have edited to include this detail explicitly.

In the "Study design" subsection, the second paragraph states "We analyzed the most recently available six weeks of data from September 6, 2020 to October 18, 2020", but Wave 4 of the survey (containing the occupation and testing questions) was only deployed on September 8, 2020. If data from September 6 and 7 was included, I assume it was left out of the study, because the respondents would not have answered the relevant questions.

Response: Good point, we have updated to text to reflect the days use only Wave 4 data, and shifted the data end date to still include precisely 6 weeks of data. This resulted in minor changes to many of our results, but no changes to our substantive findings.

- *It may help readers to be explicit about the survey text and its location. The survey documentation site contains the full text of each survey wave, and referring to this could help readers who want to read the survey text and flow.*

Response: Thank you for suggesting this, we have added a reference to this documentation.

Competing Interests: As stated in manuscript.