Summary of methodological issues in epidemiology

1st November 2020

Mike Hearn <mike@plan99.net>

Abstract. Problematic practices within epidemiology are presented, along with suggestions for improvement.

Lack of public review. The Imperial College London Report 9 paper that largely drove UK public policy contained internally inconsistent/non-replicable numbers[1], didn't use data from arguably the best datasets then available that indicated a 40% lower fatality rate[2], and relied on unpublished model code that only its author understood[3]. These problems were caught after the work had already altered government policy. Whilst many researchers have embraced open access, preprints and public code/data, these practices are not a requirement for research relied on by the civil service. When external review from outside the field did occur it was rejected with the justification that cross-discipline review is inherently illegitimate[4].

Poor characterisation of statistical uncertainty. Policy was driven by modelling that used insufficiently large data sets to derive critical inputs[5] and uncertainty bounds were either not reported at all[6] or had extremely wide ranges[7]. Uncertainty ranges were sometimes widened post-publication, e.g. days after the release of ICL Report 9 the lead author altered his prediction to be "could be 20,000 deaths *or much lower*"[8], thus rendering the predictions unfalsifiable in one direction and adding a wide uncertainty bound post-facto.

Non-existent or circular model validation. Validation of epidemiological models is rare. Some scientists have argued that few healthcare models can ever be validated against reality, yet they should still be used to make decisions[9]. The COVID model produced by Imperial College London is derived from a flu model first published in 2005[10]. Despite many outbreaks of seasonal influenza having occurred since then, no evidence was provided in Report 9 or its citations showing that the model accurately predicts epidemics. Models are frequently considered validated if their predictions match the results of other models[11,12] rather than the actual course of an epidemic. This is invalid because testing predictions against themselves is circular reasoning.

Research papers may pre-suppose their own conclusions, for example, *Nature* published a modelling paper from ICL (Flaxman et al) which claimed lockdowns had saved 3.1 million lives[13]. In fact it used circular logic by pre-allocating all the reductions in R to government intervention (NPI)[14] and encoded the output conclusion in the input parameters via statistical forcing and parameter choice[15]. A related issue is how the reliability of COVID PCR testing is determined by calibrating the test against itself[16]. Peer review appears to only rarely prevent these kinds of problems.

Particularly concerning is the use in some papers of subjective Bayesian priors, which encode the scientist's pre-existing intuitive beliefs about the likelihood of certain answers as inputs. As the result of science is itself evidence used to update those intuitive beliefs, this is another form of circular reasoning.

No code quality processes. Standard epidemiological practice is to peer review the intended assumptions and conclusions of a model, but not the implementation[17]. There are no academic processes that recognise the possibility of implementation error. Despite 15 years of continuous development the code behind Report 9 was only made public in 2020 after public pressure and FOIA requests. Once public review was possible bugs were found in its code that impacted its predictions[18], for example it was found that predictions depended on arbitrary factors like what kind of computer was used to run it[19], that it contained data corruption bugs[20,21,22], and that predictions of bed demand changed between versions by more than the size of the Nightingale emergency hospital deployment[18]. No standard regression test system was in place. Although professional software engineers were brought in to work on the code, this occurred only after it had already altered government policy. The British Computing Society criticised the lack of code quality processes in academic modelling[23].

Misleading press statements. In their paper Flaxman et al stated that the claim of 3.1 million lives saved was "*illustrative only*", and that "*in reality even in the absence of government interventions we would expect Rt to decrease and therefore [we] would overestimate deaths in the no-intervention model*"[13]. But to the press Flaxman said, "*Lockdown averted millions of deaths, those deaths would have been a tragedy*"[24]. After concerns were raised by software engineers that the ICL COVID-Sim model did not repeatedly generate the same predictions, ICL published a press release[25] in which a third party researcher stated "*I was able to reproduce the results… from Report 9*". Nature claimed "*it dispels some misapprehensions about the code, and shows that others can repeat the original findings"*[26]. Models generate predictions, not findings. In fact every prediction he got out of the model was different, three of them

showing "*significant differences*" of 10-25%[27]. The press release also stated that Report 9 was built "*on code originally developed, published and peer-reviewed in 2005 and 2006*", although the code had never been published or externally/peer reviewed until 2020.

Excessive freedom in choosing input data. Researchers may freely select data and add assumptions without regard to quality. *The Lancet* published a modelling paper in August[28] that used fatality rate data gathered in January[29], likewise for a paper modelling the impacts of contact tracing[11], although observed CFRs at that time ranged between 2.8% (higher than the Spanish Flu) and 0.18%[30]. It was already known since 2012 that it can take several months of observation for fatality ratios to become accurate enough to be usable[5]. More recent data would have lowered predicted deaths significantly. The Lancet paper also claimed "*the data are sparse*" using a citation from March, although a month earlier in July a literature review by doctors stated the opposite[31]. The ICL COVID-Sim model has over 200 user-specifiable parameters, many of which appear to be guesses[32]. As an example it assumed individuals hardly vary in their chances of catching COVID; the projected number of infections is far lower if the assumption is modified for non-uniform susceptibility[33].

Lack of cost/benefit analysis. The quality adjusted life year (QALY) is a standard metric used for analysis of healthcare interventions in the NHS[34]. NICE suggests a limit of about £20,000 - £30,000 spent per QALY gained[35]. However, QALY analysis in academic output is rare - none of the papers discussed in this report uses it. Although non-pharmaceutical interventions were a topic of the original ICL paper from 2005[10], modelling efforts then and since appear uninterested in the question of whether they are cost effective. Nor are physical and mental health losses caused by NPI accounted for. One paper with "*Modelling the health and economic impacts of … strategies for COVID-19*" in the title declined to do a cost/benefit analysis, because the idea of a tradeoff between GDP and health outcomes would be contested[36]. Yet cost/benefit analysis is routine for pharmaceutical interventions and is especially critical for COVID-19 due to the high rate of comorbidities, high average age of the victims and high cost of lockdowns.

Silencing of disagreement. A model that calculated lower herd immunity thresholds (i.e. a quicker end to the epidemic) was rejected for publication because if people felt less at risk, government intervention might be reduced[37]. The journal *Science* considered rejecting a similar paper for similar reasons[38]. Journals have refused to publish a large-scale field study of whether masks are effective[39]; the author said it would be published "*as soon as a journal is brave enough*"[40]. A Nobel prize winner in biophysics was barred from speaking at an academic conference due to his anti-lockdown views[41]. A member of SAGE obtained pre-agreement from BBC Radio 4 that a debate between her and an opposing epidemiologist would be rigged[42]. A professor of epidemiology at Stanford had a paper rejected on the basis that "*no infectious disease expert thinks this way*"[43].

Suggestions for improvement

Although this paper focuses on epidemiology, questionable research practices are widespread across many academic fields which inform public policy[44]. The following suggestions are therefore neutral with respect to field of study:

1. Before research is presented to ministers or the civil service it should be pre-vetted by a new Office of Research Integrity, that:
    a. Seeks out disagreement both within and outside the academic community. Commission Tenth Man[45] / red team reports from those people so they can make their case directly to the government.
    b. Is trained in how to critically review research papers using in-house statistical expertise, under time pressure. Papers found to be using obsolete data, containing logical fallacies, questionable causative and/or statistical models, or insufficiently supported or biased assumptions, should not be approved for use.
    c. Requires evidence of model validation against reality. Validation studies should be performed by a third party outside the domain being validated (i.e. researchers in a field would not be allowed to validate for government use research produced by researchers in that same field)
    d. Has the power to disbar researchers from being on projects that receive public money in case of detected research fraud.

2. Code quality controls:
    a. Publishing anything about a model requires publishing at the same time all code and data utilised, with a clear explanation of all assumptions made. Exceptions for datasets licensed from commercial organisations (universities may not sub-license data they collected to get around this requirement).
    b. Pre-registration of modelling efforts prior to publication, in which commitments to software engineering practices are made, e.g.
        i. Minimum levels of unit test coverage (recommendation: >= 80%)
        ii. Internal peer review of code changes
        iii. Use of memory safe languages
    c. Hiring or contracting of qualified software engineers to implement or review model code. In case of hiring for review, the comments and consequent changes must be co-published with the code itself.

3. All modelling used to argue for or against specific policies must demonstrate rigorous cost benefit analysis, backed by data collected outside the domain being studied (i.e. researchers in a field may not provide their own *de novo* figures for costs or benefits).

4. Prediction markets have proven successful at predicting which papers will successfully replicate. Similar markets may prove beneficial for estimating the accuracy of forecasts. The field of superforecasting may also have insight to contribute.

Contributors

The author is indebted to Nicholas Lewis and Harrison Comfort for their careful review and analysis.

Mike Hearn has been programming computers since 1990. Between 2006-2014 he worked at Google as a senior software engineer on Maps, Gmail and account security. Since then he has been developing database and encryption technology, primarily for the finance and trade/shipping sectors. He has no connection with academia or the field of epidemiology.

References

1. Imperial College UK COVID-19 numbers don't seem to add up, Nicholas Lewis, 2020

2. Report 9 used infection fatality rate estimates (adjusting them up by 17%) from another ICL study, Verity et al., which were based exclusively on Chinese data despite the authors also analysing the Diamond Princess data and obtaining a 39% lower best estimate. See also COVID-19: Data from the Diamond Princess cruise ship implies that UK modelling hugely overestimates the expected death rates from infection, Nicholas Lewis, 2020

3. "*Yet for other scientists the big problem with Ferguson's model is that they cannot tell how it works. It consists of several thousand lines of dense computer code, with no description of which bits of code do what. Ferguson agreed this is a problem. "For me the code is not a mess, but it's all in my head, completely undocumented. Nobody would be able to use it . . . and I don't have the bandwidth to support individual users." He plans to put that right by working with Carmack and others to publish the entire program as an interactive website*", Neil Ferguson interview: No 10's infection guru recruits game developers to build coronavirus pandemic model, Sunday Times, March 2020.

4. "*Epidemiology is not a branch of computer science and the conclusions around lockdown rely not on any mathematical model but on the scientific consensus*", statement by Imperial College London, Coding that led to lockdown was 'totally unreliable' and a 'buggy mess', say experts, Daily Telegraph, May 2020.

5. The Time Required to Estimate the Case Fatality Ratio of Influenza Using Only the Tip of an Iceberg: Joint Estimation of the Virulence and the Transmission Potential, Computational and Mathematical Methods in

Medicine, 2012

6. ICL Report 9 proposed several scenarios given varying values of R0 but did not present any uncertainty intervals for its predictions, although the underlying model was both intentionally stochastic and unintentionally computationally unstable.

7. "*Last week IHME projected that Covid-19 deaths in the U.S. would total about 60,000 by August 4; this week that was revised to 68,000, with 95% certainty that the actual toll would be between 30,188 and 175,965*", Influential Covid-19 model uses flawed methods and shouldn't guide U.S. policies, critics say, STAT News, April 2020

8. UK has enough intensive care units for coronavirus, expert predicts, New Scientist, March 2020

9. "*After reviewing the conditions required for predictive validation, such as constancy of the situation over time and across variations of conditions not specified in the model as well as availability of sufficient data to make predictive tests, [Weinstein et al] concluded that few models in healthcare could ever be validated for predictive use. This, however, does not disqualify such models from being used as aids to decision making. Philips et al state that since a decision-analytic model is an aid to decision making at a particular point in time, there is no empirical test of predictive validity. From a similar premise, Sculpher et al argue that prediction is not an appropriate test of validity for such models*", Validation of population-based disease simulation models: a review of concepts and methods, BMC Public Health, 2010

10. Strategies for containing an emerging influenza pandemic in Southeast Asia, Nature, 2005

11. "*We believe in checking models against each other, as it's the best way to understand which models work best in what circumstances*", documentation for a COVID model used in the paper "Modelling the health and economic impacts of Population-wide Testing, contact Tracing and Isolation (PTTI) strategies for COVID-19 in the UK", preprint at SSRN, 2020

12. "*There is agreement in the literature that comparing the results of different models provides important evidence of validity and increases model credibility*", Validation of population-based disease simulation models: a review of concepts and methods, BMC Public Health, 2010

13. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe, Nature, 2020

14. "*They estimate R(t) from daily deaths associated with SARS-CoV-2 using as an a priori restriction that R(t) may only change at those dates where interventions become effective. Such an approach does not prove that NPIs were effective but rather begs the result, i.e., involves circular logic*", Comment on Flaxman et al: The illusory effects of non-pharmaceutical interventions on COVID-19 in Europe, preprint at Advance, Homburg & Kuhbandner, 2020 (doi: 10.31124/advance.12479987.v1)

15. Did lockdowns really save 3 million COVID-19 deaths, as Flaxman et al. claim?, Nicholas Lewis, 2020.

16. *"the test being evaluated forms part of the reference standard, and this would tend to inflate the measured sensitivity of these tests*", Interpreting a COVID-19 test result, BMJ, 2020

17. "*Evaluation of the source code by external experts is rare because of intellectual property concerns [11,15]. This practice, however, adds to the impression of simulation models as black boxes.*", Validation of population-based disease simulation models: a review of concepts and methods, BMC Public Health, 2010

18. Second analysis of Ferguson's model, Mike Hearn (writing as Sue Denim), 2020

19. COVID-Sim Issue #30, GitHub, 2020

20. Fix a bug where variables weren't getting initialised to 0, COVID-Sim change facc5127b35b71ab9b6208961e08138f56448643

21. This fixes an uninitialised memory usage when using schools as input, COVID-Sim change 581ca0d8a12cddbd106a580beb9f5e56dbf3e94f

22. Fix issue causing invalid output from age distribution code, COVID-Sim change 3d4e9a4ee633764ce927aecfbbcaa7091f3c1b98

23. "*Computer code used to model the spread of diseases including coronavirus must meet professional standards … the quality of the software implementations of scientific models appear to rely too much on the individual coding practices of the scientists who develop them*", Computer modelling of epidemics must meet 'professional standards', says industry group, Daily Telegraph, May 2020.

24. Coronavirus: Lockdowns in Europe saved millions of lives, BBC News, 2020

25. Codecheck confirms reproducibility of COVID-19 model results, Imperial College London press release, 2020

26. Critiqued coronavirus simulation gets thumbs up from code-checking efforts, Nature, 2020

27. CODECHECK certificate 2020-010, https://doi.org/10.5281/zenodo.3865491

28. Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second COVID-19 epidemic wave in the UK: a modelling study, The Lancet, August 2020

29. Estimates of the severity of COVID-19 disease, Verity et al, preprint, published March 2020, calculating infection:fatality ratio based on PCR testing done in January.

30. 2019-Novel Coronavirus (2019-nCoV): estimating the case fatality rate – a word of caution, Swiss Medical Weekly, February 2020

31. "*After six months, we have a wealth of accumulating data showing that children are less likely to become infected and seem less infectious; it is congregating adults who aren't following safety protocols who are responsible for driving the upward curve*", Children rarely transmit COVID-19, Science News, July 2020.

32. e.g. that symptomatic people meet exactly half the number of random contacts as a healthy person, that asymptomatic people are 50% as infectious as symptomatic, or that the variation in individual infectiousness (overdispersion/"k") was exactly 25%.

33. Why herd immunity to COVID-19 is reached much earlier than thought, Nicholas Lewis, 2020

34. Judging whether public health interventions offer value for money, NICE, 2013

35. ibid

36. "*Our results are presented as a disaggregated impact inventory rather than a cost-benefit analysis given that the latter would require a measure of an appropriate rate of trade-off between reductions in deaths and reduced GDP, which are available but likely to be contested*", "Modelling the health and economic impacts of Population-wide Testing, contact Tracing and Isolation (PTTI) strategies for COVID-19 in the UK", preprint at SSRN, 2020

37. "Given the implications for public health, it is appropriate to hold claims around the herd immunity threshold to a very high evidence bar, as these would be interpreted to justify relaxation of interventions, potentially placing people at risk", rejection reason as stated by biomathematician M Gabriela Gomez, August 2020

38. "*The relevant Science editors discussed whether it was in the public interest to publish the findings … we were concerned that forces that want to downplay the severity of the pandemic as well as the need for social distancing would seize on the results to suggest that the situation was less urgent*", Modeling herd immunity,

Science, Thorp, Vinson & Ash, June 2020

39. Virker mundbind? Toptidsskrifter afviser at trykke det danske svar ("Do masks work? Top journals refuse to print the Danish answer"), Berlingske, October 2020

40. Personal correspondence with lead scientist, published on Twitter, Alex Berenson, October 2020

41. *"... too many calls by other speakers threatening to quit if you were there. They all complained about your COVID claims"*, personal correspondence published on Twitter, Michael Levitt, October 2020

42. "*I'd got prior agreement from R4 about the framing of the item. I was assured that this would not be held as an even-handed debate given the small number of 'herd immunity strategy' advocates, but sadly this wasn't followed through*", Prof Susan Michie, SAGE member, Twitter, September 2020

43. *"I made extensive revisions, then they rejected it apparently because an expert reviewer told them no infectious disease expert thinks this way – paradoxically, I am trained and certified in infectious diseases.",* Forecasting for COVID-19 has failed, Ioannidis, Cripps & Tanner, International Institute of Forecasters, August 2020.

44. What's Wrong with Social Science and How to Fix It: Reflections After Reading 2578 Papers, Alvaro de Menard, September 2020

45. From the Israeli military intelligence practice of using professional devil's advocates. "*The Tenth Man is a devil's advocate. If there are 10 people in a room and nine agree, the role of the tenth is to disagree and point out flaws in whatever decision the group has reached*", How Israeli intelligence failures led to a 'devil's advocate' role, Why Dissent Matters, 2017