BMJ Open

Problems with evidence assessment in COVID-19 health policy impact evaluation: a systematic review of study design and evidence strength

Noah A Haber ,[1] Emma Clarke-Deelder,[2] Avi Feller,[3] Emily R Smith,[4] Joshua A. Salomon,[5] Benjamin MacCormack-Gelles,[2] Elizabeth M Stone,[6] Clara Bolster-Foucault,[7] Jamie R Daw,[8] Laura Anne Hatfield ,[9] Carrie E Fry,[10] Christopher B Boyer,[11] Eli Ben-Michael,[12] Caroline M Joyce,[7] Beth S Linas,[13,14] Ian Schmid,[15] Eric H Au,[16] Sarah E Wieten,[1] Brooke Jarrett ,[13] Cathrine Axfors,[1] Van Thu Nguyen,[1] Beth Ann Griffin,[17] Alyssa Bilinski,[18] Elizabeth A Stuart[15]

To cite: Haber NA, Clarke-Deelder E, Feller A, *et al*. Problems with evidence assessment in COVID-19 health policy impact evaluation: a systematic review of study design and evidence strength. *BMJ Open* 2022;**12**:e053820. doi:10.1136/bmjopen-2021-053820

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-053820).

Received 26 May 2021
Accepted 03 December 2021

For numbered affiliations see end of article.

Correspondence to
Dr Noah A Haber;
noahhaber@stanford.edu

ABSTRACT

Introduction Assessing the impact of COVID-19 policy is critical for informing future policies. However, there are concerns about the overall strength of COVID-19 impact evaluation studies given the circumstances for evaluation and concerns about the publication environment.

Methods We included studies that were primarily designed to estimate the quantitative impact of one or more implemented COVID-19 policies on direct SARS-CoV-2 and COVID-19 outcomes. After searching PubMed for peer-reviewed articles published on 26 November 2020 or earlier and screening, all studies were reviewed by three reviewers first independently and then to consensus. The review tool was based on previously developed and released review guidance for COVID-19 policy impact evaluation.

Results After 102 articles were identified as potentially meeting inclusion criteria, we identified 36 published articles that evaluated the quantitative impact of COVID-19 policies on direct COVID-19 outcomes. Nine studies were set aside because the study design was considered inappropriate for COVID-19 policy impact evaluation (n=8 pre/post; n=1 cross-sectional), and 27 articles were given a full consensus assessment. 20/27 met criteria for graphical display of data, 5/27 for functional form, 19/27 for timing between policy implementation and impact, and only 3/27 for concurrent changes to the outcomes. Only 4/27 were rated as overall appropriate. Including the 9 studies set aside, reviewers found that only four of the 36 identified published and peer-reviewed health policy impact evaluation studies passed a set of key design checks for identifying the causal impact of policies on COVID-19 outcomes.

Discussion The reviewed literature directly evaluating the impact of COVID-19 policies largely failed to meet key design criteria for inference of sufficient rigour to be actionable by policy-makers. More reliable evidence review is needed to both identify and produce policy-actionable evidence, alongside the recognition that actionable evidence is often unlikely to be feasible.

Strengths and limitations of this study

- This study is based on previously released review guidance for discerning and evaluating critical minimal methodological design aspects of the COVID-19 health policy impact evaluation.
- The review tool assesses critical aspects of study design grounded in impact evaluation methods that must be true for the papers to provide useful policy impact evaluation, including what type of impact evaluation method was used, graphical display of outcomes data, functional form for the outcomes, timing between policy and impact, concurrent changes to the outcomes and an overall rating.
- This study used a consensus reviewer model with three reviewers in order to obtain replicable results for study strength ratings.
- While the vast majority of studies in our sample received low ratings for useful causal policy impact evaluation, they may make other contributions to the literature.
- Because our review tool was limited to a very narrow—although critical—set of items, weaknesses in other aspects not reviewed (eg, data quality or other aspects of statistical inference) may further weaken studies that were found to meet our criteria.

INTRODUCTION

Policy decisions to mitigate the impact of COVID-19 on morbidity and mortality are some of the most important issues policymakers have had to make since January 2020. Decisions regarding which policies are enacted depend in part on the evidence base for those policies, including understanding what impact past policies had on COVID-19 outcomes.[1 2] Unfortunately, there are substantial concerns that much of the existing literature may be methodologically flawed, which

could render its conclusions unreliable for informing policy. The combination of circumstances being difficult for strong impact evaluation, the importance of the topic and concerns over the publication environment may lead to the proliferation of low strength studies.

High-quality causal evidence requires a combination of rigorous methods, clear reporting, appropriate caveats and the appropriate circumstances for the methods used.[3–6] Rigorous evidence is difficult in the best of circumstances, and the circumstances for evaluating non-pharmaceutical intervention (NPI) policy effects on COVID-19 are particularly challenging.[5] The global pandemic has yielded a combination of a large number of concurrent policy and non-policy changes, complex infectious disease dynamics, and unclear timing between policy implementation and impact; all of this makes isolating the causal impact of any particular policy or policies exceedingly difficult.[7]

The scientific literature on COVID-19 is exceptionally large and fast growing. Scientists published more than 100 000 papers related to COVID-19 in 2020.[8] There is some general concern that the volume and speed[9 10] at which this work has been produced may result in a literature that is overall low quality and unreliable.[11–15]

Given the importance of the topic, it is critical that decision-makers are able to understand what is known and knowable[5 16] from observational data in COVID-19 policy, as well as what is unknown and/or unknowable.

Motivated by concerns about the methodological strength of COVID-19 policy evaluations, we set out to review the literature using a set of methodological design checks tailored to common policy impact evaluation methods. Our primary objective was to evaluate each paper for methodological strength and reporting, based on pre-existing review guidance developed for this purpose.[17] As a secondary objective, we also studied our own process: examining the consistency, ease of use, and clarity of this review guidance.

This protocol differs in several ways from more traditional systematic review protocols given the atypical objectives and scope of the systematic review. First, this is a systematic review of methodological strength of evidence for a given literature as opposed to a review summary of the evidence of a particular topic. As such, we do not summarise and attempt to combine the results for any of the literature. Second, rather than being a comprehensive review of every possible aspect of what might be considered 'quality,' this is a review of targeted critical design features for actionable inference for COVID-19 policy impact evaluation and methods. It is designed to be a set of broad criteria for minimal plausibility of actionable causal inference, where each of the criteria is necessary but not sufficient for strong design. Issues in other domains (data, details of the design, statistics, etc) further reduce overall actionability and quality, and thorough review in those domains is needed for any studies passing our basic minimal criteria. Third, because the scope relies on guided, but difficult and subjective assessments of methodological appropriateness, we use a discussion-based consensus process to arrive at consistent and replicable results, rather than a more common model with two independent reviewers with conflict resolution. The independent review serves primarily as a starting point for discussion, but is neither designed nor expected to be a strong indicator of the overall consensus ratings of the group.

METHODS

Overview

This protocol and study was written and developed following the release of the review guidance written by the author team in September 2020 on which the review tool is based. The protocol for this study was pre-registered on OSF.io[18] in November 2020 following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[19] Deviations from the original protocol are discussed in online supplemental appendix 1, and consisted largely of language clarifications and error corrections for both the inclusion criteria and review tool, an increase in the number of reviewers per fully reviewed article from two to three, and simplification of the statistical methods used to assess the data.

For this study, we ascertain minimal criteria for studies to be able to plausibly identify causal effects of policies, which is the information of greatest interest to inform policy decisions. The causal estimand is something that, if known, would definitely help policy-makers decide what to do (eg, whether to implement or discontinue a policy). The study estimates that target causal quantity with a rigorous design and appropriate data in a relevant population/sample. For shorthand, we refer to this as minimal properties of 'actionable' evidence.

This systematic review of the strength of evidence took place in three phases: search, screening and full review.

Eligibility criteria

The following eligibility criteria were used to determine the papers to include:

- The primary topic of the article must be evaluating one or more individual COVID-19 or SARS-CoV-2 policies on direct COVID-19 or SARS-CoV-2 outcomes
  - The primary exposure(s) must be a policy, defined as a government-issued order at any government level to address a directly COVID-19-related outcome (eg, mask requirements, travel restrictions, etc).
  - Direct COVID-19 or SARS-CoV-2 outcomes are those that are specific to disease and health outcomes may include cases detected, mortality, number of tests taken, test positivity rates, Rt, etc.
  - This may NOT include indirect impacts of COVID-19 on items that are not direct COVID-19 or SARS-CoV-2 impacts such as income, childcare, economic impacts, beliefs and attitudes, etc.
- The primary outcome being examined must be a COVID-19-specific outcome, as above.

- The study must be designed as an impact evaluation study from primary data (ie, not primarily a predictive or simulation model or meta-analysis).
- The study must be peer reviewed, and published in a peer-reviewed journal indexed by PubMed.
- The study must have the title and abstract available via PubMed at the time of the study start date (November 26).
- The study must be written in English.

These eligibility criteria were designed to identify the literature primarily concerning the quantitative impact of one or more implemented COVID-19 policies on COVID-19 outcomes. Studies in which impact evaluation was secondary to another analysis (such as a hypothetical projection model) were eliminated because they were less relevant to our objectives and/or may not contain sufficient information for evaluation. Categories for types of policies were from the Oxford COVID-19 Government Response Tracker.[20]

Reviewer recruitment, training and communication

Reviewers were recruited through personal contacts and postings on online media. All reviewers had experience in systematic review, quantitative causal inference, epidemiology, econometrics, public health, methods evaluation or policy review. All reviewers participated in two meetings in which the procedures and the review tool were demonstrated. Screening reviewers participated in an additional meeting specific to the screening process. Throughout the main review process, reviewers communicated with the administrators and each other through Slack for any additional clarifications, questions, corrections and procedures. The main administrator (NH), who was also a reviewer, was available to answer general questions and make clarifications, but did not answer questions specific to any given article.

Review phases and procedures

Search strategy

The search terms combined four Boolean-based search terms: (1) COVID-19 research[17] (2) regional government units (eg, country, state, county and specific country, state or province, etc), (3) policy or policies and (4) impact or effect. The full search terms are available in online supplemental appendix 2.

Information sources

The search was limited to published articles in peer-reviewed journals. This was largely to attempt to identify literature that was high quality, relevant, prominent and most applicable to the review guidance. PubMed was chosen as the exclusive indexing source due to the prevalence and prominence of policy impact studies in the health and medical field. Preprints were excluded to limit the volume of studies to be screened and to ensure each had met the standards for publication through peer review. The search was conducted on 26 November 2020.

Study selection

Two reviewers were randomly selected to screen the title and abstract of each article for the inclusion criteria. In the case of a dispute, a third randomly selected reviewer decided on acceptance/rejection. Eight reviewers participated in the screening. Training consisted of a 1-hour instruction meeting, a review of the first 50 items on each reviewers' list of assigned articles, and a brief asynchronous online discussion before conducting the full review.

Full article review

The full article review consisted of two subphases: the independent primary review phase, and a group consensus phase. The independent review phase was designed primarily for the purpose of supporting and facilitating discussion in the consensus discussion, rather than as high stakes definitive review data on its own. The consensus process was considered the primary way in which review data would be generated, rather than synthesis from the independent reviews. A flow diagram of the review process is available in online supplemental appendix 3.

Each article was randomly assigned to 3 of the 23 reviewers in our review pool. Each reviewer independently reviewed each article on their list, first for whether the study met the eligibility criteria, then responding to methods identification and guided strength of evidence questions using the review tool, as described below. Reviewers were able to recuse themselves for any reason, in which case another reviewer was randomly selected. Once all three reviewers had reviewed a given article, all articles that weren't unanimously determined to not meet the inclusion criteria underwent a consensus process.

During the consensus round, the three reviewers were given all three primary reviews for reference, and were tasked with generating a consensus opinion among the group. One randomly selected reviewer was tasked to act as the arbitrator. The arbitrator's primary task was facilitating discussion and for moving the group toward establishing a consensus that represented the collective subjective assessments of the group. If consensus could not be reached, a fourth randomly selected reviewer was brought into the discussion to help resolve disputes.

Review tool for data collection

This review tool and data collection process was an operationalised and lightly adapted version of the COVID-19 health policy impact evaluation review guidance literature, written by the lead authors of this study and released in September 2020 as a preprint.[21] The main adaptation was removing references to the COVID-19 literature. All reviewers were instructed to read and refer to this guidance document to guide their assessments. The full guidance manuscript contains additional explanation and rationale for all parts of this review and the tool, and is available both in the adapted form as was provided to the reviewers in online supplemental file 2 'CHSPER review guidance refs removed.pdf' and in an updated version in

Haber *et al*.[17] The full review tool is attached as online supplemental file 3'review tool final.pdf'.

The review tool consisted of two main parts: methods design categorisation and full review. The review tool and guidance categorises policy causal inference designs based on the structure of their assumed counterfactual. This is assessed through identifying the data structure and comparison(s) being made. There are two main items for this determination: the number of preperiod time points (if any) used to assess prepolicy outcome trends, and whether or not policy regions were compared with non-policy regions. These, and other supporting questions, broadly allowed categorisation of methods into cross-sectional, pre/post, interrupted time series (ITS), difference-in-differences (DiD), comparative ITS (CITS), (randomised) trials or other. Given that most papers have several analyses, reviewers were asked to focus exclusively on the impact evaluation analysis that was used as the primary support for the main conclusion of the article.

Studies categorised as cross-sectional, pre/post, randomised controlled trial designs, and other were included in our sample, but set aside for no further review for the purposes of this research. Cross-sectional and pre/post studies are not considered sufficient to yield well-identified causal inference in the specific context of COVID-19 policy impact evaluation, as explained in the policy impact evaluation guidance documentation. Cross-sectional and pre–post designs were considered inappropriate for policy causal inference for COVID-19 due largely to inability to account for a large number of potential issues, including confounding, epidemic trends and selection biases. Randomised controlled trials were assumed to broadly meet key design checks. Studies categorised as 'other' received no further review, as the review guidance would be unable to assess them. Additional justification and explanation for this decision is available in the review guidance.

For the methods receiving full review (ITS, DiD and CITS), reviewers were asked to identify potential issues and give a category-specific rating. The specific study designs triggered subquestions and/or slightly altered the language of the questions being asked, but all three of the methods design categories shared these four key questions:

- Graphical presentation: 'Does the analysis provide graphical representation of the outcome over time?'
  - Graphical presentation refers to how the authors present the data underlying their impact evaluation method. This is a critical criteria for assessing the potential validity of the assumed model. The key questions here are whether any chart shows the outcome over time and the assumed models of the counterfactuals. To meet a high degree of confidence in this category, graphical displays must show the outcome and connect to the counterfactual construction method.
- Functional form: 'Is the functional form of the model used for the trend in counterfactual infectious disease outcomes (eg, linear, non-parametric, exponential, logarithmic, etc) well-justified and appropriate?'
  - Functional form refers to the statistical functional form of the trend in counterfactual infectious disease outcomes (ie, the assumptions used to construct counterfactual outcomes). This may be a linear function, non-parametric, exponential or logarithmic function, infectious disease model projection or any other functional form. The key criteria here are whether this is discussed and justified in the manuscript, and if so, is it a plausibly appropriate choice given infectious disease outcomes.
- Timing of policy impact: 'Is the date or time threshold set to the appropriate date or time (eg, is there lag between the intervention and outcome)?'
  - Timing of policy impact refers to assumptions about when we would expect to see an impact from the policy vis-a-vis the timing of the policy introduction. This would typically be modelled with leads and lags. The impact of policy can occur before enactment (eg, in cases where behavioural change after policy is announced, but before it takes place in anticipation) or long after the policy is enacted (eg, in cases where it takes time to ramp up policy implementation or impacts). The key criteria here are whether this is discussed and justified in the manuscript, and if so, whether it is a plausibly appropriate choice given the policy and outcome.
- Concurrent changes: 'Is this policy the only uncontrolled or unadjusted-for way in which the outcome could have changed during the measurement period (differently for policy and non-policy regions)?'
  - Concurrent changes refers to the presence of uncontrolled other events and changes that may influence outcomes at the same time as the policy would impact outcomes. In order to assess the impact of one policy or set of policies, the impact of all other forces that differentially impact the outcome must either be negligible or controlled for. The key criteria here are whether it is likely that there are substantial other uncontrolled forces (eg, policies, behavioural changes) which may be differentially impacting outcomes at the same time as the policy of interest.

For each of the four key questions, reviewers were given the option to select 'No,' 'Mostly no,' 'Mostly yes,' and 'Yes' with justification text requested for all answers other than 'Yes.' Each question had additional prompts as guidance, and with much more detail provided in the full guidance document. Ratings are, by design, subjective assessments of the category according to the guidance. We do not use numerical scoring, for similar reasons as Cochrane suggests that the algorithms for summary judgements for the RoB2 tool are merely 'proposed' assessments, which reviewers should change as they believe appropriate.[22] It is entirely plausible, for example, for a study to meet all but one criteria but for the one remaining to be sufficiently violated that the entire

collective category is compromised. Alternatively, there could be many minor violations of all of the criteria, but that they were collectively not sufficiently problematic to impact overall ratings. Further, reviewers were also tasked with considering room for doubt in cases where answers to these questions were unclear.

The criteria were designed to establish minimal plausibility of actionable evidence, rather than certification of high quality. Graphical representation is included here primarily as a key way to assess the plausibility and justification of key model assumptions, rather than being necessary for validity by itself. For example, rather than having the 'right' functional form or lag structure, the review guidance asks whether the functional form and lags is discussed at all and (if discussed) reasonable.

These four questions were selected and designed being critical to evaluating strength of study design for policy impact evaluation in general, direct relevance for COVID-19 policy, feasibility for use in guided review. These questions are designed as minimal and key criteria for plausibly actionable impact evaluation design for COVID-19 policy impact evaluation, rather than as a comprehensive tool assessing overall quality. Thorough review of data quality, statistical validity, and other issues are also critical points of potential weakness in study designs, and would be needed in addition to these criteria, if these key design criteria are met. A thorough justification and explanation of how and why these questions were selected is available in the provided guidance document and in Haber *et al*.[17]

Finally, reviewers were asked a summary question:

- Overall: 'Do you believe that the design is appropriate for identifying the policy impact(s) of interest?'

Reviewers were asked to consider the scale of this question to be both independent/not relative to any other papers, and that any one substantial issue with the study design could render it a 'No' or 'Mostly no.' Reviewers were asked to follow the guidance and their previous answers, allowing for their own weighting of how important each issue was to the final result. A study could be excellent on all dimensions except for one, and that one dimension could render it inappropriate for causal inference. As such, in addition to the overall rating question, we also generated a 'weakest link' metric for overall assessment, representing the lowest rating among the four key questions (graphical representation, functional form, timing of policy impact and concurrent changes). A 'mostly yes' or 'yes' is considered a passing rating, indicating that the study was not found to be inappropriate on the specific dimension of interest.

A 'yes' rating does not necessarily indicate that the study is strongly designed, conducted or is actionable; it only means that it passes a series of key design checks for policy impact evaluation and should be considered for further evaluation. The papers may contain any number of other issues that were not reviewed (eg, statistical issues, inappropriate comparisons, generalisability). As such, this should only be considered an initial assessment of plausibility that the study is well designed, rather than confirmation that it is appropriate and applicable.

Heterogeneity

Inter-rater reliability (IRR) was assessed using Krippendorff's alpha.[23 24] Rather than more typical uses intended as an examination of the 'validity' of ratings, the IRR statistic in this case is being used as a heuristic indicator of heterogeneity between reviewers during the independent phase, where heterogeneity is both expected and not necessarily undesirable. As a second examination of reviewer heterogeneity, we also show the distribution of category differences between primary reviewers within a study (eg, if primary reviewers rated 'Yes,' 'Mostly no,' and 'Mostly yes' there are two pairs of answers that were one category different, and one pair that was two categories different).

Statistical analysis

Statistics provided are nearly exclusively counts and percentages of the final dataset. Analyses and graphics were performed in R.[25] Krippendorff's alpha was calculated using the IRR package.[26]Relative risks were estimated using the epitools package.[27]

Citation counts for accepted articles were obtained through Google Scholar[28] on 11 January 2021. Journal impact factors were obtained from the 2019 Journal Citation Reports.[29]

Data sharing

Data, code, the review tool and the review guidance are stored and available at the OSF.io repository for this study[30] here: https://osfio/9xmke/files/. The dataset includes full results from the search and screening and all review tool responses from reviewers during the full review phase.

Patient and public involvement statement

Patients or public stakeholders were not consulted in the design or conduct of this systematic evaluation.

RESULTS

Search and screening

Figure 1 PRISMA diagram of systematic review process.

After search and screening of titles and abstracts, 102 articles were identified as likely or potentially meeting our inclusion criteria (figure 1). Of those 102 articles, 36 studies met inclusion after independent review and deliberation in the consensus process. The most common reasons for rejection at this stage were that the study did not measure the quantitative direct impact of specific policies and/or that such an impact was not the main purpose of the study. Many of these studies implied that they measured policy impact in the abstract or introduction, but instead measured correlations with secondary outcomes (eg, the effect of movement reductions, which are influenced by policy) and/or performed cursory

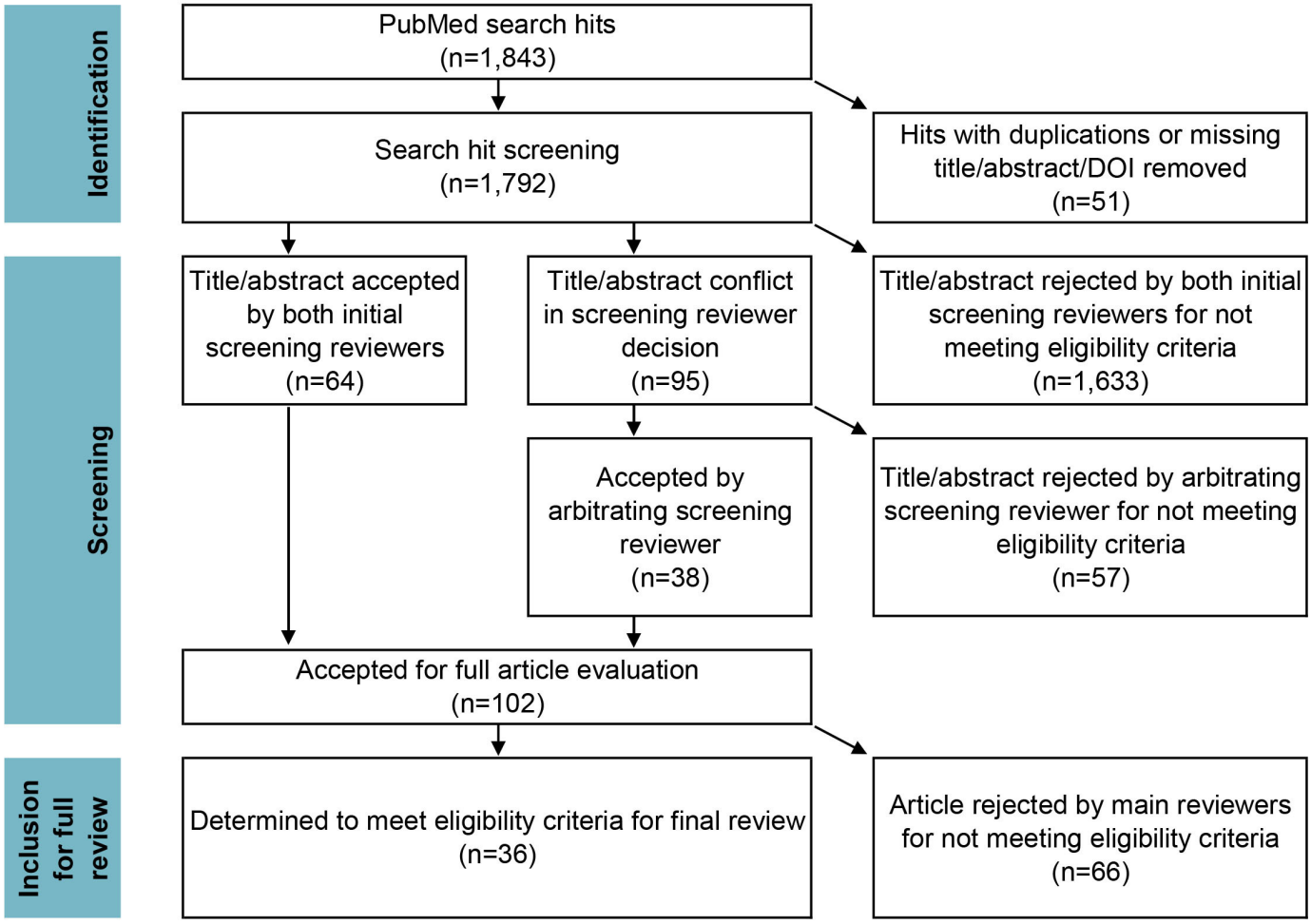


Figure 1 PRISMA diagram of systematic review process. This chart shows the PRISMA diagram for the process of screening the literature from search to the full review phase. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses

policy impact evaluation secondary to projection modelling efforts.

Descriptive statistics

Figure 2 Descriptive sample statistics (n=36).

Publication information from our sample is shown in figure 2. The articles in our sample were generally published in journals with high impact factors (median impact factor: 3.6, 25th percentile: 2.3, 75th percentile: 5.3 IQR: 3.0) and have already been cited in the academic literature (median citation count: 5.0, 25th percentile: 2.0, 75th percentile: 26.8, IQR 24.8, on 1 November 2021). The most commonly evaluated policy type was stay at home requirements (64% n=23/36). Reviewers noted that many articles referenced 'lockdowns,' but did not define the specific policies to which this referred. Reviewers specified mask mandates for three of the studies, and noted either a combination of many interventions or unspecified specific policies in seven cases.

Reviewers most commonly selected interrupted time-series (39% n=14/36) as the methods design, followed by DiD (9% n=9/36) and pre–post (8% n=8/36). There were no randomised controlled trials of COVID-19 health policies identified (0% n=0/36), nor were any studies identified that reviewers could not categorise based on the review guidance (0% n=0/36).

The identified articles and selected review results are summarised in table 1.

Strength of methods assessment

Figure 3 Main consensus results summary for key and overall questions.

Graphical representation of the outcome over time was relatively well-rated in our sample, with 74% (n=20/27) studies being given a 'mostly yes' or 'yes' rating for appropriateness. Reasons cited for non-'yes' ratings included a lack of graphical representation of the data, alternative scales used, and not showing the dates of policy implementation.

Functional form issues appear to have presented a major issue in these studies, with only 19% receiving a 'mostly yes' or 'yes' rating, 78% (n=21/27) receiving a 'no' rating, and 4% (n=1/27) 'unclear.' There were two common themes in this category: studies generally using scales that were broadly considered inappropriate for infectious disease outcomes (eg, linear counts), and/

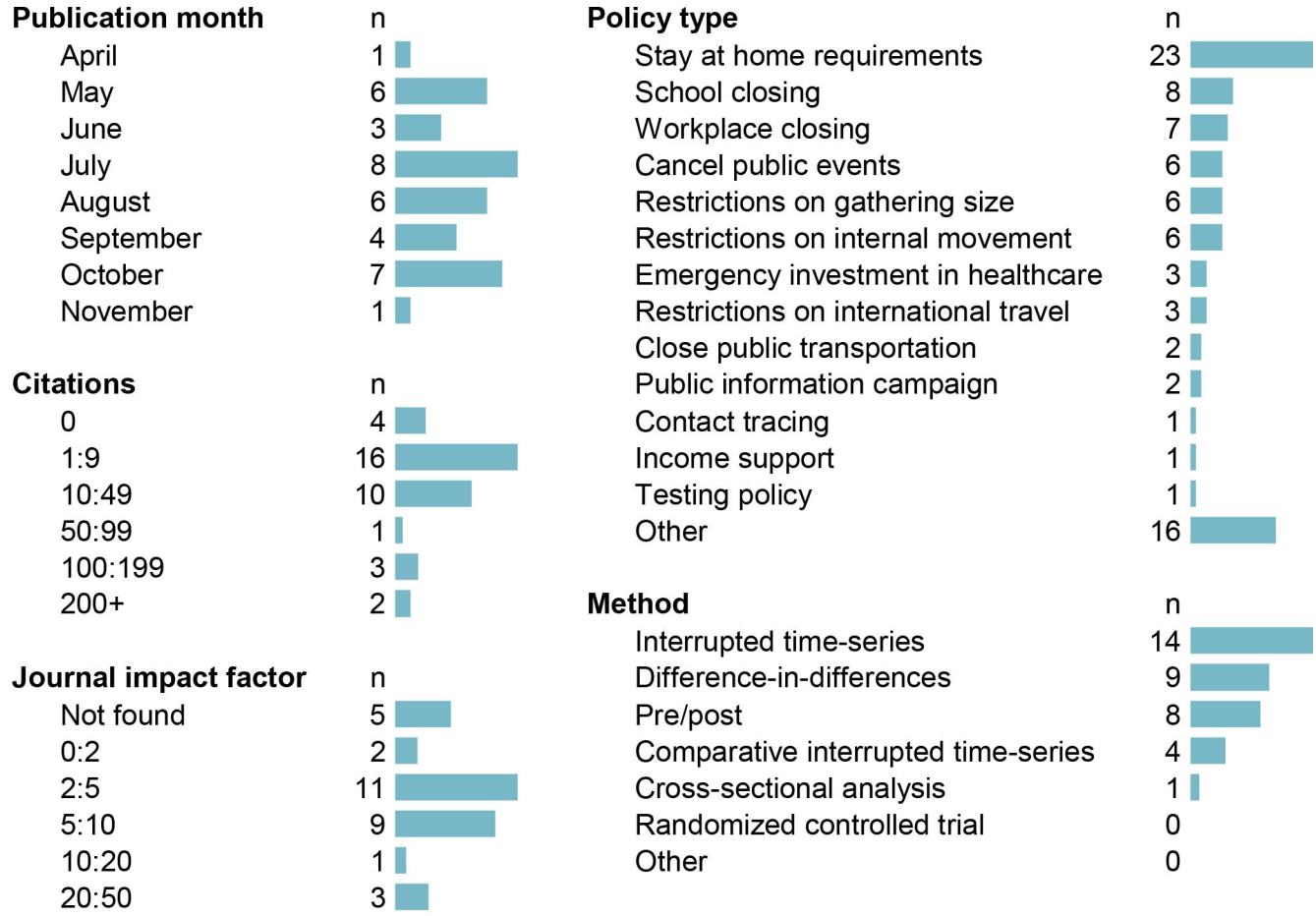


Figure 2 Descriptive sample statistics (n=36). This chart shows descriptive statistics of the 36 studies entered into our systematic evidence review.

or studies lacking stated justification for the scale used. Reviewers also noted disconnects between clear curvature in the outcomes in the graphical representations and the analysis models and outcome scales used (eg, linear). In one case, reviewers could not identify the functional form actually used in analysis.

Reviewers broadly found that these studies dealt with timing of policy impact (eg, lags between policy implementation and expected impact) relatively well, with 70% (n=19/27) rated 'yes' or 'mostly yes.' Reasons for non-'yes' responses included not adjusting for lags and a lack of justification for the specific lags used.

Concurrent changes were found to be a major issue in these studies, with only 11% (n=3/27) studies receiving passing ratings ('yes' or 'mostly yes') with regard to uncontrolled concurrent changes to the outcomes. Reviewers nearly ubiquitously noted that the articles failed to account for the impact of other policies that could have impacted COVID-19 outcomes concurrent with the policies of interest. Other issues cited were largely related to non-policy-induced behavioural and societal changes.

When reviewers were asked if sensitivity analyses had been performed on key assumptions and parameters, about half (56% n=15/27) answered 'mostly yes' or 'yes.' The most common reason for non-'yes' ratings was that, while sensitivity analyses were performed, they did not address the most substantial assumptions and issues.

Overall, reviewers rated only four studies (11%, n=4/36,) as being plausibly appropriate ('mostly yes' or 'yes') for identifying the impact of specific policies on COVID-19 outcomes, as shown in figure 3. 25% (n=9/36) were automatically categorised as being inappropriate due to being either cross-sectional or pre/post in design, 33% (n=12/36) of studies were given a 'no' rating for appropriateness, 31% 'mostly no' (n=11/36), 8% 'mostly yes' (n=3/36), and 3% 'yes' (n=1/36). The most common reason cited for non-'yes' overall ratings was failure to account for concurrent changes (particularly policy and societal changes).

Figure 4 Comparison of independent reviews, weakest link and direct consensus review.

As shown in figure 4, the consensus overall proportion passing ('mostly yes' or 'yes') was a quarter of what it was from the initial independent reviews. Forty-five per cent (n=34/75) of studies were rated as 'yes' or 'mostly yes' in the initial independent review, as compared with 11% (n=4/36) in the consensus round (RR 0.25, 95% CI 0.09 to 0.64). The issues identified and discussed in combination during consensus discussions, as well as additional clarity on the review process, resulted in reduced overall

Table 1 Summary of articles reviewed and reviewer ratings for key and overall questions

Category ratings order		Legend for colour-coded ratings						
Graphical presentation	Timing of policy impact							
Functional form	Concurrent changes	N/A	Unclear	No*	No **	Mostly no	Mostly yes	Yes

Method determined to me inappropriate by: * guidance (cross sectional or pre/post) or ** reviewer consensus

Citation	Title	Journal	Publication date	Methods design	Category ratings		Overall rating
Cobb and Seale, 2020[38]	Examining the effect of social distancing on the compound growth rate of COVID-19 at the county level (USA) using statistical analyses and a random forest machine learning model.	Public Health	4/28/2020	Pre/post			
Lyu and Wehby, 2020[39]	Comparison of Estimated Rates of Coronavirus Disease 2019 (COVID-19) in Border Counties in Iowa Without a Stay-at-Home Order and Border Counties in Illinois With a Stay-at-Home Order.	JAMA Network Open	5/1/2020	Difference-in-differences			
Tam *et al* 2020[40]	Effect of mitigation measures on the spreading of COVID-19 in hard-hit states in the USA.	PloS One	5/1/2020	Interrupted time-series			
Courtemanche *et al* 2020[41]	Strong Social Distancing Measures In The US Reduced The COVID-19 Growth Rate.	Health Affairs	5/14/2020	Difference-in-differences			
Crokidakis 2020[42]	COVID-19 spreading in Rio de Janeiro, Brazil: Do the policies of social isolation really work?	Chaos, Solitons and Fractals	5/23/2020	Interrupted time-series			
Hyafil and Moriña, 2020[43]	Analysis of the impact of lockdown on the reproduction number of the SARS-Cov-2 in Spain.	Gaceta Aanitaria	5/23/2020	Pre/post			
Castillo *et al*, 2020[44]	The effect of state-level stay-at-home orders on COVID-19 infection rates.	American Journal of infection control	5/24/2020	Pre/post			
Alfano and Ercolano, 2020[45]	The Efficacy of Lockdown Against COVID-19: A Cross-Country Panel Analysis.	Applied Health Economics and Health Policy	6/3/2020	Difference-in-differences			
Lyu and Wehby, 2020b[46]	Community Use Of Face Masks And COVID-19: Evidence From A Natural Experiment Of State Mandates In The US.	Health Affairs	6/16/2020	Difference-in-differences			
Zhang *et al*, 2020[47]	Identifying airborne transmission as the dominant route for the spread of COVID-19.	PNAS	6/30/2020	Interrupted time-series			
Xu *et al*, 2020[48]	Associations of Stay-at-Home Order and Face-Masking Recommendation with Trends in Daily New Cases and Deaths of Laboratory-Confirmed COVID-19 in the USA.	Exploratory research and hypothesis in medicine	7/8/2020	Interrupted time-series			
Lyu and Wehby, 2020[49]	Shelter-In-Place Orders Reduced COVID-19 Mortality And Reduced The Rate Of Growth In Hospitalisations.	Health Affairs	7/9/2020	Difference-in-differences			
Wagner *et al*, 2020[50]	Social distancing merely stabilised COVID-19 in the USA.	Stat (International Statistical Institute)	7/13/2020	Interrupted time-series			
Di Bari *et al*, 2020[51]	Extensive Testing May Reduce COVID-19 Mortality: A Lesson From Northern Italy.	Frontiers in Medicine	7/14/2020	Comparative interrupted time-series			
Islam *et al*, 2020[52]	Physical distancing interventions and incidence of coronavirus disease 2019: natural experiment in 149 countries.	BMJ (Clinical research ed.)	7/15/2020	Interrupted time-series			
Wong *et al*, 2020[53]	Impact of National Containment Measures on Decelerating the Increase in Daily New Cases of COVID-19 in 54 countries and 4 Epicentres of the Pandemic: Comparative Observational Study.	Journal of Medical Internet Research	7/22/2020	Pre/post			

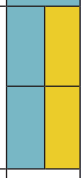
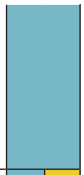
Continued

Table 1 Continued

Citation	Title	Journal	Publication date	Methods design	Category ratings	Overall rating
Liang *et al*, 2020[54]	Effects of policies and containment measures on control of COVID-19 epidemic in Chongqing.	World Journal of Clinical Cases	7/26/2020	Pre/post		
Banerjee and Nayak, 2020[55]	US county level analysis to determine If social distancing slowed the spread of COVID-19.	Pan American Journal of Public Health	7/31/2020	Difference-in-differences		
Dave *et al*, 2020[56]	When Do Shelter-in-Place Orders Fight COVID-19 Best? Policy Heterogeneity Across States and Adoption Time.	Economic inquiry	8/3/2020	Difference-in-differences		
Hsiang *et al*, 2020[57]	The effect of large-scale anticontagion policies on the COVID-19 pandemic.	Nature	8/22/2020	Interrupted time-series		
Lim *et al*, 2020[58]	Revealing regional disparities in the transmission potential of SARS-CoV-2 from interventions in Southeast Asia.	Proceedings. Biological sciences	8/26/2020	Interrupted time-series		
Arshed *et al*, 2020[59]	Empirical assessment of government policies and flattening of the COVID19 curve.	Journal of Public Affairs	8/27/2020	Cross-sectional analysis		
Wang *et al*, 2020[60]	Fangcang shelter hospitals are a One Health approach for responding to the COVID-19 outbreak in Wuhan, China.	One Health	8/29/2020	Interrupted time-series		
Kang and Kim, 2020[61]	The Effects of Border Shutdowns on the Spread of COVID-19.	Journal of Preventive Medicine and Public Health	8/30/2020	Comparative interrupted time-series		
Auger *et al*, 2020[62]	Association Between Statewide School Closure and COVID-19 Incidence and Mortality in the US.	JAMA	9/1/2020	Interrupted time-series		
Santamaría *et al*, 2020[63]	COVID-19 effective reproduction number dropped during Spain's nationwide dropdown, then spiked at lower-incidence regions.	The Science of the Total Environment	9/9/2020	Interrupted time-series		
Bennett, 2020[64]	All things equal? Heterogeneity in policy effectiveness against COVID-19 spread in chile.	World Development	9/24/2020	Comparative interrupted time-series		
Yang *et al*, 2020[65]	Lessons Learnt from China: National Multidisciplinary Healthcare Assistance.	Risk Management and Healthcare Policy	9/30/2020	Difference-in-differences		
Padalabalanarayanan *et al*, 2020[66]	Association of State Stay-at-Home Orders and State-Level African American Population With COVID-19 Case Rates.	JAMA Network Open	10/1/2020	Comparative interrupted time-series		
Edelstein *et al*, 2020[67]	SARS-CoV-2 infection in London, England: changes to community point prevalence around lockdown time, March-May 2020.	Journal of Epidemiology and Community Health	10/1/2020	Pre/post		
Tsai *et al*, 2020[68]	COVID-19 transmission in the U.S. before vs after relaxation of statewide social distancing measures.	Clinical Infectious Diseases	10/3/2020	Interrupted time-series		
Singh *et al*, 2020[69]	Public health interventions slowed but did not halt the spread of COVID-19 in India.	Transboundary and Emerging Diseases	10/4/2020	Pre/post		
Gallaway *et al*,2020[70]	Trends in COVID-19 Incidence After Implementation of Mitigation Measures - Arizona, January 22-August 7, 2020.	Morbidity and Mortality Weekly Report	10/9/2020	Pre/post		
Castex *et al*, 2020[71]	COVID-19: The impact of social distancing policies, cross-country analysis.	Economics of Disasters and Climate Change	10/15/2020	Interrupted time-series		
Silva *et al*, 2020[72]	The effect of lockdown on the COVID-19 epidemic in Brazil: evidence from an interrupted time series design.	Cadernos de Saude Publica	10/19/2020	Interrupted time-series		

Continued

Table 1 Continued

Citation	Title	Journal	Publication date	Methods design	Category ratings	Overall rating
Dave *et al*, 2020[73]	Were Urban Cowboys Enough to Control COVID-19? Local Shelter-in-Place Orders and Coronavirus Case Growth.	Journal of Urban Economics	11/6/2020	Difference-in-differences		

confidence in the findings. Increased clarity on the review guidance with experience and time may also have reduced these ratings further.

The large majority of studies had at least one 'no' or 'unclear' rating in one of the four categories (74% n=20/27), with only one study whose lowest rating was a 'mostly yes,' no studies rated 'yes' in all four categories. Only one study was found to pass design criteria in all four key questions categories, as shown in the 'weakest link' column in figure 4.

Review process assessment

During independent review, all three reviewers independently came to the same conclusions on the main methods design category for 33% (n=12/36) articles, two out of the three reviewers agreed for 44% (n=16/36) articles, and none of the reviewers agreed in 22% (n=8/36) cases. One major contributor to these discrepancies were the 31% (n=11/36) cases where one or more reviewers marked the study as not meeting eligibility criteria, 64% (n=7/11) of which the other two reviewers agreed on the methods design category.

Reviewers' initial independent reviews were heterogeneous for key rating questions. For the overall scores, Krippendorff's alpha was only 0.16 due to widely varying opinions between raters. The four key categorical questions had slightly better IRR than the overall question, with Krippendoff's alphas of 0.59 for graphical representation, 0.34 for functional form, 0.44 for timing of policy impact, and 0.15 for concurrent changes, respectively.For the main summary rating, primary reviewers within each study agreed in 26% of cases (n=16), were one category different in 45% (n=46), two categories different in 19% (n=12), and three categories (ie, the maximum distance, 'Yes' vs 'No') in 10% of cases (n=6).

The consensus rating for overall strength was equal to the lowest rating among the independent reviews in 78% (n=21/27) of cases, and only one higher than the lowest in the remaining 22% (n=6/27). This strongly suggests that the multiple reviewer review, discussion, and consensus process more thoroughly identifies issues than independent review alone. There were two cases for which reviewers requested an additional fourth reviewer to help resolve standing issues for which the reviewers felt they were unable to come to consensus.

The most consistent point of feedback from reviewers was the value of having a three reviewer team with whom to discuss and deliberate, rather than two as initially planned. This was reported to help catch a larger number of issues and clarify both the papers and the interpretation of the review tool questions. Reviewers also expressed that one of the most difficult parts of this process was assessing the inclusion criteria, some of the implications of which are discussed below.

DISCUSSION

This systematic review of evidence strength found that only four (or only one by a stricter standard) of the 36

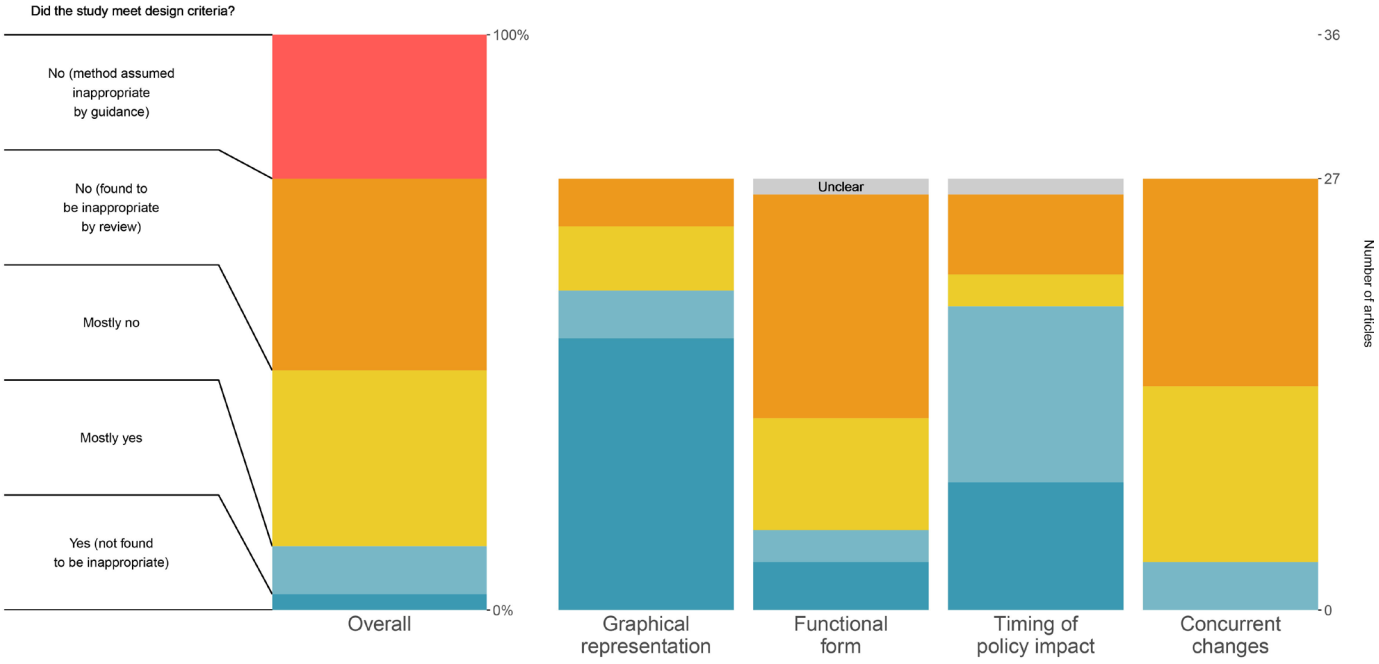


Figure 3 Main consensus results summary for key and overall questions. This chart shows the final overall ratings (left) and the key design question ratings for the consensus review of the 36 included studies, answering the degree to which the articles met the given key design question criteria. The key design question ratings were not asked for the nine included articles which selected methods assumed by the guidance to be non-appropriate. The question prompt in the figure is shortened for clarity, where the full prompt for each key question is available in the Methods section.

identified published and peer-reviewed health policy impact evaluation studies passed a set of key checks for identifying the causal impact of policies on COVID-19 outcomes. Because this systematic review examined a limited set of key study design features and did not address more detailed aspects of study design, statistical issues, generalisability and any number of other issues, this result may be considered an upper bound on the overall strength of evidence within this sample. Two major problems are nearly ubiquitous throughout this literature: failure to isolate the impact of the policy(s) of interest from other changes that were occurring contemporaneously, and failure to appropriately address the functional form of infectious disease outcomes in a population setting. While policy decisions are being made on the backs of high impact-factor papers, we find that the citation-based metrics do not correspond to 'quality' research as used by Yin *et al*.[31] Similar to other areas in the COVID-19 literature,[32] we found the current literature directly evaluating the impact of COVID-19 policies largely fails to meet key design criteria for actionable inference to inform policy decisions.

The framework for the review tool is based on the requirements and assumptions built into policy evaluation methods. Quasi-experimental methods rely critically on the scenarios in which the data are generated. These assumptions and the circumstances in which they are plausible are well-documented and understood,[2 4–6 17 33] including one paper discussing application of DiD methods specifically for COVID-19 health policy, released in May 2020.[5] While 'no uncontrolled concurrent changes' is a difficult bar to clear, that bar is fundamental to inference using these methods.

The circumstances of isolating the impact of policies in COVID-19 - including large numbers of policies, infectious disease dynamics and massive changes to social behaviours—make those already difficult fundamental assumptions broadly much less likely to be met. Some of the studies in our sample were nearly the best feasible studies that could be done given the circumstances, but the best that can be done often yields little actionable inference. The relative paucity of strong studies does not in any way imply a lack of impact of those policies; only that we lack the circumstances to have evaluated their effects.

Because the studies estimating the harms of policies share the same fundamental circumstances, the evidence of COVID-19 policy harms is likely to be of similarly poor strength. Identifying the effects of many of these policies, particularly for the spring of 2020, is likely to be unknown and perhaps unknowable. However, there remains additional opportunities with more favourable circumstances, such as measuring overall impact of NPIs as bundles, rather than individual policies. Similarly, studies estimating the impact of reopening policies or policy cancellation are likely to have fewer concurrent changes to address.

The review process itself demonstrates how guided and targeted peer review can efficiently evaluate studies in ways that the traditional peer review systems do not. The studies in our sample had passed the full

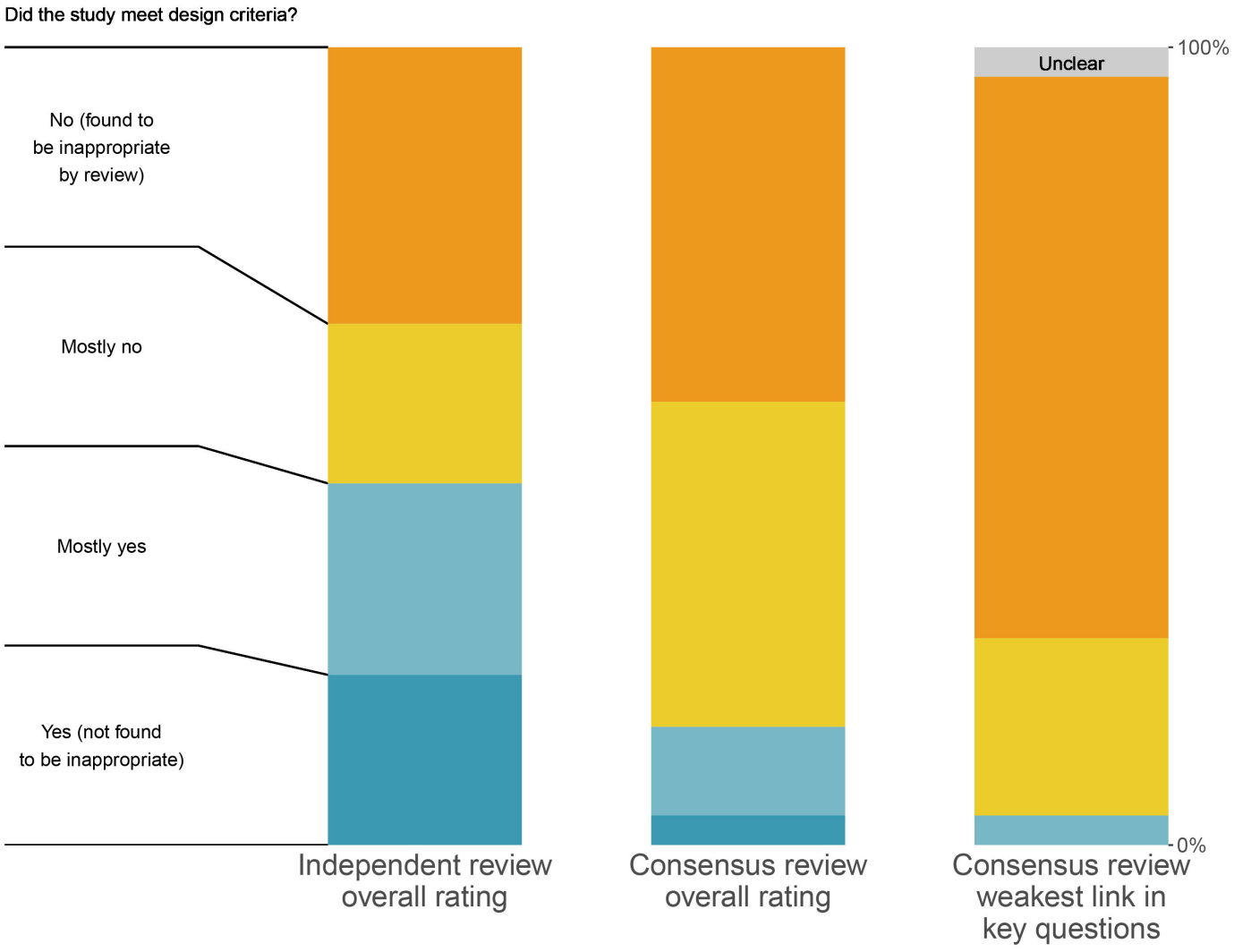


Figure 4 Comparison of independent reviews, weakest link and direct consensus review. This chart shows the final overall ratings by three different possible metrics. The first column contains all of the independent review ratings for the 27 studies which were eventually included in our sample, noting that reviewers who either selected them as not meeting inclusion criteria or selected a method that did not receive the full review did not contribute. The middle column contains the final consensus reviews among the 27 articles which received full review. The last column contains the weakest link rating, as described in the Methods section. The question prompt in the figure is shortened for clarity, where the full prompt for each key question is available in the Methods section.

peer review process, were published in largely high-profile journals, and are highly cited, but contained substantial flaws that rendered their inference utility questionable. The relatively small number of studies included, as compared with the size of the literature concerning itself with COVID-19 policy, may suggest that there was relative restraint from journal editors and reviewers for publishing these types of studies. The large number of models, but relatively small number of primary evaluation analyses is consistent with other areas of COVID-19.[34 35] At minimum, the flaws and limitations in their inference could have been communicated at the time of publication, when they are needed most. In other cases, it is plausible that many of these studies would not have been published had a more thorough or more targeted methodological review been performed.

This systematic review of evidence strength has limitations. The tool itself was limited to a very narrow—although critical—set of items. Low ratings in our study should not be interpreted as being overall poor studies, as they may make other contributions to the literature that we did not evaluate. While the guidance and tool provided a well-structured framework and our reviewer pool was well qualified, strength of evidence review is inherently subjective. It is plausible and likely that other sets of reviewers would come to different conclusions for each study, but unlikely that the overall conclusions of our assessment would change substantially. However, the consensus process was designed with subjectivity in

mind, and demonstrates the value of consensus processes for overcoming hurdles with subjective and difficult decisions.

While subjective assessments are inherently subject to the technical expertise, experiences, and opinions of reviewers, we argue they are both appropriate and necessary to reliably assess strength of evidence based on theoretical methodological issues. With the exception of the graphical assessment, proper assessment of the core methodological issues requires that reviewers are able to weigh the evidence as they see fit. Much like standard institutional peer review, reviewers independently had highly heterogeneous opinions, attributable to differences in opinion or training, misunderstandings/learning about the review tool and process, and expected reliance on the consensus process. Unlike traditional peer review, there was subject-matter-specific guidance and a process to consolidate and discuss those heterogeneous initial opinions. The reduction in ratings from the initial highly heterogeneous ratings to a lower heterogeneity in ratings indicates that reviewers had initially identified issues differently, but that the discussion and consensus process helped elucidate the extent of the different issues that each reviewer detected and brought to discussion. This also reflects reviewer learning over time, where reviewers were better able to identify issues at the consensus phase than earlier. It is plausible that stronger opinions had more weight, but we expect that this was largely mitigated by the random assignment of the arbitrator, and reviewer experiences did not indicate this as an issue.

Most importantly, this review does not cover all policy inference in the scientific literature. One large literature from which there may be COVID-19 policy evaluation otherwise meeting our inclusion criteria are preprints. Many preprints would likely fare well in our review process. Higher strength papers often require more time for review and publication, and many high-quality papers may be in the publication pipeline now. Second, this review excluded studies that had a quantitative impact evaluation as a secondary part of the study (eg, to estimate parameters for micro-simulation or disease modelling). Third, the review does not include policy inference studies that do not measure the impact of a specific policy. For instance, there are studies that estimate the impact of reduced mobility on COVID-19 outcomes but do not attribute the reduced mobility to any specific policy change. A considerable number of studies that present analyses of COVID-19 outcomes to inform policy are excluded because they do not present a quantitative estimate of specific policies' treatment effects. Importantly, this study was designed to assess a minimal set of criteria critical to the design of impact evaluation studies of COVID-19 policies. Studies found meeting these criteria would require further and more comprehensive review for assessing overall quality and actionability. Unfortunately, exceedingly few studies we reviewed, taken largely from the high-profile literature, were found to meet these minimal criteria.

While COVID-19 policy is one of the most important problems of our time, the circumstances under which those policies were enacted severely hamper our ability to study and understand their effects. Claimed conclusions are only as valuable as the methods by which they are produced. Replicable, rigorous, intense and methodologically guided review is needed to both communicate our limitations and make more actionable inference. Weak, unreliable and overconfident evidence leads to poor decisions and undermines trust in science.[15 36] In the case of COVID-19 health policy, a frank appraisal of the strength of the studies on which policies are based is needed, alongside the understanding that we often must make decisions when strong evidence is not feasible.[37]

Author affiliations
[1]Meta Research Innovation Center at Stanford University (METRICS), Stanford University, Stanford, California, USA
[2]Department of Global Health and Population, Harvard University T H Chan School of Public Health, Boston, Massachusetts, USA
[3]Department of Statistics, Goldman School of Public Policy, University of California Berkeley, Berkeley, California, USA
[4]Department of Global Health, George Washington University School of Public Health and Health Services, Washington, District of Columbia, USA
[5]Department of Health Policy, Stanford University, Stanford, CA, USA
[6]Department of Health Policy and Management, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA
[7]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Québec, Canada
[8]Health Policy and Management, Columbia University Mailman School of Public Health, New York, New York, USA
[9]Department of Biostatistics, Harvard Medical School, Boston, Massachusetts, USA
[10]Department of Health Policy, Vanderbilt University, Nashville, Tennessee, USA
[11]Department of Epidemiology, Harvard University T H Chan School of Public Health, Boston, Massachusetts, USA
[12]Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA
[13]Department of Epidemiology, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA
[14]Center for Applied Public Health and Research, RTI International, Washington, DC, USA
[15]Department of Mental Health, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA
[16]School of Public Health, The University of Sydney, Sydney, New South Wales, Australia
[17]RAND Corp, Santa Monica, California, USA
[18]Interfaculty Initiative in Health Policy, Harvard University Graduate School of Arts and Sciences, Cambridge, Massachusetts, USA

Twitter Noah A Haber @NoahHaber and Brooke Jarrett @theoriginalbrk

Acknowledgements We would like to thank Keletso Makofane for assisting with the screening, Dr Steven Goodman and Dr John Ioannidis for their support during the development of this study, and Dr Lars Hemkins and Dr Mario Malicki for helpful comments in the protocol development.

Contributors NH led the protocol development, study design, administration, data curation, data management, statistical analysis, graphical design, manuscript writing and manuscript editing, and serves as the primary guarantor of the study. NH, EC-D, JS, AF and EMS cowrote the review guidance on which the design of the study review tool is based. NH, EC-D, JS, AF, ERS and EMS designed, wrote and supported the preregistered protocol. NH, CMJ, SEW, CBB, CA, CB-F, VTN and Keletso Makofane were the screening reviewers for this study, analysing the abstracts and titles for inclusion criteria. NH, EC-D, AF, BM-G, EAS, CB-F, JRD, LAH,

CEF, CBB, EB-M, CMJ, BL, IS, EHA, SEW, BJ, CA, VTN, BAG, AB and EAS were the main reviewers for this study, and contributed to the analysis and evaluation of the studies entering into the main review phase. NH, EC-D, JS, AF, BM-G, ERS, CB-F, JRD, LAH, CEF, CB-F, EB-M, CMJ, BSL, IS, EHA, SEW, BJ, CA, VTN, BAG, AB and EAS all contributed to and supported the manuscript editing.

Funding This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. EMS receives funding under the National Institutes of Health grant T32MH109436. IS receives funding under the National Institutes of Health grant T32MH122357. BJ receives funding under the National Institutes of Health grant MH121128. CBB receives funding under the National Institutes of Health grant T32HL098048. CA receives funding from the Knut and Alice Wallenberg Foundation, grant KAW 2019.0561. BAG and EAS were supported by award number P50DA046351 from the National Institute on Drug Abuse. EAS's time was also supported by the Bloomberg American Health Initiative. Caroline Joyce receives funding from the Ferring Foundation. Meta-Research Innovation Center at Stanford (METRICS), Stanford University is supported by a grant from the Laura and John Arnold Foundation

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data, code, the review tool and the review guidance are stored and available at the OSF.io repository for this study[30] here: https://osf.io/9xmke/files/. The dataset includes full results from the search and screening and all review tool responses from reviewers during the full review phase.

ORCID iDs
Noah A Haber http://orcid.org/0000-0002-5672-1769
Laura Anne Hatfield http://orcid.org/0000-0003-0366-3929
Brooke Jarrett http://orcid.org/0000-0003-2966-3521

REFERENCES

1 Fischhoff B. Making decisions in a COVID-19 world. *JAMA* 2020;324:139.
2 COVID-19 Statistics, Policy modeling, and Epidemiology Collective. Defining high-value information for COVID-19 decision-making. *Health Policy* 2020.
3 Hernán MA, Robins JM. *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC.
4 Angrist J, Pischke J-S. *Mostly harmless econometrics: an empiricist's companion*. 1st edn. Princeton University Press, 2009. https://EconPapers.repec.org/RePEc:pup:pbooks:8769
5 Goodman-Bacon A, Marcus J. Using difference-in-differences to identify causal effects of COVID-19 policies. *SSRN Journal* 2020.
6 Bärnighausen T, Oldenburg C, Tugwell P, *et al*. Quasi-experimental study designs series-paper 7: assessing the assumptions. *J Clin Epidemiol* 2017;89:53–66.
7 Haushofer J, Metcalf CJE. Which interventions work best in a pandemic? *Science* 2020;368:1063–5.
8 Else H. How a torrent of COVID science changed research publishing - in seven charts. *Nature* 2020;588:553.
9 Palayew A, Norgaard O, Safreed-Harmon K, *et al*. Pandemic publishing poses a new COVID-19 challenge. *Nat Hum Behav* 2020;4:666–9.
10 Bagdasarian N, Cross GB, Fisher D. Rapid publications risk the integrity of science in the era of COVID-19. *BMC Med* 2020;18:192.
11 Yeo-Teh NSL, Tang BL. An alarming retraction rate for scientific publications on coronavirus disease 2019 (COVID-19). *Account Res* 2020;0:1–7.
12 Abritis A, Marcus A, Oransky I. An "alarming" and "exceptionally high" rate of COVID-19 retractions? *Account Res* 2021;28:58–9.
13 Zdravkovic M, Berger-Estilita J, Zdravkovic B, *et al*. Scientific quality of COVID-19 and SARS CoV-2 publications in the highest impact medical journals during the early phase of the pandemic: a case control study. *PLoS One* 2020;15:e0241826.
14 Elgendy IY, Nimri N, Barakat AF, *et al*. A systematic bias assessment of top-cited full-length original clinical investigations related to COVID-19. *Eur J Intern Med* 2021;86:104–6.
15 Glasziou PP, Sanders S, Hoffmann T. Waste in covid-19 research. *BMJ* 2020;369:m1847.
16 Powell M, Koenecke A, Byrd JB. A how-to guide for conducting retrospective analyses: example COVID-19 study. *Open Science Framework* 2020.
17 Haber NA, Clarke-Deelder E, Salomon JA. COVID-19 policy impact evaluation: a guide to common design issues. *Am J Epidemiol* 2021:kwab185.
18 Haber N. Systematic review of COVID-19 policy evaluation methods and design. Available: https://osf.io/7nbk6 [Accessed 15 Jan 2021].
19 PRISMA. Available: http://www.prisma-statement.org/PRISMAStatement/ [Accessed 15 Jan 2021].
20 Petherick A, Kira B, Hale T. Variation in government responses to COVID-19. Available: https://www.bsg.ox.ac.uk/research/publications/variation-government-responses-covid-19 [Accessed 24 Nov 2020].
21 Haber NA, Clarke-Deelder E, Salomon JA. Policy evaluation in COVID-19: a guide to common design issues. *arXiv:200901940 [stat]* http://arxiv.org/abs/2009.01940
22 Chapter 8: Assessing risk of bias in a randomized trial. Available: https://training.cochrane.org/handbook/current/chapter-08 [Accessed 8 Sep 2021].
23 Krippendorff KH. *Content analysis: an introduction to its methodology*. SAGE Publications, 1980.
24 Zhao X, Liu JS, Deng K. Assumptions behind Intercoder reliability indices. *Ann Int Commun Assoc* 2013;36:419–80.
25 R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R foundation for statistical computing, 2019. Available: https://www.R-project.org/
26 Gamer M, Lemon J, Fellows I. Irr: various coefficients of interrater reliability and agreement. Available: https://cran.r-project.org/web/packages/irr/index.html
27 Aragon TJ, Fay MP, Wollschlaeger D. Epitools, 2017. Available: https://cran.r-project.org/web/packages/epitools/epitools.pdf
28 About Google Scholar. Available: https://scholar.google.com/intl/en/scholar/about.html [Accessed 15 Jan 2021].
29 Clarivate analytics. *J Citation Report* 2019.
30 Haber N. Data repository for systematic review of COVID-19 policy evaluation methods and design, 2020. Available: https://osf.io/9xmke/files [Accessed 9 Nov 2021].
31 Yin Y, Gao J, Jones BF, *et al*. Coevolution of policy and science during the pandemic. *Science* 2021;371:128–30.
32 Wynants L, Van Calster B, Collins GS, *et al*. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020;369:m1328.
33 Clarke GM, Conti S, Wolters AT, *et al*. Evaluating the impact of healthcare interventions using routine data. *BMJ* 2019;365:l2239.
34 Krishnaratne S, Pfadenhauer LM, Coenen M. Measures implemented in the school setting to contain the COVID-19 pandemic: a rapid scoping review. *Cochrane Database System Rev*.
35 Raynaud M, Zhang H, Louis K, *et al*. COVID-19-related medical research: a meta-research and critical appraisal. *BMC Med Res Methodol* 2021;21:1.
36 Casigliani V, De Nard F, De Vita E, *et al*. Too much information, too little evidence: is waste in research fuelling the covid-19 infodemic? *BMJ* 2020;370:m2672.
37 Greenhalgh T. Will COVID-19 be evidence-based medicine's nemesis? *PLoS Med* 2020;17:e1003266.
38 Cobb JS, Seale MA. Examining the effect of social distancing on the compound growth rate of COVID-19 at the County level (United States) using statistical analyses and a random forest machine learning model. *Public Health* 2020;185:27–9.
39 Lyu W, Wehby GL. Comparison of estimated rates of coronavirus disease 2019 (COVID-19) in border counties in Iowa without a Stay-at-Home order and border counties in Illinois with a Stay-at-Home order. *JAMA Netw Open* 2020;3:e2011102.

40 Tam K-M, Walker N, Moreno J. Effect of mitigation measures on the spreading of COVID-19 in hard-hit states in the U.S. *PLoS One* 2020;15:e0240877.
41 Courtemanche C, Garuccio J, Le A. Strong social distancing measures in the United States reduced the COVID-19 growth rate: study evaluates the impact of social distancing measures on the growth rate of confirmed COVID-19 cases across the United States. *Health Affairs* 2020;39:1237–46.
42 Crokidakis N. COVID-19 spreading in Rio de Janeiro, Brazil: do the policies of social isolation really work? *Chaos Solitons Fractals* 2020;136:109930.
43 Hyafil A, Moriña D. Analysis of the impact of lockdown on the reproduction number of the SARS-Cov-2 in Spain. *Gac Sanit* 2021;35:S0213911120300984.
44 Castillo RC, Staguhn ED, Weston-Farber E. The effect of state-level stay-at-home orders on COVID-19 infection rates. *Am J Infect Control* 2020;48:958–60.
45 Alfano V, Ercolano S. The efficacy of Lockdown against COVID-19: a Cross-Country panel analysis. *Appl Health Econ Health Policy* 2020;18:509–17.
46 Lyu W, Wehby GL. Community use of face masks and COVID-19: evidence from a natural experiment of state mandates in the US: study examines impact on COVID-19 growth rates associated with state government mandates requiring face mask use in public. *Health Affairs* 2020;39:1419–25.
47 Zhang R, Li Y, Zhang AL, *et al*. Identifying airborne transmission as the dominant route for the spread of COVID-19. *Proc Natl Acad Sci U S A* 2020;117:14857–63.
48 Xu J, Hussain S, Lu G, *et al*. Associations of Stay-at-Home order and Face-Masking recommendation with trends in daily new cases and deaths of Laboratory-Confirmed COVID-19 in the United States. *Explor Res Hypothesis Med* 2020;000:1–10.
49 Lyu W, Wehby GL. Shelter-In-Place orders reduced COVID-19 mortality and reduced the rate of growth in hospitalizations. *Health Aff* 2020;39:1615–23.
50 Wagner AB, Hill EL, Ryan SE, *et al*. Social distancing merely stabilized COVID-19 in the United States. *Stat*2020;9.
51 Di Bari M, Balzi D, Carreras G, *et al*. Extensive testing may reduce COVID-19 mortality: a lesson from northern Italy. *Front Med* 2020;7:402.
52 Islam N, Sharp SJ, Chowell G, *et al*. Physical distancing interventions and incidence of coronavirus disease 2019: natural experiment in 149 countries. *BMJ* 2020;370:m2743.
53 Wong LP, Alias H. Temporal changes in psychobehavioural responses during the early phase of the COVID-19 pandemic in Malaysia. *J Behav Med* 2021;44:1–11.
54 Liang X-H, Tang X, Luo Y-T, *et al*. Effects of policies and containment measures on control of COVID-19 epidemic in Chongqing. *World J Clin Cases* 2020;8:2959–76.
55 Banerjee T, Nayak A. U.S. county level analysis to determine if social distancing slowed the spread of COVID-19. *Revista Panamericana de Salud Pública* 2020;44:1.
56 Dave D, Friedson AI, Matsuzawa K. When do shelter-in-place orders fight COVID-19 best? Policy heterogeneity across states and adoption time. *Econ Inq*.
57 Hsiang S, Allen D, Annan-Phan S, *et al*. The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature* 2020;584:262–7.
58 Lim JT, Dickens BSL, Choo ELW, *et al*. Revealing regional disparities in the transmission potential of SARS-CoV-2 from interventions in Southeast Asia. *Proc Biol Sci* 2020;287:20201173.
59 Arshed N, Meo MS, Farooq F. Empirical assessment of government policies and flattening of the COVID 19 curve. *J Public Aff*;7.
60 Wang K-W, Gao J, Song X-X, *et al*. Fangcang shelter hospitals are a one health approach for responding to the COVID-19 outbreak in Wuhan, China. *One Health* 2020;10:100167.
61 Kang N, Kim B. The effects of border shutdowns on the spread of COVID-19. *J Prev Med Public Health* 2020;53:293–301.
62 Auger KA, Shah SS, Richardson T, *et al*. Association between statewide school closure and COVID-19 incidence and mortality in the US. *JAMA* 2020;324:859.
63 Santamaría L, Hortal J. COVID-19 effective reproduction number dropped during Spain's nationwide dropdown, then spiked at lower-incidence regions. *Sci Total Environ* 2021;751:142257.
64 Bennett M. All things equal? Heterogeneity in policy effectiveness against COVID-19 spread in Chile. *World Dev* 2021;137:105208.
65 Yang T, Shi H, Liu J, *et al*. Lessons learnt from China: national multidisciplinary healthcare assistance. *Risk Manag Healthc Policy* 2020;13:1835–7.
66 Padalabalanarayanan S, Hanumanthu VS, Sen B. Association of state stay-at-home orders and state-level African American population with COVID-19 case rates. *JAMA Netw Open* 2020;3:e2026010.
67 Edelstein M, Obi C, Chand M, *et al*. SARS-CoV-2 infection in London, England: changes to community point prevalence around lockdown time, March–May 2020. *J Epidemiol Community Health* 2020;2:jech-2020-214730.
68 Tsai AC, Harling G, Reynolds Z. COVID-19 transmission in the U.S. before vs. after relaxation of statewide social distancing measures. *Clin Infect Dis*.
69 Singh BB, Lowerison M, Lewinson RT. Public health interventions slowed but did not halt the spread of COVID-19 in India. *Transbound Emerg Dis*.
70 Gallaway MS, Rigler J, Robinson S. Trends in COVID-19 incidence after implementation of mitigation measures — Arizona, January 22–August 7, 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:1460–3.
71 Castex G, Dechter E, Lorca M. COVID-19: the impact of social distancing policies, cross-country analysis. *Econ Disaster Clim Chang* 2020:1–25.
72 Silva L, Figueiredo Filho D, Fernandes A. The effect of lockdown on the COVID-19 epidemic in Brazil: evidence from an interrupted time series design. *Cad Saúde Pública* 2020;36:e00213920.
73 Dave D, Friedson A, Matsuzawa K, *et al*. Were urban cowboys enough to control COVID-19? Local shelter-in-place orders and coronavirus case growth. *J Urban Econ* 2020;103294:103294.

Appendix 1: Changes from pre-registered protocol and justifications

The full, original pre-registered protocol is available here: https://osf.io/7nbk6

Inclusion criteria

Minor language edits were made to the inclusion criteria to improve clarity and fix grammatical and typographical errors. This largely centered around improving clarity that a study must estimate the quantitative impact of policies that had already been enacted. The word "quantitative" was not explicitly stated in the original version.

Procedures

The original protocol specified that each article would receive two independent reviewers. This was increased to three reviewers per article once it became clear both that the number of articles which would be accepted for full review was lower than expectations, and that there would be substantial differences in opinion between reviewers.

Statistical analysis

Firstly, the original protocol specified that 95% confidence intervals would be calculated. However, after further discussion and review, we determined that sampling-based confidence intervals were not appropriate. Our results are not indicative nor intended to be representative of any super- or target-population, and as such sampling-based error is not an appropriate metric for the conclusions of this study.

Secondly, the original protocol specified Kappa-based interrater reliability statistics. However, using three reviewers, rather than the originally registered two, meant that most Kappa statistics would not be appropriate for our review process. Given the three-rater, four-level ordinal scale used, we opted instead to use Krippendorff's Alpha.

Review tool

A number of changes were made to the review tool during the course of the review process. While the original protocol included logic to allow pre/post for review in some of the key questions, this was removed for consistency with the guidance document.

The remaining changes to the review tool were error corrections and clarifications (e.g. correcting the text for the concurrent changes sections in difference-in-differences so that it

stated “uncontrolled” concurrent changes, and distinguishing the DiD/CITS requirements from the ITS requirements to emphasize differential concurrent changes).

Appendix 2: Full search terms

Note: The search filter for COVID-19 and SARS-CoV-2 were the exact search terms used for the National Library of Medicine one-click search option at the time of the protocol development and when the search took place. This reflects that some of the early literature referred to Wuhan specifically (both in geographic reference for where the SARS-CoV-2 was initially found, and unfortunately also early naming of the virus/disease) before official naming conventions became ubiquitous in the literature. In order to comprehensively capture the literature and use searching best practices, we used the most standard and recommended terms.

((((wuhan[All Fields] AND ("coronavirus"[MeSH Terms] OR "coronavirus"[All Fields])) AND 2019/12[PDAT] : 2030[PDAT]) OR 2019-nCoV[All Fields] OR 2019nCoV[All Fields] OR COVID-19[All Fields] OR SARS-CoV-2[All Fields])

AND ("impact*"[TIAB] OR "effect*"[TIAB])

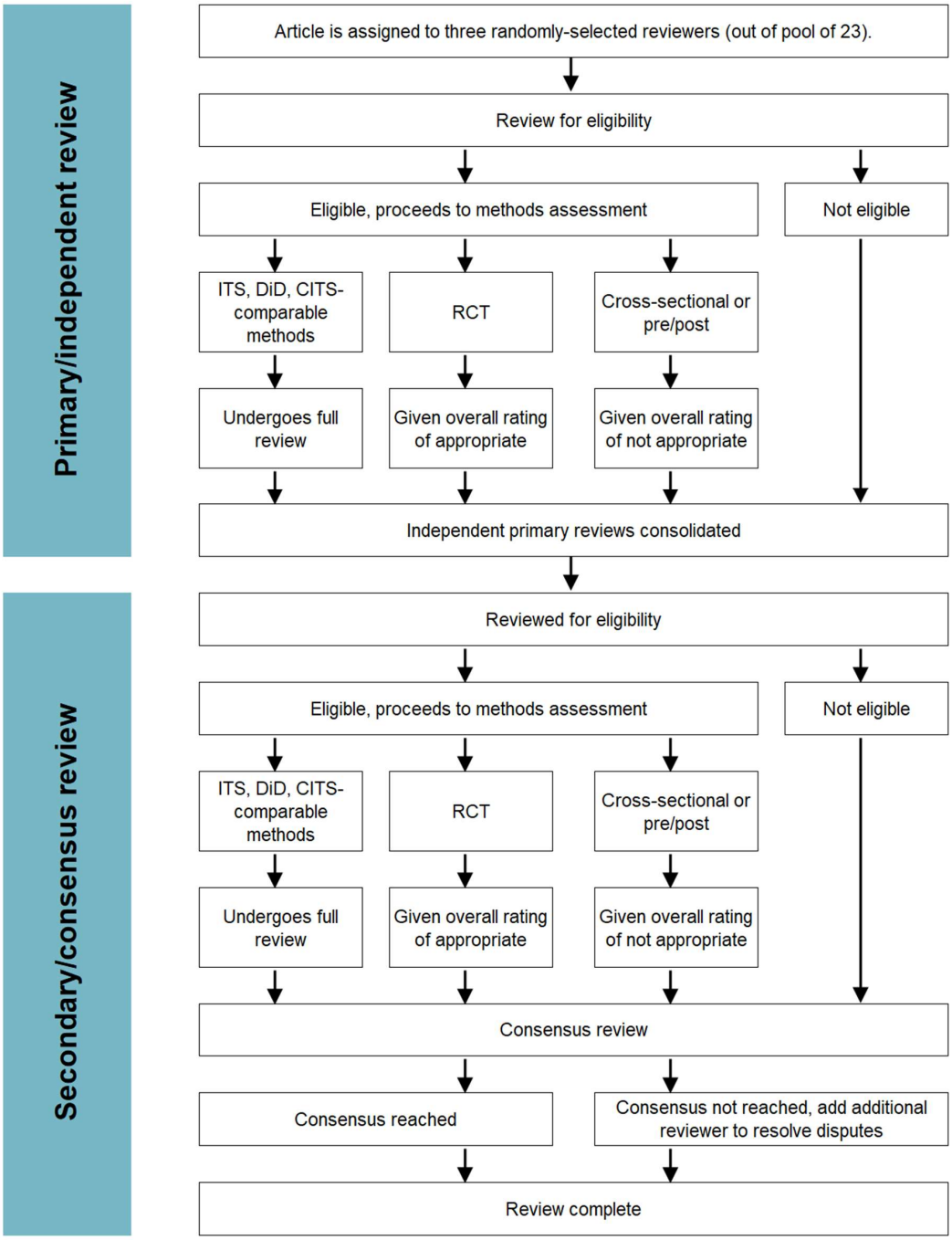
AND ("policy"[TIAB] OR "policies"[TIAB] OR "order*"[TIAB] OR "mandate*"[TIAB])

AND ("countries"[TIAB] OR "country"[TIAB] OR "state"[TIAB] OR "provinc*"[TIAB] OR "county"[TIAB] OR "parish"[TIAB] OR "region*"[TIAB] OR "city"[TIAB] OR "cities"[TIAB] OR "continent*"[TIAB] "Asia*"[TIAB] OR "Europe*"[TIAB] OR "Africa*"[TIAB] OR "America*"[TIAB] OR "Australia"[TIAB] OR "Antarctica"[TIAB] OR "Afghanistan"[TIAB] OR "Aland Islands"[TIAB] OR "Åland Islands"[TIAB] OR "Albania"[TIAB] OR "Algeria"[TIAB] OR "American Samoa"[TIAB] OR "Andorra"[TIAB] OR "Angola"[TIAB] OR "Anguilla"[TIAB] OR "Antarctica"[TIAB] OR "Antigua"[TIAB] OR "Argentina"[TIAB] OR "Armenia"[TIAB] OR "Aruba"[TIAB] OR "Australia"[TIAB] OR "Austria"[TIAB] OR "Azerbaijan"[TIAB] OR "Bahamas"[TIAB] OR "Bahrain"[TIAB] OR "Bangladesh"[TIAB] OR "Barbados"[TIAB] OR "Barbuda"[TIAB] OR "Belarus"[TIAB] OR "Belgium"[TIAB] OR "Belize"[TIAB] OR "Benin"[TIAB] OR "Bermuda"[TIAB] OR "Bhutan"[TIAB] OR "Bolivia"[TIAB] OR "Bonaire"[TIAB] OR "Bosnia"[TIAB] OR "Botswana"[TIAB] OR "Bouvet Island"[TIAB] OR "Brazil"[TIAB] OR "British Indian Ocean Territory"[TIAB] OR "Brunei"[TIAB] OR "Bulgaria"[TIAB] OR "Burkina Faso"[TIAB] OR "Burundi"[TIAB] OR "Cabo Verde"[TIAB] OR "Cambodia"[TIAB] OR "Cameroon"[TIAB] OR "Canada"[TIAB] OR "Cayman Islands"[TIAB] OR "Central African Republic"[TIAB] OR "Chad"[TIAB] OR "Chile"[TIAB] OR "China"[TIAB] OR "Christmas Island"[TIAB] OR "Cocos Islands"[TIAB] OR "Colombia"[TIAB] OR "Comoros"[TIAB] OR "Congo"[TIAB] OR "Congo"[TIAB] OR "Cook Islands"[TIAB] OR "Costa Rica"[TIAB] OR "Côte d'Ivoire"[TIAB] OR "Croatia"[TIAB] OR "Cuba"[TIAB] OR "Curaçao"[TIAB] OR "Cyprus"[TIAB] OR "Czechia"[TIAB] OR "Denmark"[TIAB] OR "Djibouti"[TIAB] OR "Dominica"[TIAB] OR "Dominican Republic"[TIAB] OR "Ecuador"[TIAB] OR "Egypt"[TIAB] OR "El Salvador"[TIAB] OR "Equatorial Guinea"[TIAB] OR "Eritrea"[TIAB] OR "Estonia"[TIAB] OR "Eswatini"[TIAB] OR "Ethiopia"[TIAB] OR "Falkland Islands"[TIAB] OR "Faroe Islands"[TIAB] OR "Fiji"[TIAB] OR "Finland"[TIAB] OR "France"[TIAB] OR "French Guiana"[TIAB] OR "French Polynesia"[TIAB] OR "French Southern

Territories"[TIAB] OR "Futuna"[TIAB] OR "Gabon"[TIAB] OR "Gambia"[TIAB] OR "Georgia"[TIAB] OR "Germany"[TIAB] OR "Ghana"[TIAB] OR "Gibraltar"[TIAB] OR "Greece"[TIAB] OR "Greenland"[TIAB] OR "Grenada"[TIAB] OR "Grenadines"[TIAB] OR "Guadeloupe"[TIAB] OR "Guam"[TIAB] OR "Guatemala"[TIAB] OR "Guernsey"[TIAB] OR "Guinea"[TIAB] OR "Guinea-Bissau"[TIAB] OR "Guyana"[TIAB] OR "Haiti"[TIAB] OR "Heard Island"[TIAB] OR "Herzegovina"[TIAB] OR "Holy See"[TIAB] OR "Honduras"[TIAB] OR "Hong Kong"[TIAB] OR "Hungary"[TIAB] OR "Iceland"[TIAB] OR "India"[TIAB] OR "Indonesia"[TIAB] OR "Iran"[TIAB] OR "Iraq"[TIAB] OR "Ireland"[TIAB] OR "Isle of Man"[TIAB] OR "Israel"[TIAB] OR "Italy"[TIAB] OR "Jamaica"[TIAB] OR "Jan Mayen Islands"[TIAB] OR "Japan"[TIAB] OR "Jersey"[TIAB] OR "Jordan"[TIAB] OR "Kazakhstan"[TIAB] OR "Keeling Islands"[TIAB] OR "Kenya"[TIAB] OR "Kiribati"[TIAB] OR "Korea"[TIAB] OR "Korea"[TIAB] OR "Kuwait"[TIAB] OR "Kyrgyzstan"[TIAB] OR "Lao People's Democratic Republic"[TIAB] OR "Laos"[TIAB] OR "Latvia"[TIAB] OR "Lebanon"[TIAB] OR "Lesotho"[TIAB] OR "Liberia"[TIAB] OR "Libya"[TIAB] OR "Liechtenstein"[TIAB] OR "Lithuania"[TIAB] OR "Luxembourg"[TIAB] OR "Macao"[TIAB] OR "Madagascar"[TIAB] OR "Malawi"[TIAB] OR "Malaysia"[TIAB] OR "Maldives"[TIAB] OR "Mali"[TIAB] OR "Malta"[TIAB] OR "Malvinas"[TIAB] OR "Marshall Islands"[TIAB] OR "Martinique"[TIAB] OR "Mauritania"[TIAB] OR "Mauritius"[TIAB] OR "Mayotte"[TIAB] OR "McDonald Islands"[TIAB] OR "Mexico"[TIAB] OR "Micronesia"[TIAB] OR "Moldova"[TIAB] OR "Monaco"[TIAB] OR "Mongolia"[TIAB] OR "Montenegro"[TIAB] OR "Montserrat"[TIAB] OR "Morocco"[TIAB] OR "Mozambique"[TIAB] OR "Myanmar"[TIAB] OR "Namibia"[TIAB] OR "Nauru"[TIAB] OR "Nepal"[TIAB] OR "Netherlands"[TIAB] OR "Nevis"[TIAB] OR "New Caledonia"[TIAB] OR "New Zealand"[TIAB] OR "Nicaragua"[TIAB] OR "Niger"[TIAB] OR "Nigeria"[TIAB] OR "Niue"[TIAB] OR "Norfolk Island"[TIAB] OR "North Macedonia"[TIAB] OR "Northern Mariana Islands"[TIAB] OR "Norway"[TIAB] OR "Oman"[TIAB] OR "Pakistan"[TIAB] OR "Palau"[TIAB] OR "Panama"[TIAB] OR "Papua New Guinea"[TIAB] OR "Paraguay"[TIAB] OR "Peru"[TIAB] OR "Philippines"[TIAB] OR "Pitcairn"[TIAB] OR "Poland"[TIAB] OR "Portugal"[TIAB] OR "Principe"[TIAB] OR "Puerto Rico"[TIAB] OR "Qatar"[TIAB] OR "Réunion"[TIAB] OR "Romania"[TIAB] OR "Russian Federation"[TIAB] OR "Rwanda"[TIAB] OR "Saba"[TIAB] OR "Saint Barthélemy"[TIAB] OR "Saint Helena"[TIAB] OR "Saint Kitts"[TIAB] OR "Saint Lucia"[TIAB] OR "Saint Martin"[TIAB] OR "Saint Pierre and Miquelon"[TIAB] OR "Saint Vincent"[TIAB] OR "Samoa"[TIAB] OR "San Marino"[TIAB] OR "Sao Tome"[TIAB] OR "Sark"[TIAB] OR "Saudi Arabia"[TIAB] OR "Senegal"[TIAB] OR "Serbia"[TIAB] OR "Seychelles"[TIAB] OR "Sierra Leone"[TIAB] OR "Singapore"[TIAB] OR "Sint Eustatius"[TIAB] OR "Sint Maarten"[TIAB] OR "Slovakia"[TIAB] OR "Slovenia"[TIAB] OR "Solomon Islands"[TIAB] OR "Somalia"[TIAB] OR "South Africa"[TIAB] OR "South Georgia"[TIAB] OR "South Sandwich Islands"[TIAB] OR "South Sudan"[TIAB] OR "Spain"[TIAB] OR "Sri Lanka"[TIAB] OR "State of Palestine"[TIAB] OR "Sudan"[TIAB] OR "Suriname"[TIAB] OR "Svalbard"[TIAB] OR "Sweden"[TIAB] OR "Switzerland"[TIAB] OR "Syria"[TIAB] OR "Syrian Arab Republic"[TIAB] OR "Tajikistan"[TIAB] OR "Thailand"[TIAB] OR "Timor-Leste"[TIAB] OR "Tobago"[TIAB] OR "Togo"[TIAB] OR "Tokelau"[TIAB] OR "Tonga"[TIAB] OR "Trinidad"[TIAB] OR "Tunisia"[TIAB] OR "Turkey"[TIAB] OR "Turkmenistan"[TIAB] OR "Turks and Caicos"[TIAB] OR "Tuvalu"[TIAB] OR "Uganda"[TIAB] OR "UK"[TIAB] OR "Ukraine"[TIAB] OR "United Arab Emirates"[TIAB] OR "United Kingdom"[TIAB] OR "United Republic of Tanzania"[TIAB] OR "United States Minor Outlying Islands"[TIAB] OR "United States of America"[TIAB] OR

"Uruguay"[TIAB] OR "USA"[TIAB] OR "Uzbekistan"[TIAB] OR "Vanuatu"[TIAB] OR "Venezuela"[TIAB] OR "Viet Nam"[TIAB] OR "Vietnam"[TIAB] OR "Virgin Islands"[TIAB] OR "Virgin Islands"[TIAB] OR "Wallis"[TIAB] OR "Western Sahara"[TIAB] OR "Yemen"[TIAB] OR "Zambia"[TIAB] OR "Zimbabwe"[TIAB] OR "Alabama"[TIAB] OR "Alaska"[TIAB] OR "Arizona"[TIAB] OR "Arkansas"[TIAB] OR "California"[TIAB] OR "Colorado"[TIAB] OR "Connecticut"[TIAB] OR "Delaware"[TIAB] OR "Florida"[TIAB] OR "Georgia"[TIAB] OR "Hawaii"[TIAB] OR "Idaho"[TIAB] OR "Illinois"[TIAB] OR "Indiana"[TIAB] OR "Iowa"[TIAB] OR "Kansas"[TIAB] OR "Kentucky"[TIAB] OR "Louisiana"[TIAB] OR "Maine"[TIAB] OR "Maryland"[TIAB] OR "Massachusetts"[TIAB] OR "Michigan"[TIAB] OR "Minnesota"[TIAB] OR "Mississippi"[TIAB] OR "Missouri"[TIAB] OR "Montana"[TIAB] OR "Nebraska"[TIAB] OR "Nevada"[TIAB] OR "New Hampshire"[TIAB] OR "New Jersey"[TIAB] OR "New Mexico"[TIAB] OR "New York"[TIAB] OR "North Carolina"[TIAB] OR "North Dakota"[TIAB] OR "Ohio"[TIAB] OR "Oklahoma"[TIAB] OR "Oregon"[TIAB] OR "Pennsylvania"[TIAB] OR "Rhode Island"[TIAB] OR "South Carolina"[TIAB] OR "South Dakota"[TIAB] OR "Tennessee"[TIAB] OR "Texas"[TIAB] OR "Utah"[TIAB] OR "Vermont"[TIAB] OR "Virginia"[TIAB] OR "Washington"[TIAB] OR "West Virginia"[TIAB] OR "Wisconsin"[TIAB] OR "Wyoming"[TIAB] OR "Ontario"[TIAB] OR "Quebec"[TIAB] OR "Nova Scotia"[TIAB] OR "New Brunswick"[TIAB] OR "Manitoba"[TIAB] OR "British Columbia"[TIAB] OR "Prince Edward Island"[TIAB] OR "Saskatchewan"[TIAB] OR "Alberta"[TIAB] OR "Newfoundland"[TIAB] OR "Labrador"[TIAB])

Appendix 3: Article review flow diagram



Review version with references removed; NOT FOR DISTRIBUTION

Policy evaluation in COVID-19: A guide to common design issues

Noah A Haber, Emma Clarke-Deelder, Joshua A Salomon, Avi Feller, Elizabeth A Stuart

Noah A Haber, ScD*
noahhaber@stanford.edu
Meta Research Innovation Center at Stanford University
Stanford University
1265 Welch Rd
Palo Alto, CA 94305
(650) 497-0811

Emma Clarke-Deelder, MPhil
Department of Global Health & Population
Harvard T. H. Chan School of Public Health
665 Huntington Avenue
Building 1, room 1104
Boston, Massachusetts 02115

Joshua A Salomon, PhD
Department of Medicine
Center for Health Policy and Center for Primary Care and Outcomes Research
Stanford University School of Medicine
Encina Commons, Room 118
615 Crothers Way
Stanford, CA 94305-6019

Avi Feller, PhD
Goldman School of Public Policy
University of California, Berkeley
2607 Hearst Avenue
Room 309
Berkeley, CA 94720

Elizabeth A Stuart, PhD
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health
624 N. Broadway
Hampton House 839
Baltimore, MD 21205

* corresponding author

Review version with references removed; NOT FOR DISTRIBUTION

Abstract

Policy responses to COVID-19, particularly those related to non-pharmaceutical interventions, are unprecedented in scale and scope. Researchers and policymakers are striving to understand the impact of these policies on a variety of outcomes. Policy impact evaluations always require a complex combination of circumstance, study design, data, statistics, and analysis. Beyond the issues that are faced for any policy, evaluation of COVID-19 policies is complicated by additional challenges related to infectious disease dynamics and lags, lack of direct observation of key outcomes, and a multiplicity of interventions occurring on an accelerated time scale.

In this paper, we (1) introduce the basic suite of policy impact evaluation designs for observational data, including cross-sectional analyses, pre/post, interrupted time-series, and difference-in-differences analysis, (2) demonstrate key ways in which the requirements and assumptions underlying these designs are often violated in the context of COVID-19, and (3) provide decision-makers and reviewers a conceptual and graphical guide to identifying these key violations. The overall goal of this paper is to help policy-makers, journal editors, journalists, researchers, and other research consumers understand and weigh the strengths and limitations of evidence that is essential to decision-making.

Introduction

The response to the global COVID-19 pandemic has demanded urgent decision making in the face of substantial uncertainties. Policies to arrest transmission, including stay-at-home orders and other non-pharmaceutical interventions (NPIs), have wide-reaching consequences that touch many aspects of well being. Decision-making in the public interest requires evaluating and weighing the evidence on both intended and unintended consequences in order to best predict outcomes. The wide range of policy interventions implemented by different jurisdictions may yield opportunities for learning from what has already happened to inform future policymaking, and we have observed a proliferation of studies aimed at such policy evaluations. However, policy evaluation requires a complex combination of circumstance, data, study design, analysis, and interpretation in order to be informative.

Policy impact evaluation aims to answer questions about the extent to which the realized outcomes given a particular policy would have been different in the absence of that policy. Estimating the causal impact of the policy with observational data is challenging because what would have happened in the absence of the policy change (the “counterfactual”) is, by definition, unobserved. Randomized controlled trials (RCTs) of policies related to COVID-19 interventions may not always be practical or ethical. In this context, a large and growing number of studies have attempted to evaluate the impact of COVID-19 policies using observational data. There

are many potential pitfalls in the use of observational data for evaluation generally, and some additional methodological design challenges relating to COVID-19 policies in particular.

This paper provides a graphical guide to policy impact evaluations for COVID-19, targeted to decision-makers, researchers and evidence curators. Our aim is to provide a coherent framework for conceptualizing and identifying common pitfalls in COVID-19 policy evaluation. Importantly, this should not be taken either as a comprehensive guide to policy evaluation more broadly or as guidance on performing analysis, which may be found elsewhere. Rather, we review relevant study designs for policy evaluations — including pre/post, interrupted time series, and difference-in-difference approaches — and provide guidance and tools for identifying key issues with each type of study as they relate to NPIs and other COVID-19 policy interventions. Improving our ability to identify key pitfalls will enhance our ability to identify and produce valid and useful evidence for informing policymaking.

Common policy evaluation designs and their pitfalls in COVID-19

Identifying the type of design

Review version with references removed; NOT FOR DISTRIBUTION

Table 1: Summary definitions of policy impact evaluation designs commonly used for COVID-19

Design	**Units (e.g., regions of comparison)**		**Time points measured per unit**		**Assumed counterfactual.** **"If not for the intervention, ___"**
	With intervention	**Without intervention**	**Before intervention**	**After intervention**	
Cross-sectional	At least one	At least one	N/A	One time point	Outcome in intervention units would have been the same as the outcome in the non-intervention units.
Pre/post Figure 1A	At least one	None	At least one (typically one)	At least one (typically one)	Outcome would have stayed the same from the pre period to the post period.
Interrupted time-series (ITS) Figure 1B	At least one	None	More than one	At least one (typically several)	Outcome slope and level* would have continued along the same modelled trajectory from the pre-period to the post period.
Difference-in-differences (DiD) Figure 1C	At least one	At least one†	At least one (typically one)	At least one (typically one)	Outcome in intervention units would have changed as much as (or in parallel with) the outcome in the non-intervention units.
Comparative interrupted time series (CITS) Figure 1D	At least one	At least one†	More than one (typically several)	At least one (typically several)	Outcome slope and level* would have changed as much as non-intervention group's slope and level* changed.

* Assessing both slope and level only applicable if there are multiple data points during the post period
† Units without the intervention may be the pre-period of a different unit that eventually receives the intervention.

Identifying the underlying design in a given analysis often requires using a combination of the methods as reported and evaluating the data structure that is used for the main analysis, as shown in Table 1. COVID-19-related policy evaluation analyses typically fall under these categories. In most cases, the design can be categorized using a combination of whether there are also units that did not receive the treatment (columns 2-3) and whether there are time points both before and after intervention for those units (columns 4-5). The final column describes the implied counterfactual, discussed further in subsequent sections. Cross sectional designs typically compare units with vs without the treatment at single time points. Pre/post studies typically compare within units who received the intervention at two points: before and after a policy. Interrupted time-series analyses compare outcomes within units within units who received the intervention at greater than two time points before the intervention vs with at least one (typically multiple) after the intervention. Difference-in-differences analysis compares the outcome change in units which received the intervention with those that did not (or have not yet), with at least one point before and one after the intervention. In cases with multiple periods, that may involve a comparison with the pre-policy period of one region with the post-period of a different region, even though all regions eventually receive the intervention.

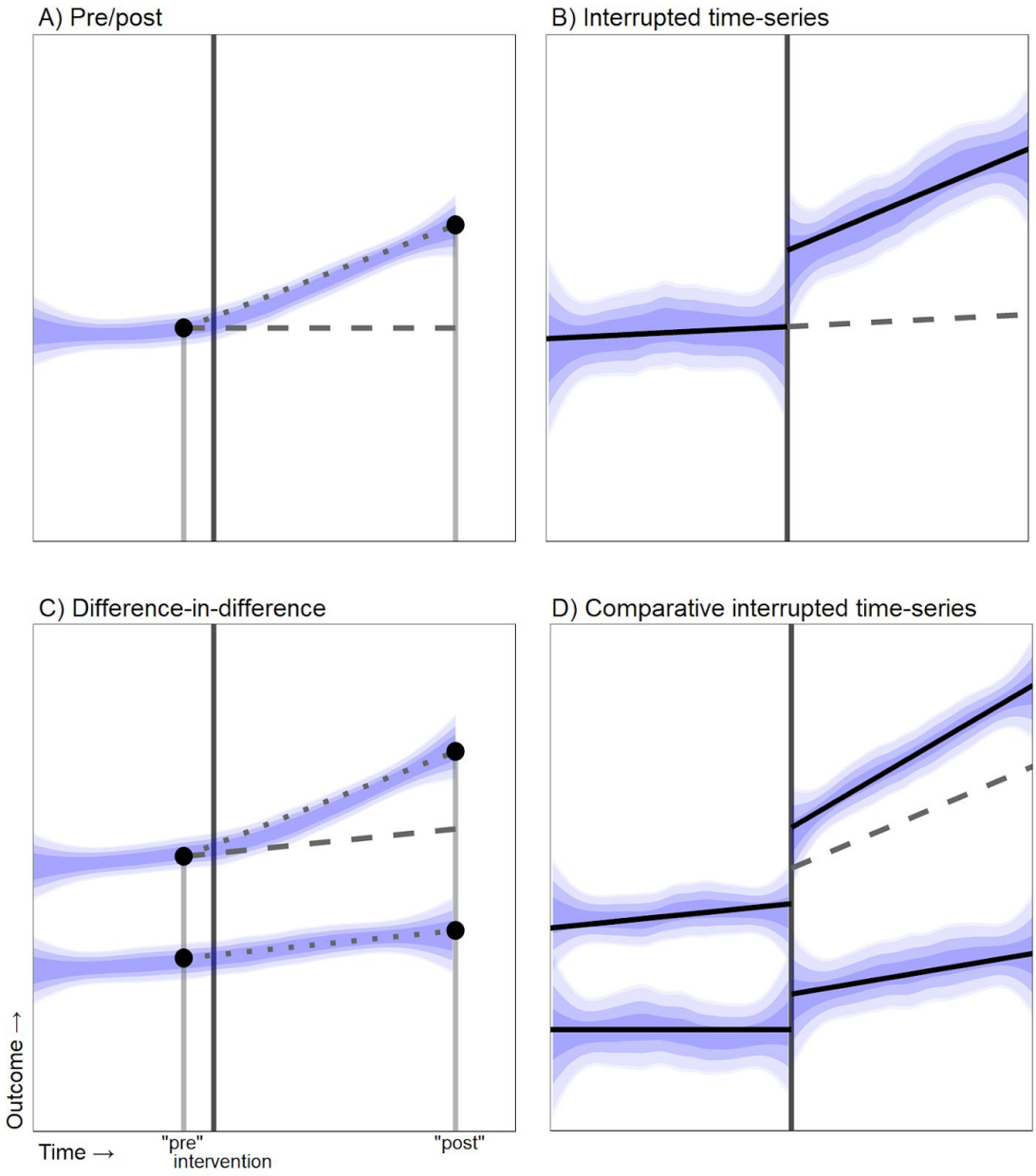
Review version with references removed; NOT FOR DISTRIBUTION

Methods descriptions may not always provide a precise or reliable guide to which of the design approaches has been used. Some studies do not explicitly name these designs (or may classify them differently); and these are only a small fraction of designs and frameworks that are possible to use for policy evaluation. Studies may have data at multiple time points but are effectively cross-sectional. DiD, ITS, and CITS designs based on repeated cross-sectional data are sometimes described as “cross-sectional” instead of longitudinal. The term “event study” is often used to refer to studies with a single unit and one change over time resembling ITS, but may refer to other designs. Although ITS is often used to describe changes in one unit, it may also refer to settings in which many treated units adopt an intervention over time. Studies will also frequently employ multiple designs, while others use more complex methods of generating counterfactuals. Definitions of these terms vary widely, and the definitions above should be considered as guidance only.

Policy impact evaluation design foundations for COVID-19

The simplest design is the cross-sectional analysis, which compares COVID-19 outcomes between units of observation (e.g., cities) at a single calendar time or time since an event, typically post-intervention. These studies are unlikely to be appropriate for COVID-19-related policy evaluations, but provide a useful starting point for reasoning about different designs. Just as with comparisons of non-randomized medical treatments, the localities that adopt a particular policy likely differ substantially from those that don't on both observed and unobserved characteristics on a number of dimensions, including epidemic status and timing.

Figure 1: Longitudinal designs overview



This chart shows four canonical longitudinal designs. In all cases: the blue shading represents the underlying data trends, the solid vertical grey line represents the time of intervention, the grey dashed lines represent the assumed counterfactual in the absence of the intervention, as discussed in the text. The impact estimate is obtained by comparing the outcomes observed for the treated unit in the post period (the solid line) with the implied counterfactual line (the dashed line). In the case of the pre/post and

Review version with references removed; NOT FOR DISTRIBUTION

difference-in-differences panels the large black dots represent the time of measurement, connected by the grey dotted lines.

Given the challenges in a simple cross-sectional comparison, which compare post-intervention outcomes, it is important to consider longitudinal designs, which instead look at differences or trends across time, as summarized in Figure 1. These can be distinguished by the data used and the construction of the counterfactual. Pre/post, for example, has only one unit, measured at two time points. Two common strategies expand on the logic and data requirements of the pre/post design. Interrupted time series designs (Figure 1B) incorporate multiple time points before the intervention, and usually multiple time points after the intervention, to enable a more complete view on changes in levels and trends that are temporally related to the intervention. Difference-in-difference designs (Figure 1C) add a set of comparison points from a group or location that did not have the intervention. Another related design (comparative interrupted time-series, Figure 1D, discussed only briefly here), uses both aspects — a change over time and a comparison group — to compare the observed change in slopes for the intervention group with the change in slope for the comparison group.

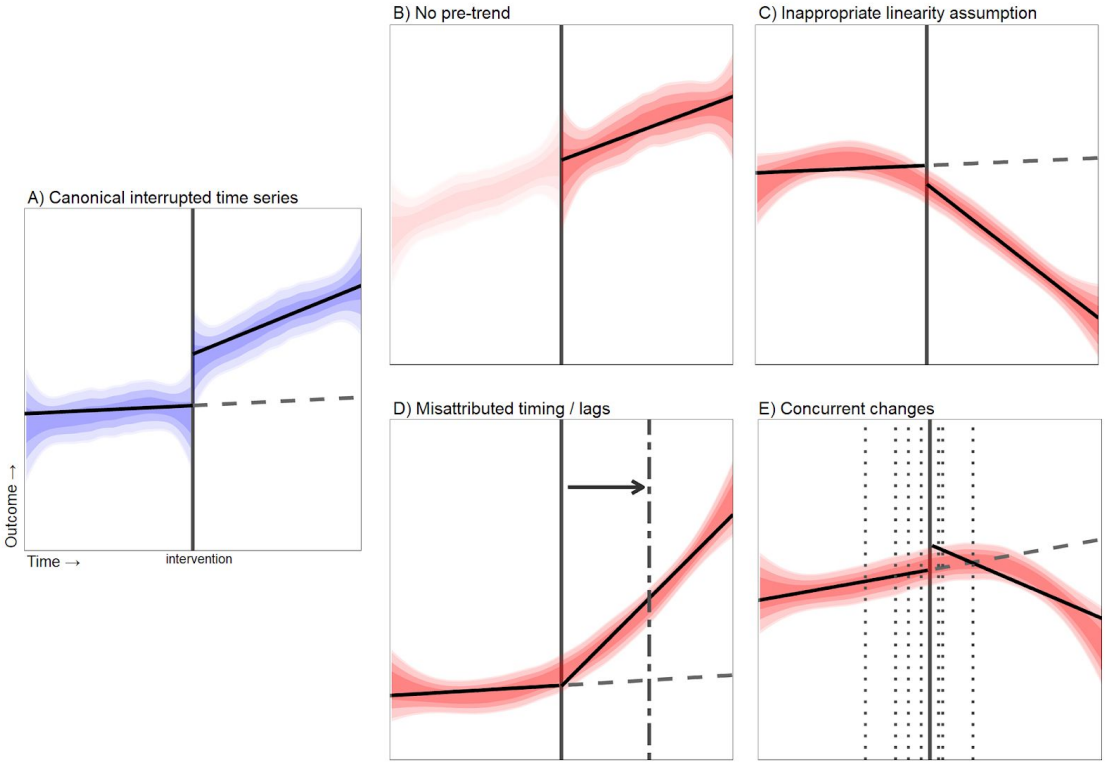
Pre/post studies

The simplest longitudinal design is a pre/post analysis, where some outcome is observed before policy implementation, and again after, in a single group (Figure 1A). Pre/post studies are analogous to a single arm trial with no control and only a single follow-up observation after treatment.. This effectively imposes the assumption that the counterfactual trend is completely flat (i.e., that the outcome in the post-period in the absence of the policy change is the same as the value of the outcome before the policy change) without accounting for pre-existing underlying trends, and attributing all outcome changes completely to the intervention of interest. Just as the outcomes for an individual patient might be expected to change before and after treatment, for reasons unrelated to the treatment, outcomes related to policy interventions will change for reasons not caused by the policy. Infection rates, for example, would not be expected to remain stationary except in very specific circumstances, but a pre/post measurement would assume that any changes in infection rates are attributable to the policy.

Interrupted time-series

Figure 2: Interrupted time-series graphical guidance for identifying common pitfalls

Review version with references removed; NOT FOR DISTRIBUTION



This chart shows one canonical design for ITS (blue, Panel A) and four panels demonstrating common issues with ITS analysis (red, panels B-E) discussed in the text. In all cases: the lag/red shading represents the underlying data trends, the vertical grey line represents the time of intervention, the grey dashed lines represent the assumed counterfactual in the absence of the intervention. In panel D) the dash-dot line represents the time at which the policy is expected to impact the outcome. In panel E), the vertical dotted lines represent concurrent events and changes.

Interrupted time-series (ITS) is a strategy that uses a projection of the pre-policy outcome trend as a counterfactual for how the outcome would have changed if the policy had not been introduced. In other words, in the absence of the policy change, ITS assumes the outcome would have continued on its pre-policy trend during the study period. ITS can be a useful tool in policy evaluation because it allows researchers to account for underlying trends in the outcome and, by comparing the treated unit (or location) to itself; it can therefore eliminate some of the confounding concerns that arise in cross-sectional or pre-post studies.

However, the validity of ITS depends critically on how well counterfactual trends in the outcome are modelled, and whether the policy of interest is the only relevant change during the study period. In the canonical setting (Figure 2A), the pre-policy trend is stable and can be feasibly modelled with the available data; the researcher appropriately models the timing of the change in the slope and/or level of the outcome; the researcher has sufficient information to conclude

that there were no other changes during the study period that would be expected to influence the outcome. These elements are largely not satisfied in studies of COVID-related policy, as described below.

ITS relies critically on modelled trends of the outcome over time. Key components of ITS analyses include both visual and statistical examination of trends, preferentially alongside a theoretical justification of the model used. At a minimum, analyses should provide graphical representation of the data and model over time to examine whether pre-trend outcomes are stable, all trends are well-fit to the data, “interrupted” at the appropriate time point, and sensibly modelled (Figure 2B). In the case where an ITS includes a large number of units (e.g. states), it can be difficult to display this information graphically.

One common pitfall in ITS is adoption of inappropriate assumptions on the outcome trend (Figure 2C). The estimate of policy impact will be biased if a linear trend is assumed but the outcome and response to interventions instead follow nonlinear trends (either before or after the policy). In some cases, transformation of the outcome, for example using a log scale, may improve the suitability of a linear model. Imposing linearity inappropriately is a serious risk in the context of COVID-19, as trends in infectious disease dynamics are inherently non-linear. For intuition, terms such as “exponential growth,” “flattening,” and “s-curves” all refer to non-linear infectious disease trends. Depending on the particular situation, non-linearity or other modelled trends can have complicated and counterintuitive impact on policy impact. Apparent linearity may also be temporary and an artifact of testing, which may give a misleading impression that linear models for infectious disease trends are appropriate indefinitely. While some use linear progression in order to avoid more complex infectious disease models, in fact, linear projections impose strict and often unrealistic models, generally resulting in an inappropriate counterfactual.

Researchers can easily misattribute the timing of the policy impact, resulting in spurious inference and bias (Figure 2D). Some public health policies can be expected to translate into immediate results (e.g., smoking bans and acute coronary events). In contrast, nearly every outcome of interest in COVID-19 exhibits complex and difficult to infer time lags typically in the realm of many weeks. The time between policy implementation and expected effect in the data can be large and highly variable. For example, in order to see the impact of a mask order, first the mask order takes effect, then people change their behaviors over time to comply with the order (or sometimes the reverse in the case of anticipation effects), mask use behavior produces changes in infections, then infections later result in symptoms, symptoms induce people to seek testing, the tests must then be processed in labs, and then finally the results get reported in data monitoring efforts. Selection of lead/lag time should be justifiable *a priori* or external data. Selecting a lag based on the data risks issues comparable to p-hacking.

Finally, and perhaps most concerningly in the context of COVID-19, ITS fails when the policy of interest coincides in time with other changes that affect the outcome (Figure 2E). For example, if both mask and bar closure orders are rolled out together as a package, ITS cannot isolate the impact of bar closures specifically. These changes do not need to have taken place exactly

Review version with references removed; NOT FOR DISTRIBUTION

concurrently with the policy implementation date of interest; they merely need to have some effect within the time period of measurement to result in potentially serious bias in effect estimates if unaddressed. ITS will also likely be biased if, during the study period, there is a change in the way the outcome data is collected or measured. This might occur if the introduction of a COVID-19 control policy is combined with an effort to collect better data on infection or mortality cases. Analogously, if an RCT involves randomizing people to a group receiving both A and B vs. control, we typically can't disentangle the effects of A from the effects of B, unless we also have separate A- and B-only arms. Ultimately, if multiple things are changing at the same time, ITS may not be an appropriate design for policy evaluation.

COVID-19 policies rarely arrive alone; they are typically created alongside other policies, unofficial action, and large scale behavior changes which themselves impact COVID-19-related outcomes. In some cases, anticipation of a policy may induce behavior change before the actual policy takes effect. The policies themselves may have been chosen due to the expectation of change in disease outcomes, which introduces additional biases related to “reverse” causality.

Table 2: Checklist for identifying common pitfalls for ITS to evaluate COVID-19 policy

Key design questions. If any answer is “no,” this analysis is unlikely to be appropriate or useful for estimating the impact of the intervention of interest.	**Details and suggestions for identifying issues:**
Does the analysis provide graphical representation of the outcome over time?	-Check for a chart that shows the outcome over time, with the dates of interest. Outcomes may be aggregated for clarity (e.g. means and CIs at discrete time points).
Is there sufficient pre-intervention data to characterize pre-trends in the data?	-Check the chart(s) to see if there are several time points over a reasonable period of time over which to establish stability and curvature in the pre-trends.
Is the pre-trend stable?	-Check if there are sufficient data to reasonably determine a stable functional form for the pre-trends, and that they follow a modelable functional form.
Is the functional form of the counterfactual (e.g. linear) well-justified and appropriate?	-Check whether the authors explain and justify their choice of functional form. -Check if there is any curvature in the pre-trend. -Consider the nature of the outcome. Is it sensible to measure the trend of this outcome on the scale and form used? Note: infectious disease dynamics are rarely linear. -Consider that while pre-trend fit is a necessary condition for an appropriate linear counterfactual model, it is not sufficient. Check if the authors provide justification for the functional form to continue to be of the same functional form (e.g. linear).
Is the date or time threshold set to the appropriate date or time (e.g. is there lag between the intervention and outcome)?	-Check whether the authors justify the use of the date threshold relative to the date of the intervention. -Trace the process between the intervention being put in place to when observable effects in the outcome might appear over time. -Consider whether there are anticipation effects (e.g. do people change behaviors before the date when the intervention begins?) -Consider whether there are lag effects. (e.g. does it take time for behaviors to change, behavior change to impact infections, infections to impact testing, and data to be collected, etc?)

Review version with references removed; NOT FOR DISTRIBUTION

	-Check if authors appropriately and directly account for these time effects.
Is this policy the only thing to happen which could have impacted the outcome during the measurement period, differently for policy and non-policy regions??	-Consider other policies or interventions which could impact the outcome during this time. -Consider social behaviors changed which could meaningfully impact the outcome during this time. -Consider economic conditions changed which could meaningfully impact the outcome during this time. -Note that the actual concurrent changes do not need to happen during the period of measurement, just their effects.

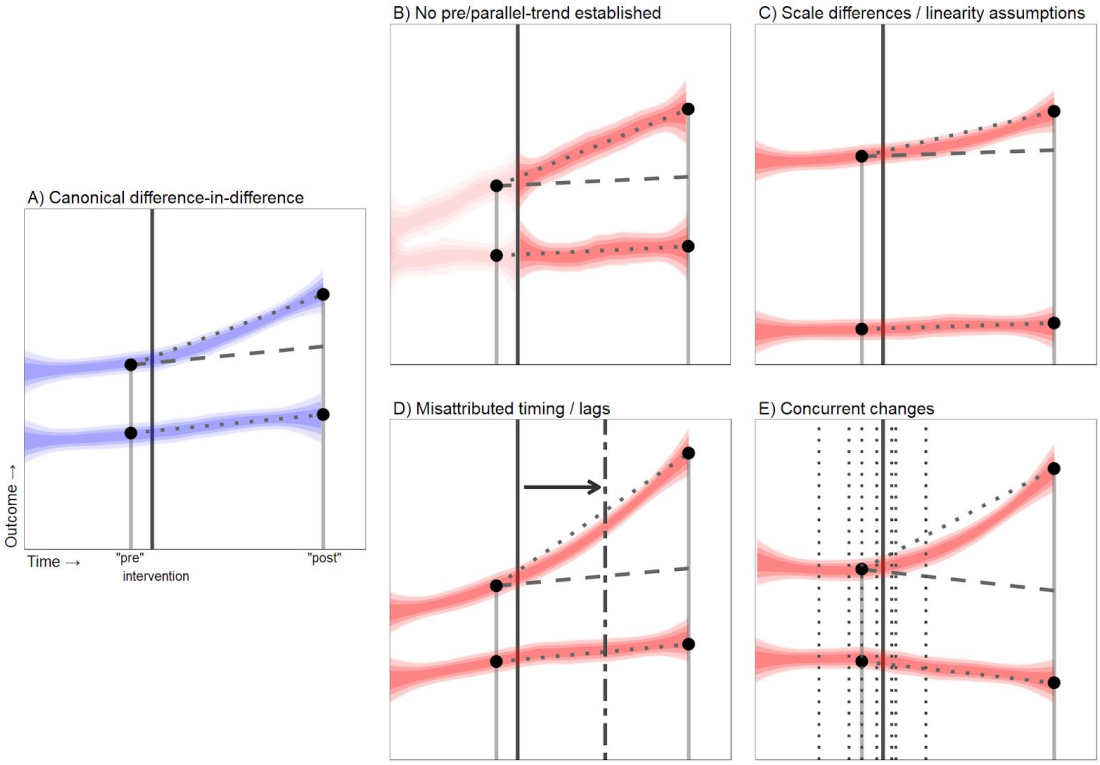
These issues are summarized as a checklist of questions to identify common pitfalls in Table 2.

Difference-in-differences

The difference-in-difference (DiD) approach uses concurrent non-intervention groups as a counterfactual. Typically, this consists of one set of units (e.g., regions) that had the intervention and one set that did not, with each measured before and after the intervention took place. DiD is more directly analogous to a non-randomized medical study with at least one treatment and control group but limited observation before and after treatment. In contrast to ITS, which compares a unit with itself over time, DiD compares differences between treatment arms or units at two observation points. In many analyses, a DiD approach is implied by comparing regions over time, without formally naming or modelling it. Other DiD approaches use interventions implemented at multiple time points.

Figure 3: Difference-in-differences graphical guidance for identifying common pitfalls

Review version with references removed; NOT FOR DISTRIBUTION



This chart shows one canonical design for DiD (blue, Panel A) and four panels demonstrating common issues with DiD analysis (red, panels B-E). In all cases: the blue/red shading represents the underlying data trends, the vertical grey line represents the time of intervention, the grey dashed lines represent the assumed counterfactual in the absence of the intervention. In panel D) the dash-dot line represents the time at which the policy is expected to impact the outcome. In panel E), the vertical dotted lines represent concurrent events and changes.

One key component of the standard DiD approach is the parallel counterfactual trends assumption: that the intervention and comparison groups would have had parallel trends over time in the absence of the intervention. In some cases, the parallel trends assumption may be referenced or examined implicitly but not named.

Ideally, pre-intervention trends would be shown to be clearly identifiable, stable, of a similar level, and parallel between groups. With only one observation before and only one after the intervention, assessment of the plausibility of the parallel counterfactual trends assumption is not possible. Absent this confirmation the evaluation runs the risk of biased estimation due to differential pre-trends (Figure 3B). Pre-trends approaching the ceiling or floor may also not be informative about stable and parallel pre-trends. Empirical assessment of whether pre-intervention trends were parallel and stable between groups is possible when multiple observations are available at multiple time points before the intervention, noting that this can

begin to resemble a CITS design. In this scenario, pre-trend data should be visually and statistically established and documented. While parallel trends before intervention (which we can observe and may be testable) do not guarantee parallel *counterfactual* trends in the post-intervention period (which we cannot observe and are generally untestable), examining pre-intervention parallel trends is a minimal requirement for DiD reliability.

It is also important to consider the scale and level on which the outcome is measured (Figure 3C). As with ITS, if the outcomes in the treatment and comparison groups are moving in parallel on a logged scale, they will not be moving in parallel on a natural scale. Level differences by themselves may be a problem for COVID-19 outcomes, as infectious disease transmission dynamics dictate that infection risks are related to the prevalence of infected people in a population, i.e. the rate of change is linked intrinsically to the level. A population with an extremely low prevalence will tend to have an inherently slower rise in infection rates than an otherwise identical population with merely a low prevalence. Just as importantly, large level differences in the outcome between intervention and comparison groups is often indicative of other important differences between comparators, which may result in other assumptions being violated.

While DiD is in some ways more robust to very specific kinds of timing effects (Figure 3D) and concurrent changes (Figure 3E), it also introduces additional risks. DiD effectively doubles the opportunity for concurrent changes to spuriously impact results, since they can occur in the treatment or comparison groups. As above, this can become even more problematic for DiD in the typical case where intervention groups enact more or very contextually different policies than non-intervention groups. Even cases where concurrent changes happen equally in both treatment and comparison groups can lead to overwhelming bias, particularly when approaching the maximum or minimum levels of the osutcome. If either the treatment or control group is approaching the floor (e.g. 0% prevalence) or ceiling for an outcome of interest due to other policies concurrent in both places (e.g. national lockdowns, but region-level differences in mask policy), this can lead to bias when comparing changes between the two groups.

Table 3: Checklist for identifying common pitfalls for DiD to evaluate COVID-19 policy

Key design questions. If any answer is “no,” this analysis is unlikely to be appropriate or useful for estimating the impact of the intervention of interest	**Details and suggestions for inspection:**
Does the analysis provide graphical representation of the outcome over time?	-Check for a graph that shows the outcome over time for all groups, with the dates of interest. Outcomes may be aggregated for clarity (e.g. mean and CI at discrete time points).
Is there sufficient pre-intervention data to observe both pre and post trends in the data?	-Check the chart(s) to see if there are several time points over a reasonable period of time over which to establish stability and curvature in the pre- and post- trends.
Are the pre-trends stable?	-Check if there are sufficient graphical data to reasonably determine a stable functional form for the pre-trends, and that they follow a modelable functional form.

Review version with references removed; NOT FOR DISTRIBUTION

Are the pre-trends parallel?	-Observe if the trends in the intervention and comparison groups appear to move together at the same rate at the same time.
Are the pre-trends at a similar level?	-Check if the trends in the intervention and comparison groups are at similar levels. -Note that non-level trends exacerbates other problems with the analysis, including linearity assumptions
Are intervention and non-groups broadly comparable?	-Consider areas where comparison groups may be dissimilar for comparison beyond just the level of the outcome.
Is the date or time threshold set to the appropriate date or time (e.g. is there lag between the intervention and outcome)?	-Check whether the authors justify the use of the date threshold relative to the date of the intervention. -Trace the process between the intervention being put in place to when observable effects in the outcome might appear over time. -Consider whether there are anticipation effects (e.g. do people change behaviors before the date when the intervention begins?) -Consider whether there are lag effects. (e.g. does it take time for behaviors to change, behavior change to impact infections, infections to impact testing, and data to be collected, etc?) -Check if authors appropriately and directly account for these time effects.
Is this policy the only uncontrolled or unadjusted-for way in which the outcome could have changed during the measurement period, differently for policy and non-policy regions?	-Consider any uncontrolled factor which could have influenced the outcome differently in policy and non-policy regions. -This may include (but is not limited to) -Other policies -Social behaviors -Economic conditions -Are these factors justified as having negligible impact? -If justified, is the argument that these have negligible impact convincing? -Note that the actual concurrent changes do not need to happen during the period of measurement, just their effects.

Similarly to the ITS section, these issues are summarized as a checklist of questions to identify common pitfalls in Table 3.

Discussion

In recent months, there has been a proliferation of research evaluating policies related to the COVID-19 pandemic. As with other areas of COVID-19 research, quality has been highly variable, with low quality studies resulting in poorly or mis-informed policy decisions, wasted resources, and undermined trust in research. To support high quality policy evaluations, in this paper we describe common approaches to evaluating policies using observational data, and describe key issues that can arise in applying these approaches. We hope that this guidance can help support researchers, editors, reviewers, and decision-makers in conducting high quality policy evaluations and in assessing the strength of the evidence that has already been published.

Policy evaluation — far from a simple task in normal circumstances — is particularly challenging during a pandemic. Cross-sectional comparisons of states or countries are likely to be biased by selection into treatment: for example, countries with worse outbreaks may be more likely to

implement policies such as mask requirements. In analyses of changes over time – such as single-unit studies using interrupted time-series or multi-unit comparisons using difference-in-differences or comparative interrupted time-series – it may not be possible to parse apart the effects of different policies implemented around the same time, such as mask mandates paired with limits on social gatherings. Analyses of changes over time may also be biased if disease or human behavioral dynamics are not modelled appropriately. This can be challenging because case counts typically do not grow linearly and there is often a lag between a policy change and a behavioral response.

This guidance should be considered minimal screening to identify low quality policy impact evaluation in COVID-19, but is in no way sufficient to identify high quality evidence or actionability. Decision-makers and researchers should pay particular attention to the relevance of the intervention as it was evaluated to relevant decisions being made. The evaluated impact of a program encouraging mask use through messages might not be informative about mask requirement orders. Differences in level of aggregation may be important, such as ecological fallacy arising from a situation in which areas with higher overall mask use have higher transmission, but transmission is actually lower for individuals wearing masks. Policy impact evaluation is only as useful as the question it asks, data it uses, and the way it is analyzed. Problems with measurement, spillover effects, generalizability, changes in measurement overtime (e.g. varying test availability), statistics, testing robustness to alternative assumptions, and many issues can undermine an otherwise robust evaluation, and are not discussed here.

While this guidance is not comprehensive, it may help inform study designs not covered here. Issues with comparative interrupted time-series and synthetic control methods, for example, are broadly similar to the issues with difference-in-differences analyses we discuss here. Other approaches may include adjustment and matching based observational causal inference designs, instrumental variables and related quasi-experimental approaches, and randomized controlled trials. Each has its own set of practical, ethical, and inferential limitations.

In the face of these challenges, we recommend careful scrutiny and attention to potential sources of bias in COVID-19-related policy evaluations, but we remain optimistic about the potential for robust evaluations to inform decision-making. Researchers and decision-makers should triangulate across a large variety of approaches from theory to evidence, invest in better data and more reliable and useful evidence wherever feasible, clearly acknowledge limitations and potential sources of bias, and acknowledge when actionable evidence is not feasible. We anticipate increasing opportunities for better examining policies moving forward, particularly if policies and interventions are designed with policy impact evaluation and data collection in mind.

The COVID-19 pandemic requires urgent decisions about policies that affect millions of people's lives in significant ways. High quality evidence on the effects of these policies is critical to informing decision-making, but is very hard to generate. Evidence-based decision-making

depends on research that carefully considers potential sources of bias, and clearly communicates underlying assumptions and sources of uncertainty.

COVID-19 Health Policy Impact Evaluation Review

Start of Block: Main form

Q10 **Administrative information**

Q8 Study DOI

_____

Q3 Reviewer number

_____

Q54 Review type/round

The first round (Primary/independent review round) is for the independent first reviews of every article; the second (Secondary/consensus round) is for the second round of review for each article.

- ○ Primary/independent review round (1)
- ○ Secondary/consensus round (2)

Q50 Screening

Q52 Do you wish to recuse yourself from reviewing this study for any reason (e.g. social or professional relationship with the authors, financial conflict of interest, etc)?

- ○ No, I do not wish to recuse myself. (1)
- ○ Yes, I recuse myself from reviewing this paper. (2)

Skip To: End of Survey If Q52 = Yes, I recuse myself from reviewing this paper.

Q51 Do you believe that this study meets the inclusion criteria for this research?

The inclusion criteria are: The primary topic of the article must be evaluating one or more individual COVID-19 policies on direct COVID-19 outcomes The primary exposure(s) must be a policy, defined as a government-issued order at any government level to address a directly COVID-19-related outcome (e.g. mask requirements, travel restrictions, etc). COVID-19 outcomes may include cases detected, mortality, number of tests taken, test positivity rates, Rt, etc. This may NOT include indirect impacts of COVID-19 on things such as income, childcare, trust in science, etc. The primary outcome being examined must be a COVID-19-specific outcome, as above. The study must be designed as an impact evaluation study from primary data (i.e. not primarily a predictive or simulation model or meta-analysis) The study must be peer reviewed, and published in a peer-reviewed journal indexed by PubMed The study must have the title and abstract available via PubMed at the time of the study start date The study must be written in English

- ○ Yes (1)
- ○ No (2) _____

Skip To: End of Survey If Q51 = No

Q7 **Study topic information**

Please consult review guidance (available here) for additional guidance on answering these questions.

Q6 Main impact sentence

Copy and paste the sentence from the abstract that best describes the main claim of the study (e.g. "Policy X had a positive impact on outcome Y")

_____

Q9 Main COVID-19 policy type evaluated

Select all that apply. Note: categorization from the Oxford Government Response Tracker

- ☐ School closing (1)
- ☐ Workplace closing (2)
- ☐ Cancel public events (3)
- ☐ Restrictions on gathering size (4)
- ☐ Close public transportation (5)
- ☐ Stay at home requirements (6)
- ☐ Restrictions on internal movement (7)
- ☐ Restrictions on international travel (8)
- ☐ Income support (9)
- ☐ Debt/contract relief for household (10)
- ☐ Fiscal measures (11)
- ☐ Giving international support (12)
- ☐ Public information campaign (13)
- ☐ Testing policy (14)
- ☐ Contact tracing (15)
- ☐ Emergency investment in healthcare (16)
- ☐ Investment in COVID-19 vaccines (17)

☐ Other policy response (fill in) (18) _____

Q12 Main COVID-19 outcome type evaluated

Select all that apply

☐ COVID-19 cases (1)

☐ COVID-19 test positivity (2)

☐ COVID-19 deaths (3)

☐ COVID-19 hospitalizations (4)

☐ SARS-CoV-2 infections and infection rate (e.g. effective R) (8)

☐ Other (fill in) (9) _____

Q13 **Method(s) identification**

For this section, consider only the data structure as it enters into the main statistical model. In other words, if the original dataset is of individuals at many time points, but the main statistical model uses a regional-level aggregated count of cases, the data as it enters into the main statistical model is a regional aggregate at one time point.

Q14 What is the level of aggregation for the main outcome data?

☐ Individual level (1)

☐ Regional aggregate (e.g. count, mean, etc.) (2)

Q16 How many regional units included in the main statistical model received the policy of interest?

If 2-20, enter the number of regional units analyzed which received the policy of interest.

- ○ One (1) (1)
- ○ Two through twenty (2-20) (2) _____
- ○ More than twenty (21+) (3)
- ○ Unclear or N/A (4) _____

Q17 How many regional units were included which did NOT receive any form of the policy of interest?

If 2-20, enter the number of regional units analyzed which did not receive the policy of interest.

- ○ Zero (0) (1)
- ○ One (1) (2)
- ○ Two through twenty (2-20) (3) _____
- ○ More than twenty (21+) (4)
- ○ Unclear or N/A (5) _____

Display This Question:

If Q17 = Zero (0)

Q25 Did different regions receive different intensities of the policy of interest for comparison?

For example, the study might compare places with more intense versions of policy or policies vs. places with less intense versions of policy or policies, rather than just places with and without the policy or policies.

- ○ Yes (regions with more intense policy were compared with regions with less intense policy) (1)
- ○ No (2)
- ○ Unclear or N/A (3)

Q18 For each regional unit, how many time point observations were in the model *before* the policy was enacted?

- ○ None (0) (1)
- ○ One (1) (2)
- ○ More than one (2+) (3)
- ○ Unclear or N/A (4) _____

Q19 For each regional unit, how many time point observations were in the model *after* the policy was enacted?

- ○ None (0) (1)
- ○ One (1) (2)
- ○ More than one (2+) (3)
- ○ Unclear or N/A (4) _____

Display This Question:

If Q19 = One (1)

And Q18 = One (1)

Or If

Q19 = More than one (2+)

Or If

Q18 = More than one (2+)

Q20 How would you describe the time intervals between observations?

- Days (1-5 days between observations) (1)
- Weeks (about 5-10 days between observations) (2)
- Multiple weeks (11-25 days between observations) (3)
- Monthly (26 or more days between observations) (4)

Display This Question:

If Q17 = Zero (0)

Q21 Did the pre-policy period for any region act as a “control” for different region post-policy enactment?

In other words, was there any pre-period in one or more region's being used to control or compare for the trends of any one or more *different* regions' post-period?

- No (pre-periods were treated as controls only within-region) (1)
- Yes (pre-periods were treated as controls with other regions) (2)
- Unclear or N/A (3)

Q22 Was any unit assigned the policy or the timing of the policy externally (i.e. as an experiment/trial)?

- ○ No (observational data only) (1)
- ○ Yes (treatment assigned as part of research or evaluation) (2)
- ○ Unclear or N/A (3)

Display This Question:

If Q22 = Yes (treatment assigned as part of research or evaluation)

Q23 Was the assignment randomized?

- ○ Yes (1)
- ○ No (2)

Q27 Based on your answers above and the guidance document, please select the type of study that best resembles the design of the main analysis.

Please note that the design(s) named in the paper may not match with the method described below, nor is this the actual exact design that was used. If you believe that the design used differs from the choices below in a way that makes this choice impossible, please contact the study administrator before selecting "other."

				Design		
				Units (e.g., regions of comparison)		
				Time points measured per unit		
				Assumed counterfactual.		
			“If not for the intervention, ___”			
				With intervention		
			Without intervention			
		Before intervention				
	After intervention					
		Cross-sectional				
	At least one					
At least one						N/A
				One time point		
		Outcome in intervention units would have been the same				

as the outcome in the non-intervention units.

Pre/post

At least one

None

At least one (typically one)

At least one (typically one)

Outcome would have stayed the same from the pre period to the post period.

Interrupted time-series (ITS)

At least one

None

More than one

At least one (typically several)

Outcome slope and level* would have continued along the same modelled trajectory from the pre-period to the post period.

Difference-in-differences (DiD)

At least one

At least one†

At least one (typically one)

At least one (typically one)

Outcome in intervention units would have changed as much as (or in parallel with) the outcome in the non-intervention units.

Comparative interrupted time series (CITS)

At least one

At least one†

More than one (typically several)

At least one (typically several)

Outcome slope and level* would have changed as much as non-intervention group's slope and level* changed.

* Assessing both slope and level only applicable if there are multiple data points during the post period † Units without the intervention may be the pre-period of a different unit that eventually receives the intervention.

○ Cross-sectional analysis (1)

○ Non-randomized experiment/trial (2)

○ Randomized controlled trial (3)

- ○ Pre/post (4)
- ○ Interrupted time-series (5)
- ○ Difference-in-differences (6)
- ○ Comparative interrupted time-series (7)
- ○ Other (please contact administrator before selecting) (8) _____

Q49 **Design evaluation**

Display This Question:
If Q27 = Interrupted time-series
Or Q27 = Difference-in-differences
Or Q27 = Comparative interrupted time-series

Q29 Does the analysis provide graphical representation of the outcome over time?

If not "Yes" please describe (three short sentences max).

-Check for a chart that shows the outcome over time, with the dates of interest, separated by policy/non policy groups if applicable. Outcomes may be aggregated for clarity (e.g. means and CIs at discrete time points).

- ○ Yes (1)
- ○ Mostly yes (2) _____
- ○ Mostly no (3) _____
- ○ No (4) _____
- ○ Unclear (5) _____

Display This Question:

If Q27 = Interrupted time-series

Or Q27 = Difference-in-differences

Or Q27 = Comparative interrupted time-series

Q30 Is there sufficient pre-intervention data to characterize pre-trends in the data?

If not "Yes" please describe (three short sentences max).

-Check the chart(s) to see if there are several time points over a reasonable period of time over which to establish stability and curvature in the pre-trends.

- ○ Yes (1)
- ○ Mostly yes (2) _____
- ○ Mostly no (3) _____
- ○ No (4) _____
- ○ Unclear (5) _____

Display This Question:

If Q27 = Interrupted time-series

Or Q27 = Difference-in-differences

Or Q27 = Comparative interrupted time-series

Q32 Is the pre-trend stable?

If not "Yes" please describe (three short sentences max).

-Check if there are sufficient data to reasonably determine a stable functional form for the pre-trends, and that they follow a modelable functional form.

- Yes (1)
- Mostly yes (2) _____
- Mostly no (3) _____
- No (4) _____
- Unclear (5) _____

Display This Question:
If Q27 = Interrupted time-series
Or Q27 = Comparative interrupted time-series

Q31 Is there sufficient post-intervention data to observe post trends in the data?

If not "Yes" please describe (three short sentences max).

-Check the chart(s) to see if there are several time points over a reasonable period of time over which to establish stability and curvature in the post- trends.

- Yes (1)
- Mostly yes (2) _____
- Mostly no (3) _____
- No (4) _____
- Unclear (5) _____

Display This Question:
If Q27 = Interrupted time-series
Or Q27 = Difference-in-differences
Or Q27 = Comparative interrupted time-series

Q33 Is the functional form of the counterfactual (e.g. linear) well-justified and appropriate?

If not "Yes" please describe (three short sentences max).

-Check whether the authors explain and justify their choice of functional form.
-Check if there is any curvature in the pre-trend.
-Consider the nature of the outcome. Is it sensible to measure the trend of this outcome on the scale and form used? Note: infectious disease dynamics are rarely linear.
-Consider that while pre-trend fit is a necessary condition for an appropriate linear counterfactual model, it is not sufficient. Check if the authors provide justification for the functional form to continue to be of the same functional form (e.g. linear).

- ○ Yes (1)
- ○ Mostly yes (2) _____
- ○ Mostly no (3) _____
- ○ No (4) _____
- ○ Unclear (5) _____

Display This Question:
If Q27 = Interrupted time-series
Or Q27 = Difference-in-differences
Or Q27 = Comparative interrupted time-series

Q34 Is the date or time threshold set to the appropriate date or time (e.g. is there lag between the intervention and outcome)?

If not "Yes" please describe (three short sentences max).

-Check whether the authors justify the use of the date threshold relative to the date of the intervention.
-Trace the process between the intervention being put in place to when observable effects in

the outcome might appear over time.
-Consider whether there are anticipation effects (e.g. do people change behaviors before the date when the intervention begins?)
-Consider whether there are lag effects. (e.g. does it take time for behaviors to change, behavior change to impact infections, infections to impact testing, and data to be collected, etc?)
-Check if authors appropriately and directly account for these time effects.

- ○ Yes (1)
- ○ Mostly yes (2) _____
- ○ Mostly no (3) _____
- ○ No (4) _____
- ○ Unclear (5) _____

Display This Question:

If Q27 = Interrupted time-series

Q36 Is this policy the only uncontrolled or unadjusted-for way in which the outcome could have changed during the measurement period?

If not "Yes" please describe (three short sentences max).

-Consider other policies or interventions which could impact the outcome during this time.
-Consider social behaviors changed which could meaningfully impact the outcome during this time.
-Consider economic conditions changed which could meaningfully impact the outcome during this time.
-Note that the actual concurrent changes do not need to happen during the period of measurement, just their effects.

- ○ Yes (1)
- ○ Mostly yes (2) _____
- ○ Mostly no (3) _____
- ○ No (4) _____
- ○ Unclear (5) _____

Display This Question:
If Q27 = Difference-in-differences
Or Q27 = Comparative interrupted time-series

Q53
Is this policy the only uncontrolled or unadjusted-for way in which the outcome could have changed during the measurement period, differently for policy and non-policy regions?
If not "Yes" please describe (three short sentences max).

-Consider any uncontrolled factor which could have influenced the outcome differently in policy and non-policy regions.

-This may include (but is not limited to)
-Other policies
-Social behaviors
-Economic conditions
-Are these factors justified as having negligible impact?
-If justified, is the argument that these have negligible impact convincing?

-Note that the actual concurrent changes do not need to happen during the period of measurement, just their effects.

- ○ Yes (1)
- ○ Mostly yes (2) _____
- ○ Mostly no (3) _____
- ○ No (4) _____
- ○ Unclear (5) _____

Display This Question:

If Q27 = Interrupted time-series

Or Q27 = Difference-in-differences

Or Q27 = Comparative interrupted time-series

Q38
Did authors provide diagnostics or show robustness and/or sensitivity of results to alternative model choices?

If not "Yes" please describe (three short sentences max).

- ○ Yes (1)
- ○ Mostly yes (2) _____
- ○ Mostly no (3) _____
- ○ No (4) _____
- ○ Unclear (5) _____

Display This Question:

If Q27 = Interrupted time-series

Or Q27 = Difference-in-differences

Or Q27 = Comparative interrupted time-series

Q39 Given the above, do you believe that the design is appropriate for identifying the policy impact(s) of interest?

This should be taken as independent of what you believe about other studies, and/or the feasibility of other designs.

If not "Yes" please describe (three short sentences max).

- Yes (1)
- Mostly yes (2) _____
- Mostly no (3) _____
- No (4) _____
- Unclear (5) _____

Display This Question:

If Q54 = Secondary/consensus round

Q55 General and/or additional comments on this paper from consensus discussion. This may include any additional information worth commenting on regarding the paper, difficulties encountered evaluating it, etc.

(three short sentences max)

_____

End of Block: Main form