

Unravelling the mess that is mask efficacy research

An introduction to mechanism-informed narrative synthesis

Trish Greenhalgh

Acknowledging UKRI, core research team, advisory group members, and the AI group in Trish's department at Oxford



The
literature
on mask
efficacy is
a mess

Synthesis challenges in complex evidence: A critical analysis of systematic reviews of face mask efficacy

Trisha Greenhalgh ¹, Sahanika Ratnayake², Rebecca Helm³, Luana Poliseli² and Jon Williamson ²

- 66 previous systematic reviews
- Most were 'PRISMA-compliant'
- 37 concluded that masks were effective or that respirators were more effective than medical masks
- 29 concluded that masks were ineffective or that respirators were no better than medical masks



Research questions

Empirical question:

What is the efficacy of face masks in reducing transmission of respiratory infections?

Methodological question:

How can we make sense of a complex body of evidence using narrative synthesis?

"Technical" question:

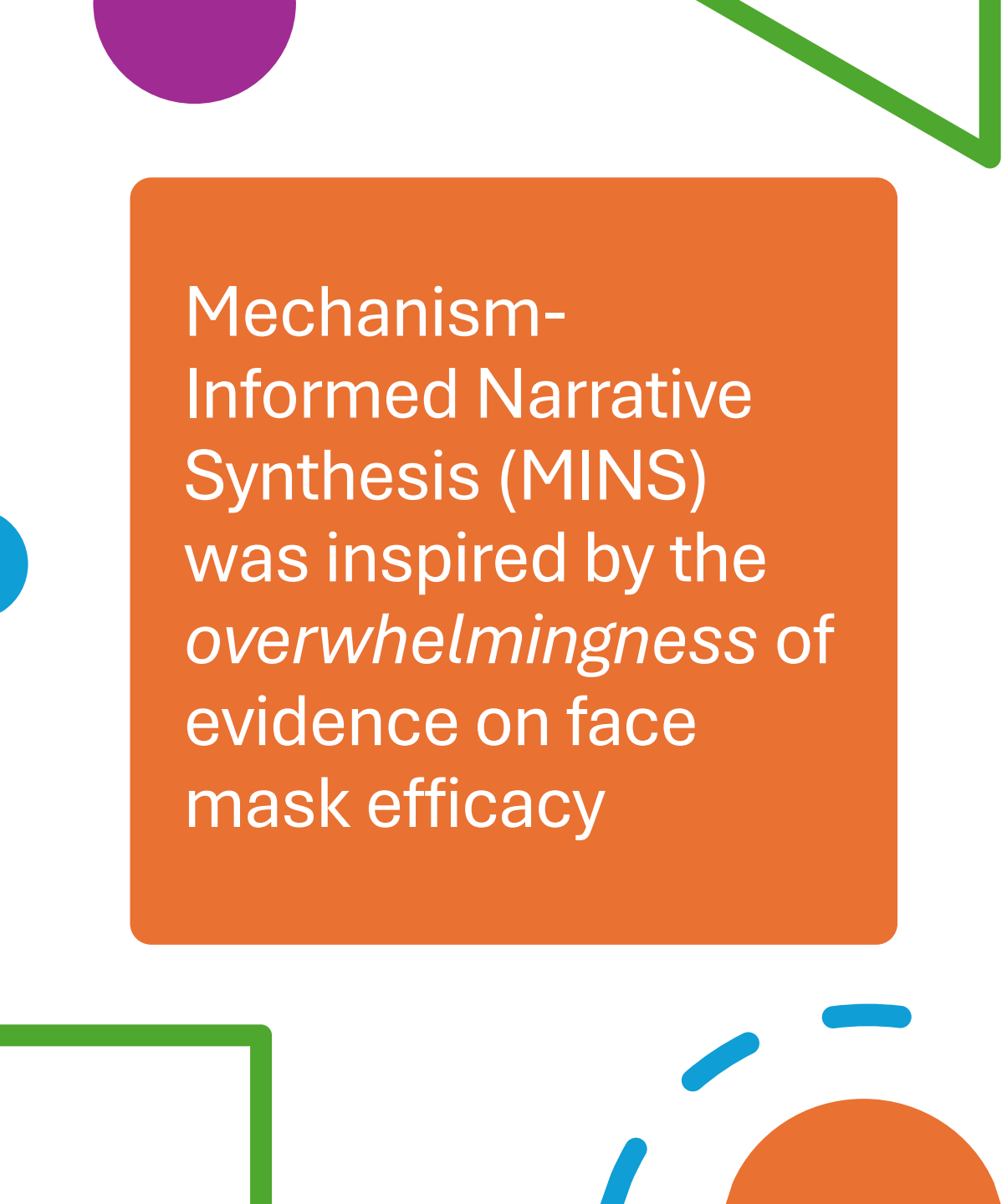
Can we use AI to help with this?

Evidential pluralism

Two kinds of evidence are needed to demonstrate causality


- Evidence of **association** (e.g. from RCTs)
- Evidence of **mechanism** (e.g. from laboratory studies)





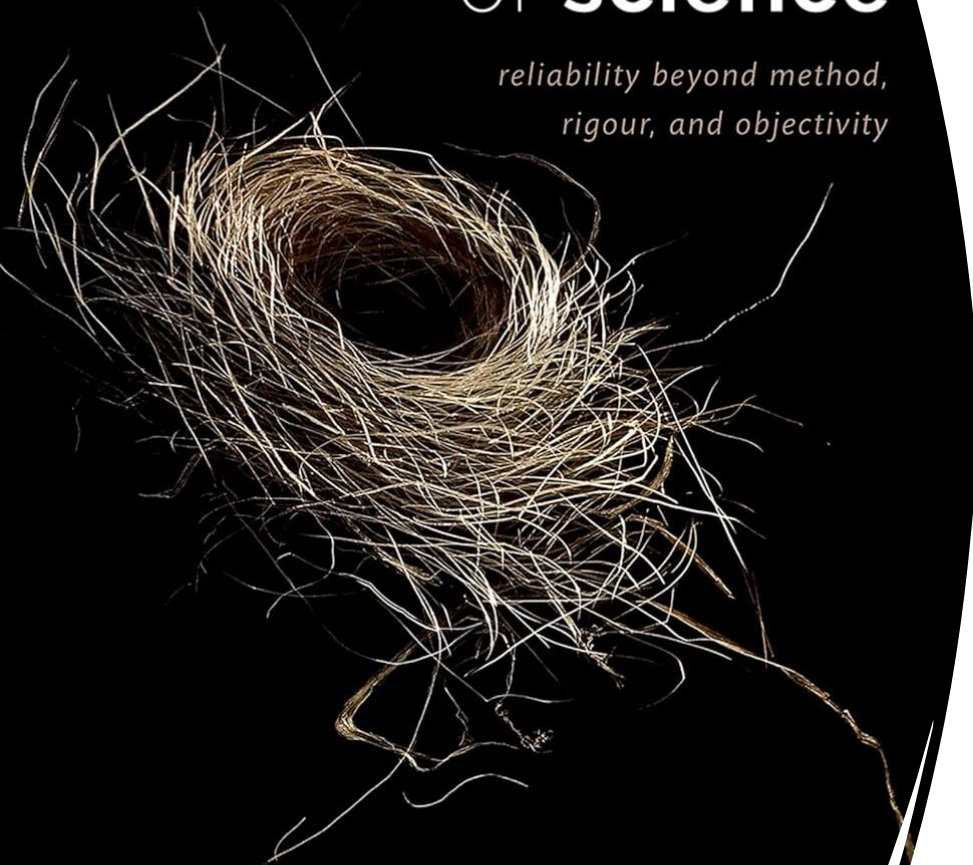
Mechanism-
Informed Narrative
Synthesis (MINS)
was inspired by the
overwhelmingness of
evidence on face
mask efficacy

OVERWHELMING BECAUSE

- Hundreds of studies
 - Multiple study designs incl weird ones
 - Crosses many disciplines (material, behavioural, epidemiological)
 - Multi-level causality ('web')
 - Emergent interactions between variables and context
 - Technically difficult to understand
- 

the tangle of science

reliability beyond method,
rigour, and objectivity

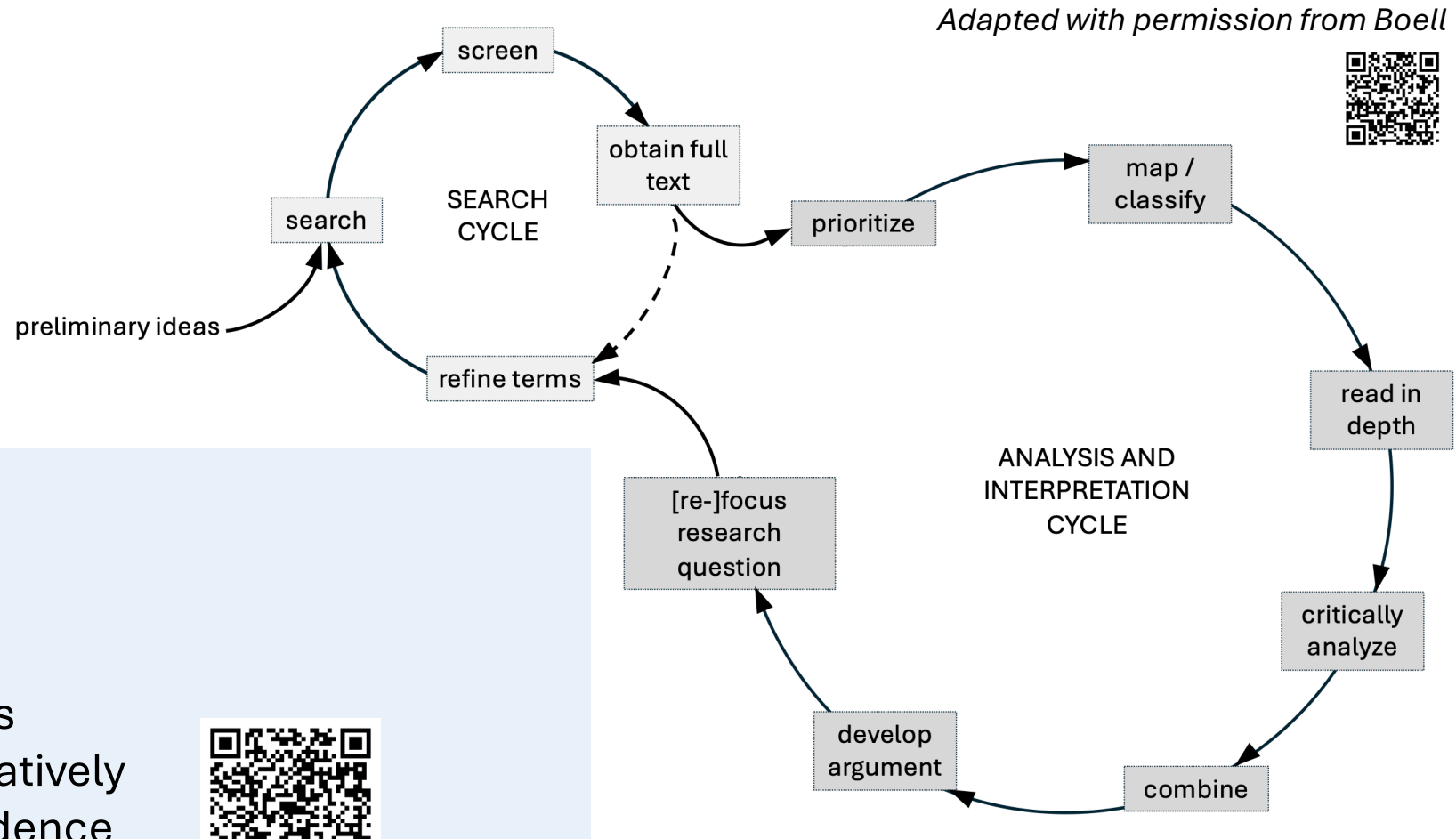
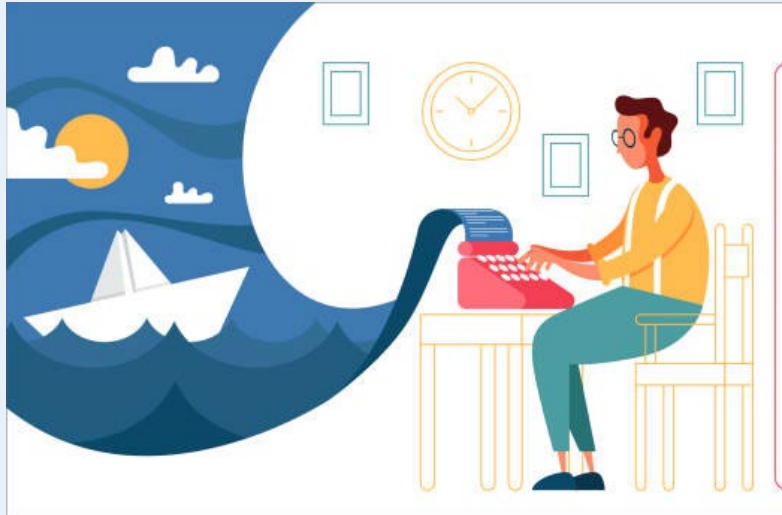


Science as a 'tangle'

- Scientific findings are not atomistic; they are nested in a 'tangle' of other related findings
- All these findings must *cohere*
- Any individual finding must *make sense* in relation to the wider body of knowledge
- Associative findings are made more robust if they cohere with mechanistic findings



Narrative (hermeneutic) review



- Seeks clarity and understanding
- Addresses broad questions
- Includes a range of study designs
- Works by hermeneutic logic: iteratively refining an account of all the evidence
- Qualitative and quantitative

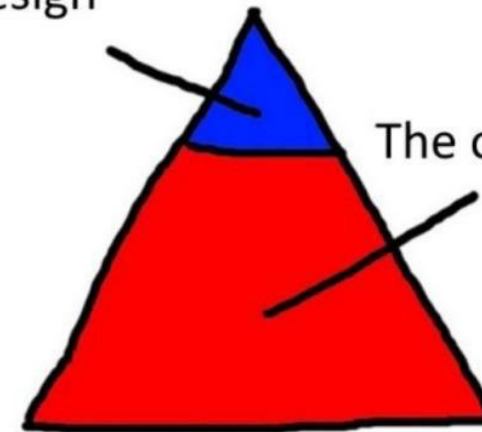


MINS:

A 5-step method,
supported by AI

STEP 1: Search

Thoughtful, well-conducted studies of
any design



The other shit



MINS:

A 5-step method,
supported by AI



STEP 1: Search

STEP 2: Explain and clarify

*...contagiousness
of different
diseases*

*...ballistics of
droplets emitted in a
sneeze*

Draft explanatory account (in words)

*...physics of
how airborne
particles spread*

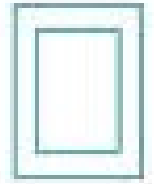
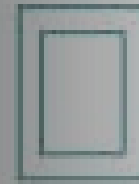
*...material science
of cloth, paper and
polypropylene
filters*

*...pathophysiology of
infectious particle
production and
shedding*

*...behavioural
science of why
people do and
don't wear
masks*

*...physiology of
how masks
affect oxygen
levels (or not)*

*...temporal dynamics
of unfolding disease
outbreaks*



Example of a *mechanistic* study

- A study from South Korea looked at over 300 people infected with SARS-CoV-2, more than one-third of whom were asymptomatic. Using serial PCR measurements, they estimated viral shedding over several days. Asymptomatic patients shed similar viral loads to symptomatic ones.





MINS:

A 5-step method,
supported by AI

STEP 1: Search

STEP 2: Explain and clarify

**STEP 3: Develop bespoke critical
appraisal tools (using AI)**

Quality assessment template for mask efficacy studies (example)

Outbreak investigations

A. Outbreak description & context

1. Outbreak setting & period: _____
2. Purpose of investigation: Descriptive Analytic Control evaluation
3. Case definition: Clear & consistent? Yes Partial No
4. Case finding completeness: Methods (line list / tracing / testing): _____
Completeness: _____% ; Asymptomatic cases included? Yes No

B. Exposure measurement (incl. masking)

5. Exposure definition explicit? Yes Partial No
6. Mask exposure detail: Type? _____ ; Fit? Yes/No ; Timing relative to infectious period? Yes/No
Misclassification risk: Low Moderate High
7. Other exposures measured? (distancing, ventilation, contact duration, activity)
 Yes Partial No

C. Environmental & spatial context

8. Setting documented? (indoor/outdoor, ventilation, room size, crowding)
 Fully Partially Not described
9. Spatial/cluster structure described? (household, seating map, proximity)
Clustering analysed? Yes No

D. Sampling & representativeness

10. Sampling approach: Census Probability Convenience
Participation rate: _____% Missing contact follow-up: _____%
Selection bias likely? Yes No Unclear

E. Descriptive epidemiology

11. Timeline / epidemic curve included? Yes No
12. Transmission sequence plausible? Yes Partial No

F. Analytic methods & confounding

13. Analytic component present? Cohort Case-control Descriptive only
14. Confounders measured/controlled? (setting, ventilation, behaviour, duration)
Control adequate? Yes Partial No
15. Cluster effects modelled? Yes No
16. Effect estimate: Measure (RR/OR): _____ ; 95% CI: _____

G. Bias assessment

17. Exposure misclassification risk: Low Moderate High
18. Selection bias: Low Moderate High
19. Temporal bias (reverse causation): Yes Possible No
20. Information bias (recall): Low Moderate High
21. Environmental bias (ventilation missing): Major Minor None

H. Missing data

22. Missing exposure data: _____%
23. Missing outcome data: _____%
Could missingness distort direction? Toward null Away from null Either Unclear

I. Interpretation & plausibility

24. Biological/aerosol plausibility considered? Yes No
25. Alternative explanations evaluated? (superspreading, airflow failure)
 Yes No

J. Reporting transparency

26. Limitations clearly stated? Yes Partial No
27. Protocol/data methods transparent? Yes Partial No

Overall quality rating: Low / Moderate / High concerns (*explain*)



MINS:

A 5-step method,
supported by AI

STEP 1: Search

STEP 2: Explain and clarify

STEP 3: Develop tools

**STEP 4: Apply tools to appraise
individual studies (using AI)**

Example of completed study appraisal template

+ HUMAN CHECKING

Benkouiten et al. 2013

Study Design Identified: Prospective cohort study with serial virological sampling of a French Hajj pilgrim cohort (2012 Hajj). Three sampling points: pre-travel, during Hajj, pre-return, plus symptom surveillance. Design is correctly classified and requires **Cohort Checklist**. Citations used <<>>

Quantitative Extraction Table

Item	Extracted Value
Initial cohort	169 recruited
Completed pre-travel swab	165 (97.6%)
Completed post-travel swab	154 (91.1%)
Symptomatic mid-Hajj sampled	70 (41.4%)
Respiratory symptoms during Hajj	90.4% reported ≥1 symptom
Cough	83.4%
Sore throat	79.7%
Rhinorrhoea	68.5%
Subjective fever	45.4%
ILI (self-reported triad)	41.0%
Pre-Hajj virus positivity	4.8% (5 HRV, 1 FLUC, 1 HAdV, 1 HEV)
During Hajj positivity	38.6% of symptomatic (19 HRV, 6 FLUA, 1 FLUC, 1 RSVB, 1 HMPV, 1 HAdV, 1 HEV; 3 coinfections)
Pre-return positivity	11%
Mask use	55.1%
Handwashing “frequent”	40.3% (associated with more symptoms)
Hand sanitizer use	46.3% (associated with higher viral detection)
Chronic conditions	57.5% (DM 27.5%, HTN 26.3%, chronic resp disease 7.8%)

Completed Checklist — Cohort Study

A. Population, Recruitment, Follow-up

- Recruitment from *single Marseille travel agency* → not representative of wider French or global pilgrim population.
- Very high follow-up for virological sampling (≈90%).
- Daily symptom reporting by accompanying physician increases internal validity.

B. Exposure Measurement (Mask, Hygiene, NPIs)

- NPIs self-reported post-travel, subject to **substantial recall bias**.
- Mask use binary (yes/no) without type, fit, frequency, correct use.
- Handwashing classified as “usual” vs “more frequent” → **reverse causation likely** (symptomatic pilgrims wash hands more).
- No observational verification of NPI adherence.

C. Outcome Measurement (Respiratory Illness & PCR Detection)

- Respiratory symptoms captured prospectively.
- Virological outcome measured objectively with rRT-PCR for **11 viruses** → strong laboratory validity.
- Sampling delays at some points (e.g., specimens stored at 20°C up to 30 days) → **possible under-detection** of RNA viruses.
- ILI definition differs from WHO standard (subjective fever).

D. Confounding

Major confounders **not controlled**:

- Density of exposure (Tawaf vs tent vs bus).
- Mask correctness, mask type.
- Ventilation quality.
- Age, chronic disease status (important given 58% had comorbidities).
- Behaviour during Hajj differing by symptom onset (reverse causation).

E. Missing Data

- Very little virological missingness, but behavioural missingness not fully reported.
- Symptom onset timing retrospective for many participants → temporal ambiguity.

F. Effect Estimate (for NPIs)

- Study **does not estimate NPI effect sizes**.
- Associations reported:
 - “Frequent handwashing” associated with *more* fever/ILI and higher viral positivity → clearly reverse causation.
 - Hand sanitizer use associated with higher viral carriage (same issue).
- No statistical analysis of mask effect at all.
→ **Cannot infer mask efficacy from this study**.

Bias Mechanisms Detected (from Bias Dictionary)

Toward the Null (attenuation)

- Self-reported mask/hygiene behavior → non-differential misclassification.
- No assessment of mask fit → expected strong attenuation (Background Science).
- Severe shared-air aerosol exposure during Tawaf and tents → dominant transmission mode.
- Delayed sample processing for some specimens → underdetection.

Away from the Null (inflation)

- Reverse causation: symptomatic pilgrims practice more hygiene.
- Behaviour clustering not measured.
- Overcrowding exposures correlate with both symptoms and behavioural changes.
- Retrospective symptom logs → recall bias inflates reported symptom clustering after Tawaf.

Dominant Bias Direction

For NPIs: **inflation and attenuation simultaneously, producing uninterpretable associations**.

For viral circulation dynamics: biases small (PCR is objective) → findings reliable for **viral acquisition**, not for behavioral effects.

Estimated Magnitude of Bias

- Exposure misclassification (NPIs): 30–60% attenuation.
- Reverse causation for hygiene variables: very large inflation.
- Outcome misclassification minimal for PCR, moderate for clinical symptoms.

Net: the study robustly documents *rapid viral acquisition*, but provides **no interpretable evidence about mask efficacy**.

Causal Diagram Assessment

Critical unmeasured nodes:

- Tawaf → extreme density → infection risk.
- Ventilation and indoor crowding pathways.
- Symptom onset → behaviour change → measured NPI adherence.
- Mask correctness and filtration.

DAG: **no identification strategy** for mask or NPI causal effects.

Missing STROBE Elements

- No description of NPI variable validity.
- No time-resolved exposure-outcome modelling.
- No adjustment for confounders.
- Missing handling for stored-sample degradation.
- No symptom severity quantification.

Plain-Language Summary

Benkouiten et al. conducted the strongest virological study of Hajj pilgrims to date, sampling a French cohort before, during, and after the 2012 Hajj. Nearly all pilgrims developed respiratory symptoms, and PCR detection of respiratory viruses—especially rhinovirus—rose from 4.8% pre-Hajj to 38.6% in symptomatic pilgrims during Hajj, and remained at 11% at departure. These findings clearly establish intense viral circulation and high acquisition risk in the crowded Hajj environment.

However, the study was **not designed to estimate mask efficacy or NPI effectiveness**. Mask use, handwashing, and sanitizer use were self-reported and confounded by reverse causation: symptomatic pilgrims adopted more hygiene. No mask-type or correct-use data were collected. Therefore, although this study is excellent for documenting viral spread, it cannot be used to infer any causal impact of masking.

Final Quality Judgment

Moderate concerns for viral dynamics; High concerns for NPI/mask inference.

The study provides high-quality evidence on *viral acquisition*, but **no valid estimate of mask or NPI efficacy** due to misclassification, confounding, and behavioural reverse causation.



MINS:

A 5-step method,
supported by AI

STEP 1: Search

STEP 2: Explain and clarify

STEP 3: Develop tools

STEP 4: Appraise individual studies

STEP 5: Synthesise

Seven cognate groups of community mask studies

Household studies ('infected family member')

RCTs in low-risk community settings

Outbreak investigations

Case-control studies e.g. from testing centres

Prospective cohort studies

Cross-sectional disease prevalence studies

Mass gatherings



Empirical findings
(so far only completed for
community based studies)

1.
Mask efficacy
relies on multiple
linked causal
pathways whose
mechanisms
include...

- Pathophysiological and environmental mechanisms (e.g., pathogen shedding, aerosolization, airborne spread)
- Physical and material mechanisms (e.g., filtration, breathability, fit)
- Behavioural mechanisms (e.g., adherence, risk avoidance, norm-setting)
- Mechanisms by which adverse effects (e.g. headache) reduce adherence
- Epidemiological mechanisms (e.g., contagiousness, incubation period, temporal dynamics, overdispersal).

2.
These
mechanisms
explain how
findings of
primary studies
were distorted
due to (e.g.)

- Design flaws, including low-filtration / low breathability masks, poor fit, misaligned exposure windows
- Underpowering due to lower-than-expected disease prevalence
- Confounding (especially when both 'maskers' and 'non-maskers' are exposed to contaminated indoor air)
- Post-allocation biases (including low adherence, crossover, differential testing)

Empirical potential

Features of context / setting that increase a study's potential to demonstrate an effect

- Disease factors: contagiousness, prevalence, overdispersal
- Air quality factors: under-ventilated spaces, crowding, time spent indoors
- Physiological factors: e.g. vocalising, exercising, close contact

Pre-existing confounders

Characteristics of participants that may distort the effect of mask-wearing on infection risk

DEMOGRAPHIC

- Age
- Occupation
- Comorbidities
- Immunity (vaccination, past infection)

BEHAVIOURAL*

- Propensity to engage in protective behaviours e.g. handwashing, distancing
- Propensity to seek testing

ENVIRONMENTAL e.g.

- Household crowding
- Area deprivation
- Public transport use

** Behaviours may also be influenced by mask use*

PRE-DATING THE CAUSALITY WINDOW

Intervention

Masking when there is potential for infection

THE FOCAL CAUSAL RELATIONSHIP

How masks reduce transmission

- Source control: ↓ exhaled dose of pathogen, ↓ turbulence and trajectory of infectious plume
- Wearer protection: ↓ inhaled dose of pathogen; altered pathogen path (smaller → larger airways)
- Social influence: ↑ mask uptake strengthens social norm, making others more likely to mask

INTERVENTION FIDELITY

How the intervention may be compromised

- Suboptimal mask properties: low filtration factor, low fit factor (leakage), low breathability
- Suboptimal adherence e.g. improper wearing, limited duration, limited settings, attrition

INTERVENTION MEASUREMENT

How the intervention as documented may differ from the intervention received

- Suboptimal measurement e.g. self-reports affected by forgetting or social desirability bias
- Lack of detail on e.g. type of mask, when/where worn ('often', 'sometimes'), mistiming relative to key exposure window, failure to capture very high exposure situations (e.g. unmasked co-sleeping)

DURING THE CAUSALITY WINDOW

Outcome

Presence of new infection

TESTING POLICY AND PRACTICE

How testing access and behaviour shape who is detected as infected

- Eligibility: e.g. only symptomatic people or contacts of cases are offered testing
- Routine screening of certain groups
- Differential test-seeking: health-conscious people and those with symptoms seek testing more
- Differential testing intensity: one arm of the study gets tested more

OUTCOME MEASUREMENT

How new infections as documented differ from actual new infections

- Symptom-based outcomes miss mild/asymptomatic cases and mix pathogens.
- Unreliable ascertainment if self-report or self-swab → misclassification
- Test misaligned with seroconversion / incubation (e.g., serology too early; no lag)
- Serology noise (e.g. from vaccination)
- Blunt population proxies: ecological metrics (e.g. R_t) are downstream, non-specific, and dilute individual-level effects

AFTER THE CAUSALITY WINDOW

3. Confounding in observational studies could sometimes be adjusted for

- Quantitative bias analysis (QBA) was useful for estimating the range of true effects in a study under plausible real-world assumptions, allowing judicious inclusion of selected non-randomised studies
- QBA could be accelerated using AI (with extensive human checking)

Empirical conclusion

There is consistent evidence that face masks can meaningfully reduce transmission of respiratory diseases under certain behavioural and environmental conditions

BUT Multiple residual biases and substantial variation in these conditions mean that **point estimates of efficacy vary widely and have limited generalisability**

Publications (mostly still ‘under review’)



Greenhalgh T et al. Mechanism-informed narrative synthesis (MINS) of complex evidence: A new systematic review methodology illustrated with face mask efficacy. *Research Synthesis Methods* 2026, under review.



Greenhalgh T et al. Explaining heterogeneous findings in community-based mask efficacy studies: Systematic review using Mechanism-Informed Narrative Synthesis. *Research Synthesis Methods* 2026, under review.



Greenhalgh T et al. The unbearable lightness of systematic review: A critical analysis of illusory visuals in evidence synthesis. *Science, Technology and Human Values* 2026, under review.



Greenhalgh T et al. When the SPIRIT moves you: Protocol changes can introduce bias in non-inferiority trials. *BMJ* 2026; 393:s725.



Greenhalgh T, Ratnayake S, Helm R, Poliseli L, Williamson J. Synthesis challenges in complex evidence: A critical analysis of systematic reviews of face mask efficacy. *Research Synthesis Methods*. 2026 Feb 6.



Greenhalgh T et al. The randomised controlled trial as cultural product: an illustrative review using mask efficacy trials. *Social Science and Medicine Qualitative* 2026; in press.

More coming on healthcare worker studies – *amuse bouche*:

OPINION



¹ University of Oxford, Oxford, UK

² Rockyview General Hospital, Calgary, AB, Canada

³ University of Calgary, Calgary, AB, Canada

⁴ Ontario School Safety

Cite this as: *BMJ* 2026;**393**:s725

<http://doi.org/10.1136/bmj.s725>

Published: 17 April 2026



When the SPIRIT moves you: protocol changes can introduce bias in non-inferiority trials

Protocol fidelity and transparency about changes are essential to the credibility of randomised controlled trials

Trisha Greenhalgh,¹ Samantha Lovell,² Joe Vipond,³ Mark Ungrin,³ Mary Jo Nabuurs⁴

In recent weeks, the UK has seen masked students queueing for emergency meningitis jabs and the publication of the module 3 of the UK covid-19 inquiry,¹ which criticised the government's reliance on flawed advice that the virus did not spread

a finding of non-inferiority reflects the intervention being tested rather than artefacts of design drift.

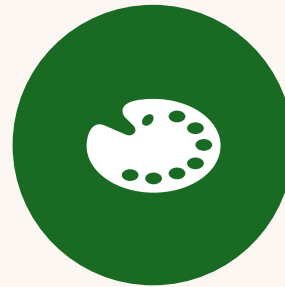
One pertinent example is retrospective changes to the protocol of a randomised controlled

non-inferiority trial published in *Annals of Internal*

MINS involves 4 shifts



In review logic: from
'extract and aggregate'
to 'explain and clarify'



In how we value
evidence: from
'hierarchy' to 'diversity'



In the tools we use:
from 'risk-of-bias' to
'mechanisms-of-bias'



In how we use AI: from
'process accelerator' to
'conversational agent'

Thank you for your attention

NB There is another lecture (and paper) on how the AI performed

trish.greenhalgh@phc.ox.ac.uk

Ackn: Jon Williamson, Rebecca Helm,
Sahanika Ratnayake, Luana Poliseli



MINS: A 5-step method, supported by AI

