The uses and abuses of systematic reviews: the case of ivermectin in Covid-19

*Edmund J Fordham PhD FInstP EurIng

Theresa A Lawrie MBBCh PhD

Katherine MacGilchrist MSc

EbMCsquared, a Community Interest Company
Northgate House, Upper Borough Walls, Bath BA1 1RG, UK

Andrew Bryant MSc
Population Health Sciences Institute, Newcastle University
Baddiley-Clark Building, Richardson Road, Newcastle upon Tyne NE2 4AX, UK

*Corresponding author.
Address: EbMCsquared CIC
Northgate House, Upper Borough Walls,
Bath BA1 1RG, UK

Word count: 2839 words excl Refs
2 Tables; 2 Figures
55 references
Background

A recent Cochrane review\(^1\) of ivermectin for the treatment and prevention of Covid-19 stands in contrast to the world-first PRISMA guidelines\(^2\) systematic review and meta-analysis by Bryant \(et\ al.\)^\(^3\) (hereinafter “Bryant”), which is criticised in the review with further commentary elsewhere\(^4\). We review here the approach of Popp \(et\ al.\)^\(^1\) (hereinafter “Popp”).

Ivermectin\(^5\), a WHO Essential Medicine\(^6\) of negligible cost\(^7\) and outstanding safety record\(^8\)–\(^13\) has attracted significant attention for its potential use both in treatment and in the prevention of Covid-19, from its known anti-viral effect over a wide class of RNA viruses\(^14\) and specific action\(^15\) \textit{in vitro} against SARS-CoV-2. These initial reports were bolstered by multiple reports\(^16\) of efficacy at all stages of the Covid-19 illness (from prophylaxis through to critical care\(^17\)), \textit{in vitro}\(^18\) and \textit{in silico}\(^19,20,21\) studies of its likely modes of action, clinical experience reports from multiple contexts\(^22\), population-level case studies\(^23\), retrospective observational studies\(^24\), through to randomised controlled clinical trials\(^3\) (RCTs). These reports occupy all levels of the so-called “Evidence Quality Pyramid” in the doctrines of Evidence-Based Medicine.

Systematic Reviews (SRs) and Meta-Analyses (MAs) occupy, in theory, the summit of the pyramid, and are considered the premier quality of evidence by Health Regulatory Authorities (HRAs) and for clinical practice guidelines by national and international health services. However, though SRs and MAs are carried out to identify, qualify or avoid the worst aspects of reporting bias, excluding all other evidence can become itself a powerful source of bias.

The Cochrane organisation\(^25\) is dedicated to the development of a library of SRs, confined to the evidence. Cochrane SRs are traditionally conducted by volunteer reviewers, reducing the potential for financial conflicts of interest. Advocacy is not properly part of a SR but of processes such as DECIDE\(^26\) in which other stakeholders including public health professionals and patients should be involved. Cochrane has a tradition of accuracy, objectivity and attention to reliability through systematised but necessarily subjective criteria. Similar criteria are more generally available through the guidelines of the PRISMA Working Group\(^2\), followed in Bryant\(^3\).
However, SR methodologies are capable of abuse in respect of (i) choice of inclusion criteria (ii) choice of outcome measures, comparisons and analyses (iii) the subjectivity inevitable in Risk of Bias assessments, even when codified procedures and criteria are observed. We analyse these choices in Popp¹.

The Cochrane Review of Popp¹: synopsis

The stated objectives are to assess efficacy and safety of ivermectin compared to: no treatment, placebo, “standard of care” (SoC) or any “proven” intervention. The analyses compare hospitalised patients and outpatients separately, without pooling of data. The objectives also included the assessment of prevention of Covid-19 infection with ivermectin. The stated intention is “to provide a complete evidence profile”.

The inclusion criteria were restricted to randomised trials (RCTs), and effectively excluded combination therapies by a requirement for co-interventions to be identical in both arms. Comparators held to be “unproven” were also excluded, as were groups with diagnoses without confirmation by PCR or antigen tests. Outcome measures had pre-specified time-points; mortality up to 28 days, and laboratory-confirmed infection at 14 days.

Only studies considered at low RoB or “some concerns” were included in the “primary” analyses; high RoB analyses are reported in “secondary” analyses and regarded as sensitivity analyses. Primary comparisons are summarised in Table 1 with the secondary comparisons for the prophylaxis comparison, where mortality at 28 days was held to be primary and the more relevant symptomatic infection is secondary.

These analyses concentrate on mortality, improvement (to discharge or resolution), deterioration (to need for mechanical ventilation, or need for oxygen or other respiratory support), and Covid-19 infection. Adverse events are compared in addition though a SR of limited scope cannot supersede the extensive pharmacovigilance already conducted over literally billions²⁷ of ivermectin treatments. Table 1 also shows the total number of patients for each comparison, with the number of events, as a rough indicator the statistical power in each. Very few are actually meta-analyses, because in the majority only a single study is included. Those that do pool data comprise just two studies each.
Unsupported assertions

The background contains unsupported assertions, e.g. that only Non-Pharmaceutical Interventions (NPIs) are available to control infection, whilst exhibiting in Analysis 3.1 a single prophylaxis trial of strong effect and narrow confidence intervals. It is claimed that “serious adverse events (SAE) including vision problems, neurotoxicity and liver damage can occur”, though the cited source\textsuperscript{28} contains no such reports. Moreover the considerable literature\textsuperscript{8-13} on safety is ignored. Professor Chris Whitty (Chief Medical Officer for England), has himself stated\textsuperscript{9}: “The drug has proven to be safe. Doses up to 10 times\textsuperscript{10} the approved limit [200 mcg/kg] are well tolerated by healthy volunteers. Adverse reactions are few and usually mild”.

The review of (potential) mechanisms of action rehearses an outdated objection that whilst Ivermectin slows the reproduction\textsuperscript{15} of SARS-CoV-2, “such effects would need major doses [e.g.\textsuperscript{29}] in humans”. This conjecture ignores other pharmacokinetic modelling\textsuperscript{30}, the absence of a full immune response in the \textit{in vitro} culture\textsuperscript{22} and has long been falsified by the results of trials and clinical experience worldwide\textsuperscript{16}, at 1 – 3 times the anti-parasitic dose\textsuperscript{31}. The sole mechanism considered is the blocking of nuclear import of viral proteins by interaction with the Importin $\alpha/\beta 1$ hetero-dimer\textsuperscript{18}; routinely cited in early reports, this conjecture has long been extended by findings on its anti-inflammatory properties\textsuperscript{17} and multiple \textit{in silico} models of its likely interference with critical viral replication enzymes\textsuperscript{19-21} and blocking of the spike protein / ACE2 interaction\textsuperscript{20}.

Inconsistencies in design

Though denigrating Bryant\textsuperscript{3} as a “bowl of colourful fruit salad”\textsuperscript{4}, Popp\textsuperscript{1} is inconsistent regarding (potentially) active comparators. They state that hydroxychloroquine “does not work” for covid-19, yet excluded trials with comparators including hydroxychloroquine, a drug they hold to be inactive, effectively placebo (which would be included). Conversely, remdesivir is asserted as an “eligible active comparator” for future updates (though latest evidence\textsuperscript{32} reports “no clinical benefit”). There is no logic in mixing inactive controls with “deemed-active” ones, whilst rejecting “unapproved” controls declared to be inactive. “HRA approval” and “clinical demonstration” are sadly not the same thing, and HRA policies on SoC vary internationally. Gorial\textsuperscript{33} reports that hydroxychloroquine was an Iraqi SoC for
Covid-19, creating an inclusion conflict, had Gorial\textsuperscript{33} been a RCT. Hashim\textsuperscript{34} states that azithromycin was also an Iraqi SoC. Babalola\textsuperscript{35} is excluded, though the lopinavir/ritonavir comparator was at the time a SoC in Nigeria.

The design in Popp\textsuperscript{1} is consistent only in rejecting “unapproved” interventions but including “approved” ones, irrespective of actual efficacy. Popp\textsuperscript{1} thus mixes clinical evidence with HRA policies that may not be sound, or even consistent internationally, and encompasses additional controversies e.g. regarding hydroxychloroquine, or remdesivir\textsuperscript{32}. Yet the evidence does not change even if HRA policy does. A research design including politically-driven HRA policies is certain to corrupt a search for objective evidence.

Similarly, the exclusion of combination therapies (e.g. ivermectin-doxycycline) if the adjunct was not present in the control, is to exclude rationally-designed combination therapies (e.g. ZIVERDOX\textsuperscript{36}, I-MASK\textsuperscript{31}, or SMDT\textsuperscript{37,38}) that are in reality the norm among clinicians using ivermectin with success.

The pragmatic approach in Bryant\textsuperscript{3}, in contrast, stated a clear comparison of “ivermectin” against “no ivermectin” and makes no judgments about the efficacy of any other drugs. In regard to potentially-active comparators, reflection should show that any bias is conservatively against ivermectin; efficacy will be understated, not overstated, with respect to controls (Figure 1). Unsuspected active agents in ivermectin combination therapies might of course contribute bias the other way, but could only bias a meta-analysis toward an illusory conclusion of ivermectin efficacy if the same adjunct were dominant among the studies or patients in the meta-analysis. The only conflict then requiring resolution would be between an effective “Adjunct X” (ivermectin presumed ineffective) and synergy between the two. Either way, an effective therapy would have been demonstrated.

The fruit-salad analogy may in fact be helpful. If the outcome were the cure of scurvy, stronger results would be expected comparing a diet of oranges to boiled rice, than if the comparator were a fruit salad of unspecified composition. Only if all comparators were equally active would meta-analyses fail to demonstrate the value of oranges for scurvy. Any bias would understate, not overstate, their value. Similarly, if the diet of oranges were in some studies combined with grapefruit, eighteenth-century systematic reviewers might argue that the effect of grapefruit was not discriminated, but only if all studies included
grapefruit could meta-analysis fail to identify the efficacy of oranges, pointing only to the combination.

Specifically in Bryant\textsuperscript{3}, the only candidate adjunct (cited in Popp\textsuperscript{1} as grounds for exclusion) is doxycycline, which contributed three included studies\textsuperscript{34,50,51} coming nowhere near dominance in the Bryant\textsuperscript{3} meta-analysis.

Popp\textsuperscript{1} has a requirement for PCR or antigen testing to confirm a diagnosis, though without any examination of their reliability, and unavailable in many of the earlier studies or in resource-limited situations. This criterion would have excluded the RECOVERY trial of dexamethasone\textsuperscript{39}, which nevertheless led to the adoption of corticosteroids as a SoC, and the resulting inclusion of dexamethasone as a “proven” comparator in the review design\textsuperscript{1}.

**Fragmentation**

The inclusion policies thus excluded much of the available trials data. Pre-defined (and essentially arbitrary) time-points for outcome measures (28-day mortality, infection within 14 days) resulted in further exclusions. High RoB studies were rejected for “primary” analyses. After multiple exclusions of available trial evidence, Popp\textsuperscript{1} performs a final fragmentation of the data by analysing inpatient and outpatient data as separate comparisons, though the patients had the same disease and hospitalisation criteria vary considerably according to local resource constraints.

The overall consequence in Popp\textsuperscript{1} is illustrated schematically in Figure 2. The number of trials and participants was minimal for each “analysis”. Actual meta-analyses are few; many comparisons reduce to a single study so that no meta-analysis is possible. This reduces their “systematic review”, though bulky, largely to a bibliography with synopsis of results.

Fragmentation by review design defeats the whole purpose of meta-analysis.
Popp\textsuperscript{1} claims to provide a “complete evidence profile” but excludes most of the available evidence, available trials not conforming to their own ideal specification.

They also exclude several important confounder-adjusted observational trials\textsuperscript{24,33,40}. There is little basis for such exclusion, as shown exhaustively in another Cochrane Review\textsuperscript{41}. Randomisation is no guarantee that unmeasured risk factors are balanced between treatment arms, especially in small trials, where the alleged superiority of randomisation may be illusory. Four trials\textsuperscript{7,42-44}, with only 32 – 72 participants each, were included in Popp\textsuperscript{1}, but reported no risk factors, and were too small for randomisation to be have been a reliable equaliser of confounders. Against these inclusions, exclusion of the early ICON study\textsuperscript{24} (with 196 participants even in its matched subset) is inconsistent with the goal of a “complete evidence profile”.

Though Bryant\textsuperscript{3} is aligned with Popp\textsuperscript{1} in a restriction to RCTs, Bryant\textsuperscript{3} does not claim to present a “complete evidence profile”, and indeed we would argue that it does not. The protocol choice in Bryant\textsuperscript{3} was governed by standards typically expected by HRAs, to show the strength of the evidence in favour of ivermectin, even in the face of a restriction to RCTs.

A “complete evidence profile” requires consideration of other evidence in the “evidence quality pyramid”, including confounder-adjusted observational evidence\textsuperscript{24,40}, the population-scale case studies\textsuperscript{16,23}, and the evidence of clinicians reporting successful interventions with ivermectin-based therapies\textsuperscript{36}. These are typically under-reported in the primary literature precisely because their experience is so often acquired in crisis situations\textsuperscript{45}.

Popp\textsuperscript{1} claims the virtue, but prevents its realisation by design.
**The mortality outcome**

We illustrate the inadequacies in Popp\(^1\) by reference to the “headline” inpatient mortality outcome. Does ivermectin save lives?

Results from Popp\(^1\) in Table 2 are compared with the findings of the landmark ICON study\(^24\), a confounder-matched retrospective study, those in Bryant\(^3\), and the WHO\(^46\). The Popp\(^1\) meta-analysis (to June 2021) recognises just two small inpatient trials (73 and 112 participants respectively). This is still smaller than even the matched cohort in ICON (196 patients, 280 overall) available a full year earlier, showing a clear mortality benefit, quoting \(p = 0.045\). Popp\(^1\) thus exhibits smaller and poorer-quality data for in-patient mortality than already available more than a year earlier. Even including the outpatient analysis in Popp\(^1\) (a single death in 422 patients, most from the widely-criticised\(^48,23\) Lopez-Medina study\(^49\)) the Popp\(^1\) evidence is still far smaller than the WHO guideline\(^46\) from 1419 participants, which though deprecating ivermectin nevertheless quotes a mortality Odds Ratio of 0.19 with narrow Confidence Intervals. Bryant\(^3\) covers 2038 patients (even after the removal of a disputed study\(^47\)).

Yet in the face of far more extensive evidence that ivermectin does indeed save lives, Popp\(^1\) contrives the plain language summary: “we don’t know whether ivermectin leads to more or fewer deaths”.

---

8
Summary

Defects in Popp¹ thus include:

1. unsupported assertions of adverse reactions to ivermectin, and the outdated claim that unsafe dosing would be needed to be effective;
2. a demand for PCR or antigen testing, without analysis of reliability and not universally available even in developed countries at the start of the pandemic;
3. contradictions in the exclusion criteria, including placebo and approved SoC comparators, but rejecting hydroxychloroquine, though held to be ineffective (and an approved SoC in some jurisdictions);
4. inclusion of “deemed active” comparators whilst excluding “potentially active” ones;
5. exclusion of combination therapies, though the norm among practising clinicians;
6. the rejection of other than RCTs when the objective is a “complete evidence profile”;
7. arbitrary time-points for outcome measures, excluding non-compliant trials;
8. fragmentation of data by location of care under varying hospitalisation criteria;
9. the resulting focus on a small fraction of the available clinical evidence, with most comparisons based on single studies with no meta-analysis possible;
10. a resulting inpatient mortality comparison with fewer patients than a June 2020 confounder-matched study;
11. no conclusion on the headline mortality outcome, when multiple lines of evidence from elsewhere (including the WHO⁴⁶) point to significant mortality advantage.

What has driven most of the exclusions in Popp¹ are the criteria excluding other than “approved” drugs, with time-point or testing specifications that existing data happens not to satisfy. Refusal to look inevitably finds little.
Conclusion

In designing any SR, one does need to set criteria so that studies are sufficiently homogeneous to make valid statistical comparisons. However, the criteria cannot be so narrow as to leave too few data, or data that cannot be generalised from. There is an inevitable tension between statistical purity, and retaining data and clinical relevance; an instance of the eternal struggle between simplicity and truth. By discarding information to preserve purity, we end up communicating untruth; conversely if we abandon all comparability requirements then our message may get lost in confusion. Systematic reviewers must ask themselves whether they have struck the right balance for a particular research question, having regard to the context within which the question has arisen.

Globally we need to identify cheap, safe and effective therapeutics, quickly, so that Covid-19 patients are not left without prompt treatment. Using repurposed medicines with established safety profiles is a pragmatic public health strategy. “We call on public health authorities to authorize treatments with known low-risk and possible benefit for outpatients”.

The approach in Popp, demanding compliance with design criteria that few existing studies satisfy, offers no insight into important research questions in a health emergency requiring rapid decision-making. Though considerable evidence exists, heterogeneity results in most of it being rejected. Popp should (i) reconsider their balance between statistical purity and delivering clinical relevance, (ii) consider the impact of their selection criteria on their conclusions and (iii) re-consider those criteria (and their search strings) prior to running the planned future updates.

We affirm the pragmatic approach in Bryant, which was designed precisely to expose key clinical effects from existing trials data, having regard to the manifest urgency. Sensitivity analyses already show that the headline mortality and prophylaxis effects are sustained, and a Bayesian analysis allowing causal hypotheses to be ranked on a probability basis shows that the hypothesis of mortality benefit remains supported with high probability even after the exclusion of two disputed trials.
Though castigated for “pooling heterogenous populations comparators and outcomes” this is to a degree required in any meta-analysis, and the effects of (i) potentially active comparators and (ii) adjuncts, are readily assessed by simple thought, as outlined above.

Popp\textsuperscript{1} adds nothing to the evidence on ivermectin in Covid-19, and indeed deems most of it not to exist, promoting an almost counter-factual impression. Meanwhile, a global health emergency continues, and lives continue to be lost, economies and social fabric are torn apart, whilst a WHO Essential Medicine\textsuperscript{6} of unparalleled safety profile\textsuperscript{8-13} is withheld by many HRAs, merely because of debates over its quantitative metrics of effect.

The responsible physician knows that “first do no harm” does not mean: “do nothing”. Doing nothing, especially for the high-risk patient, is unacceptable. It is lethal folly to withhold a safe\textsuperscript{8-13} medicine that, when viewing the totality of the evidence\textsuperscript{55}, shows great promise at negligible cost. Such policies effectively deny to dedicated clinicians the ability to practice medicine. It is high time those policies were ended.
References


### Table 1

Summary of the primary analyses in Popp\(^1\), with numbers of patients and events shown as a rough indicator of statistical power. Only those analyses marked had more than one study contributing to the comparison. Mortality was considered the “primary analysis” in the prophylaxis question though reporting no deaths; under this heading the more meaningful “infection” outcome is also shown, though considered a “secondary” analysis. Analyses considered secondary in Popp are shown in italics. There were 9 further secondary analyses not shown. In all these, the studies, patients and events are similarly sparse.

<table>
<thead>
<tr>
<th>inpatients</th>
<th>outpatients</th>
<th>prophylaxis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality (2 studies)</td>
<td>Mortality (2 studies)</td>
<td>Mortality</td>
</tr>
<tr>
<td>185 patients; 14 events</td>
<td>422 patients; 1 event</td>
<td>304 patients; 0 events</td>
</tr>
<tr>
<td>Need for mechanical ventilation</td>
<td>Need for mechanical ventilation</td>
<td>Symptomatic Covid-19</td>
</tr>
<tr>
<td>185 patients; 12 events</td>
<td>398 patients; 1 event</td>
<td>304 patients; 74 events</td>
</tr>
<tr>
<td>Need for oxygen</td>
<td>Need for non-invasive support</td>
<td></td>
</tr>
<tr>
<td>45 patients; 0 events</td>
<td>398 patients; 0 events</td>
<td></td>
</tr>
<tr>
<td>Improvement to discharge</td>
<td>Improvement to resolution</td>
<td></td>
</tr>
<tr>
<td>73 patients; 54 events</td>
<td>398 patients; 246 events</td>
<td></td>
</tr>
<tr>
<td>Adverse events</td>
<td>Adverse events (2 studies)</td>
<td>Adverse events</td>
</tr>
<tr>
<td>152 patients; 20 events</td>
<td>422 patients; 325 events</td>
<td>304 patients; 11 events</td>
</tr>
<tr>
<td>Serious adverse events (2 studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>207 patients; 1 event</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admission to ICU (2 studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>143 patients; 13 events</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration of hospital stay</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45 patients</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Viral clearance at 3, 7, 14 days (2 studies)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>170 patients; 23, 59 &amp; 26 events</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. (Schematic only). The likely effect of including studies with unsuspected or uncertain active controls in a meta-analysis. There would be a conservative bias against the intervention tested.
**Figure 2.** Overall effect of the exclusion criteria in Popp¹. The majority of the available evidence is rejected because of non-compliance with a preferred “ideal” trial design.

The data left over for the leading mortality analyses are shown.
<table>
<thead>
<tr>
<th>Study</th>
<th>Date available</th>
<th>patients</th>
<th>events</th>
<th>studies</th>
<th>Risk Ratio (OR where stated) [95% CIs]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popp(^1) (1.1.1, moderate)(^a)</td>
<td>July 2021</td>
<td>185</td>
<td>14</td>
<td>2</td>
<td>0.60 [0.14, 2.51]</td>
</tr>
<tr>
<td>ICON(^24) (inpatients, matched subset)(^b)</td>
<td>June 2020</td>
<td>196</td>
<td>37</td>
<td>1</td>
<td>0.54 [0.29, 1.00]</td>
</tr>
<tr>
<td>WHO(^46, c) (Forest plot, Risk Ratio)</td>
<td>March 2021</td>
<td>915</td>
<td>31</td>
<td>5</td>
<td>0.36 [0.17, 0.75]</td>
</tr>
<tr>
<td>WHO(^46, d) (Summary, Odds Ratio)</td>
<td>March 2021</td>
<td>1419</td>
<td>n/a</td>
<td>7</td>
<td>0.19 {OR} [0.09, 0.36]</td>
</tr>
<tr>
<td>Bryant(^3,53) (w/o Elgazzar)(^e)</td>
<td>July-Sept 2021</td>
<td>2038</td>
<td>93</td>
<td>15</td>
<td>0.51 [0.27, 0.95]</td>
</tr>
</tbody>
</table>

**Table 2.** Abstract of mortality analyses from Popp\(^1\), the ICON\(^24\) study, WHO\(^46\), and Bryant\(^53\).

\(^a\)Popp\(^1\) Risk Ratio (RR) quoted in source.

\(^b\)ICON\(^24\) total study size 280 patients; quoted figures for confounder-adjusted cohort only. RR re-calculated from source data (OR reported in source, \(p=0.045\)). “Severe” patient subset is smaller (25 patients) but the mortality advantage yet higher, quoting \(p<0.002\).

\(^c\)WHO\(^46\) RR from Forest plot in “Living Guideline”, page 18

\(^d\)WHO\(^46\) OR from summary in “Living Guideline”, page 19. Not directly convertible to RR because of insufficient data. All disease severities.

\(^e\)Bryant\(^53\) data excludes Elgazzar\(^47\). Subgrouped by severity.