UnderreportedAndTemporallyAggregated

This README file contains the currentw ork and outline for our project for inferring instantaneous reproduction number that accounts for temporally aggregated and under-reported incidence data.

Overview

Our project is broken down into 3 main sections

1. Demonstrate that inference works for typical value of under-reporting (0.4) for a large number of epidemics (1,000).
2. Fixing the true incidence, and considering different reporting rates to generate the reported incidence. Investigate R inference and see how results vary for different values of under-reporting.
3. Finally using a real world data-set from an Ebola outbreak, investigate Rt inference for different values of under-reporting.

The message from each section should take roughly the following form:

1. Inference works for a wide range of epidemics, with higher accuracy and precision when incidence is higher

For sections 2 and 3, we expect there to be a nuance between two factors. In general, higher assumed true incidence will lead to higher levels of precision in reproduction number inference. But there may be some interaction between that and the value of under-reporting assumed.

##Checklist

- Generate temporally aggregated incidence with $R_t = 1.5$ and re-infer R_t for Ebola epidemic with 'stuttering start', that is to say an epidemic that starts with 1 case on day 1. Allow epidemic to grow until it exceeds 1000 cases, then switch to $R_t = 0.75$ for final 5 weeks. Only show inference for final 10 weeks. DONE.
- Repeat inference for this epidemic with different values of M $(10^3, 10^4, 10^5)$ and 30 different times for each M. Use this to assess the robustness/consistency of inference for each value of M. DONE.
- Run 1,000 epidemics and for each one generate 6 possible reported incidences (with $\rho = 0.33, 0.43, 0.53, 0.63, 0.73, 0.83$). Infer with correct knowledge of reporting rate. R_t is sampled from a Gamma(shape = 1, scale = 3) distribution, as this ensures incidence does not get too high or too low. We then use this distribution as our prior when performing the inference. DONE.
- Using these 6,000 inferences we can look at the coverage and the error. We look at the coverage over all values of ρ and look at the distribution of correct coverage over all simulations, as well as what the total coverage is. This total coverage should be close to 95% since we use 2.5 and 97.5 credible intervals. For the error, we look at the distribution and calculate the mean error.
- We may need to qualify the significance of this distribution and the mean error. I have tried to do this by repeating the inference again but using a Naive Epi-Estim approach.

Section 1. Checking incidence is accurate.

We look at two case studies. Firstly, we look at a realistic outbreak. Secondly, we simulate a large number of epidemics, where the true Rt values are sampled from the gamma distribution that informs our prior.

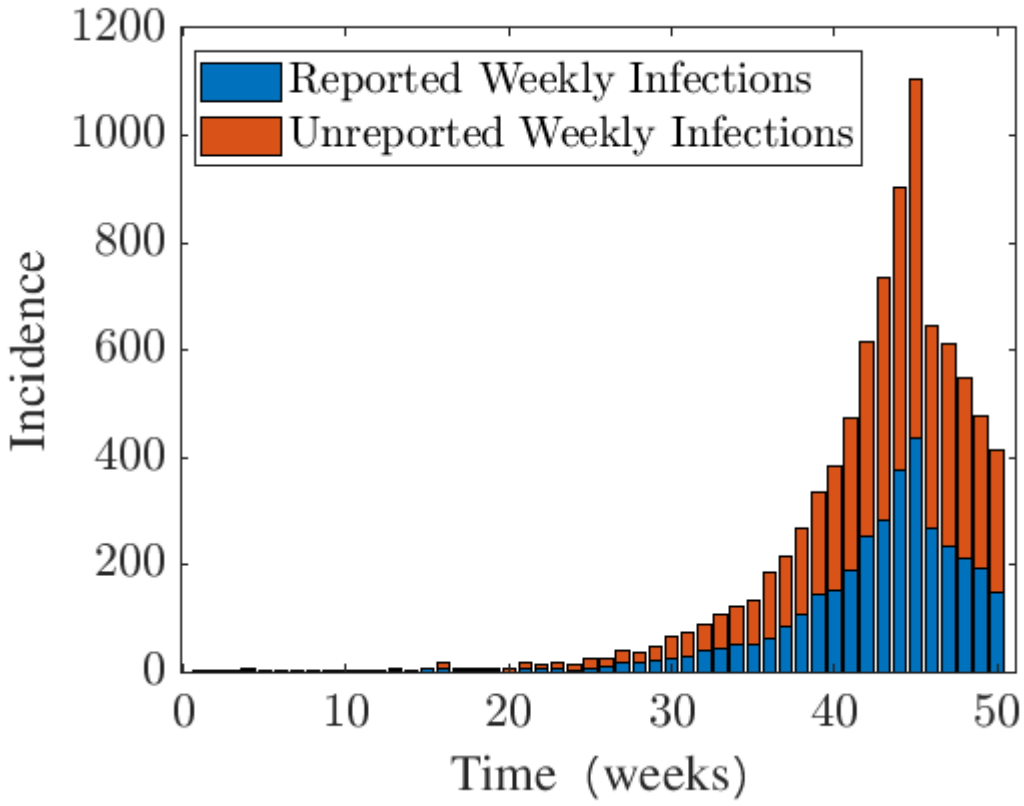
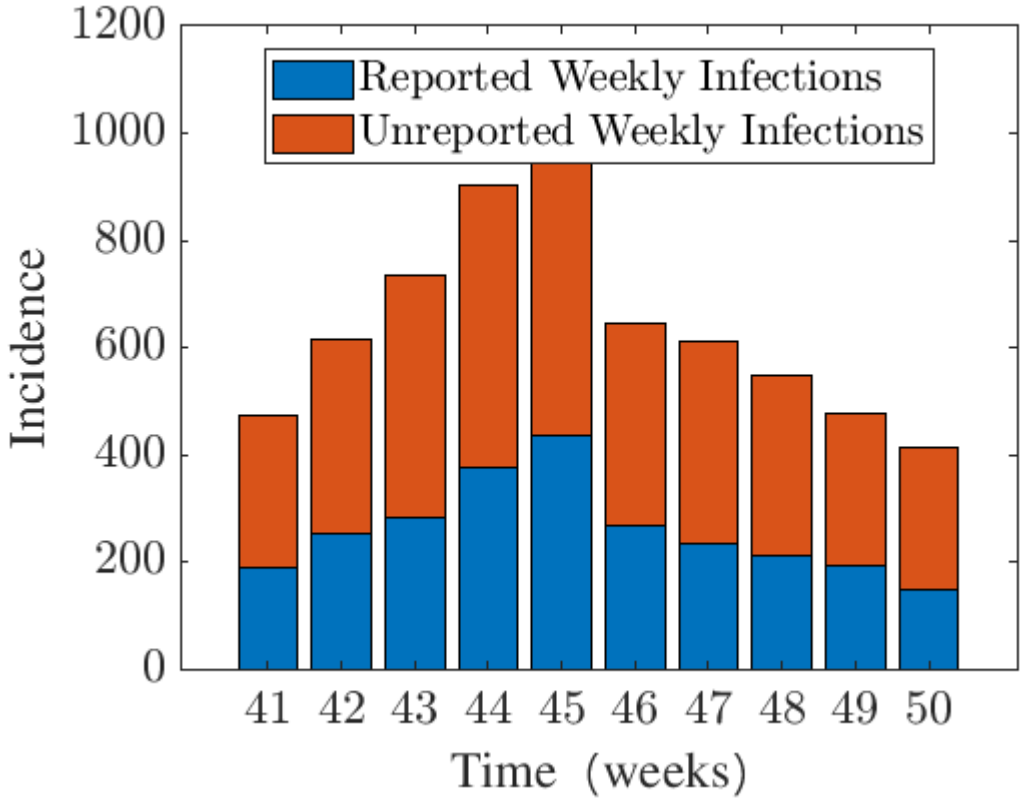


Fig 1: Example simulation with $\rho = 0.4$



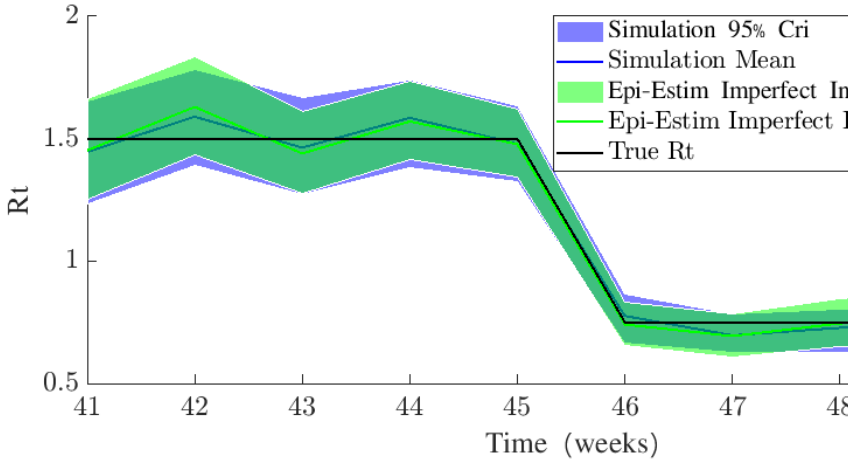


Fig 2: Same example simulation with $\rho = 0.4$

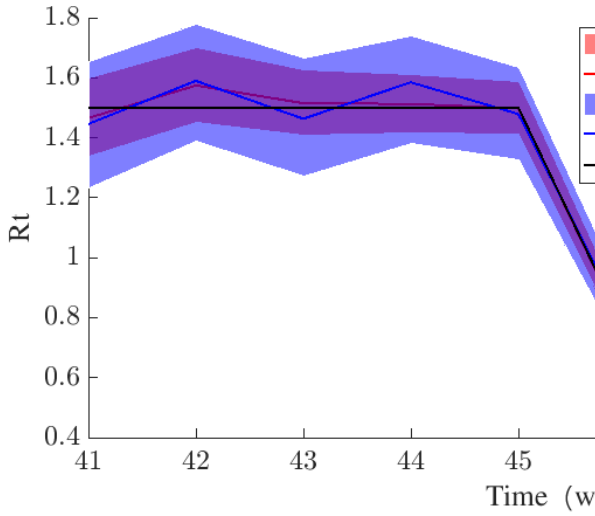


Fig 3: Comparison of inference for simulation method vs Epi-Estim
Fig 4: Comparison of inference for simulation method vs Epi-Estim (with perfect information)

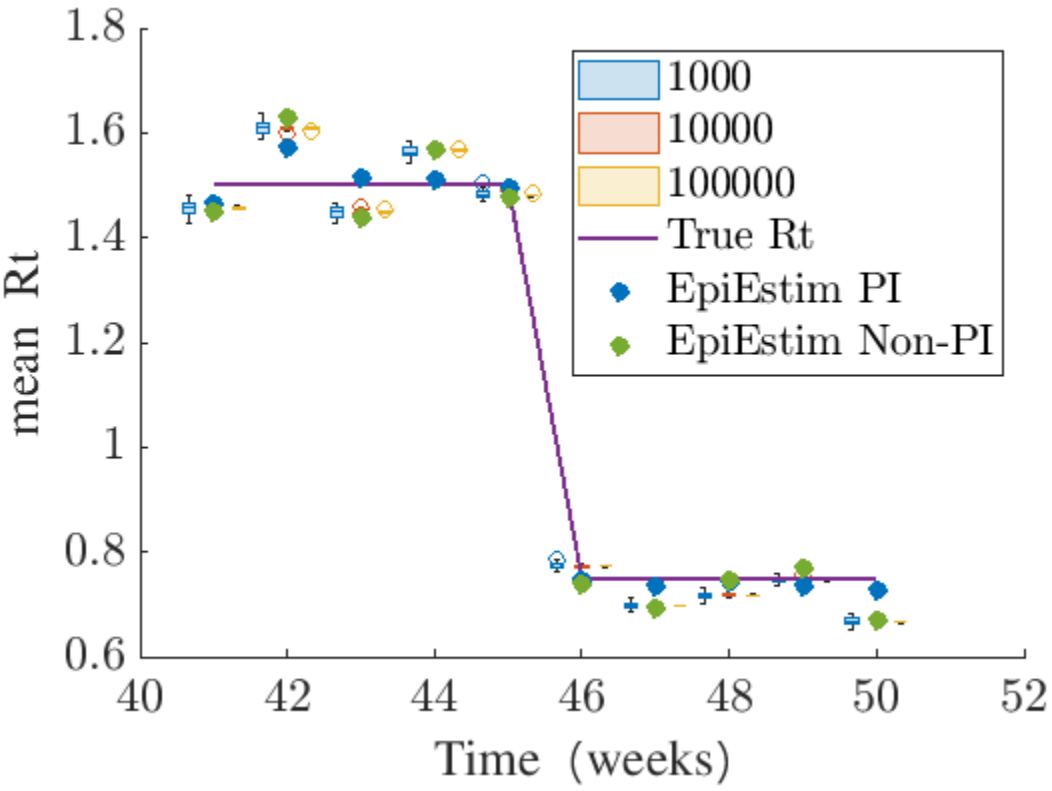
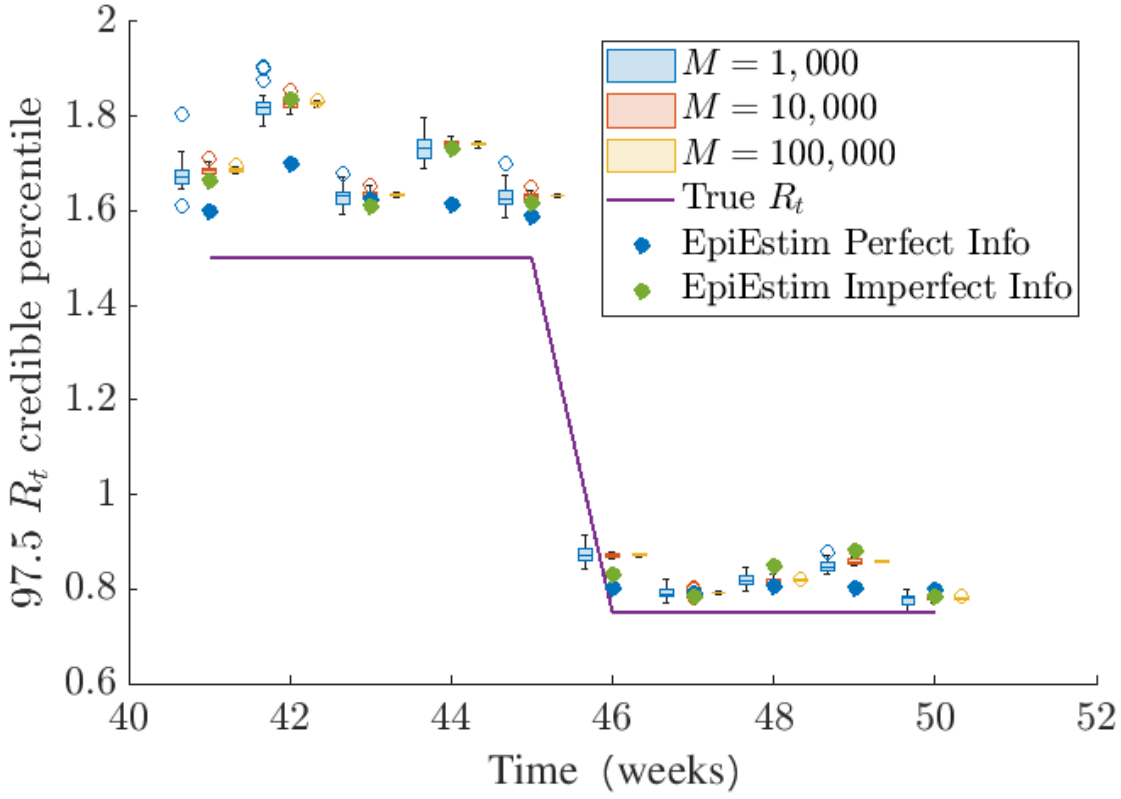
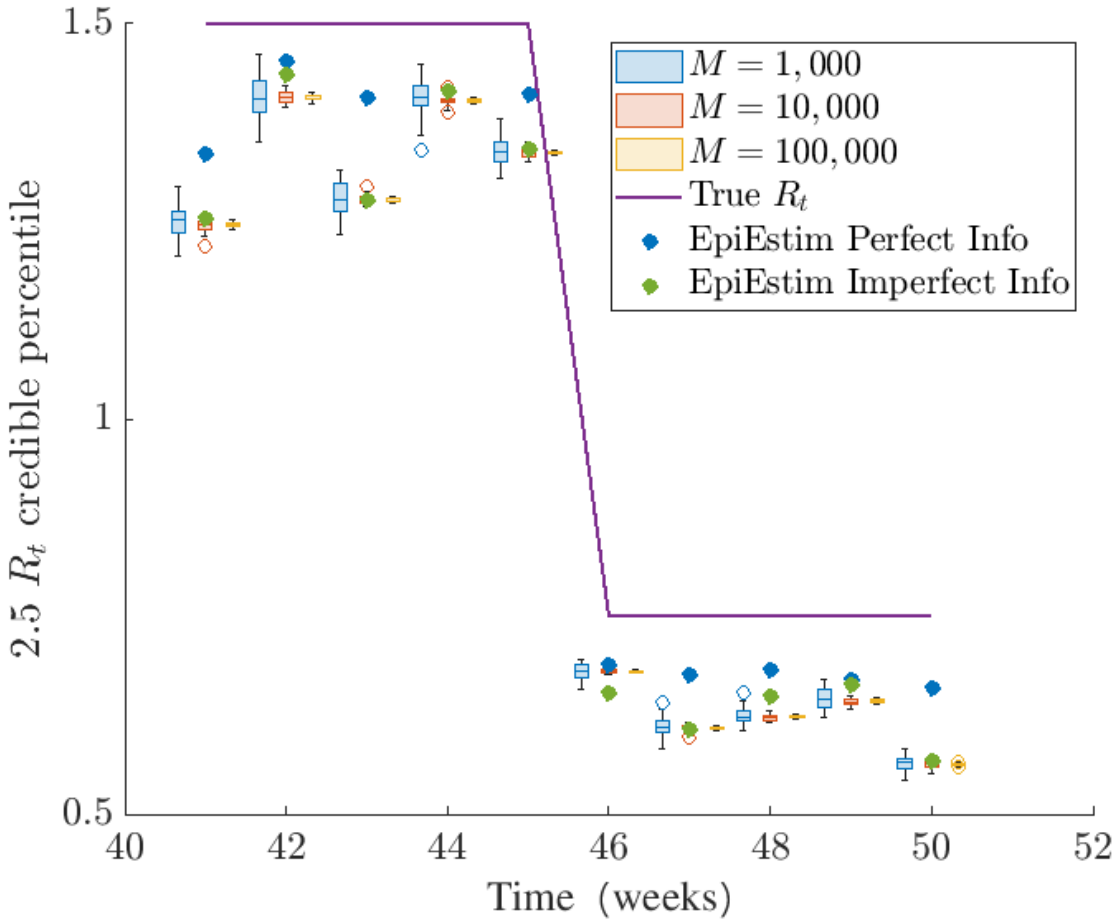


Fig 5: Robustness check 1. Inference of same epidemic 30 different times for different values of M (and EpiEstim with perfect and imperfect information) Message for Fig 5: As M increases, the mean estimate becomes more



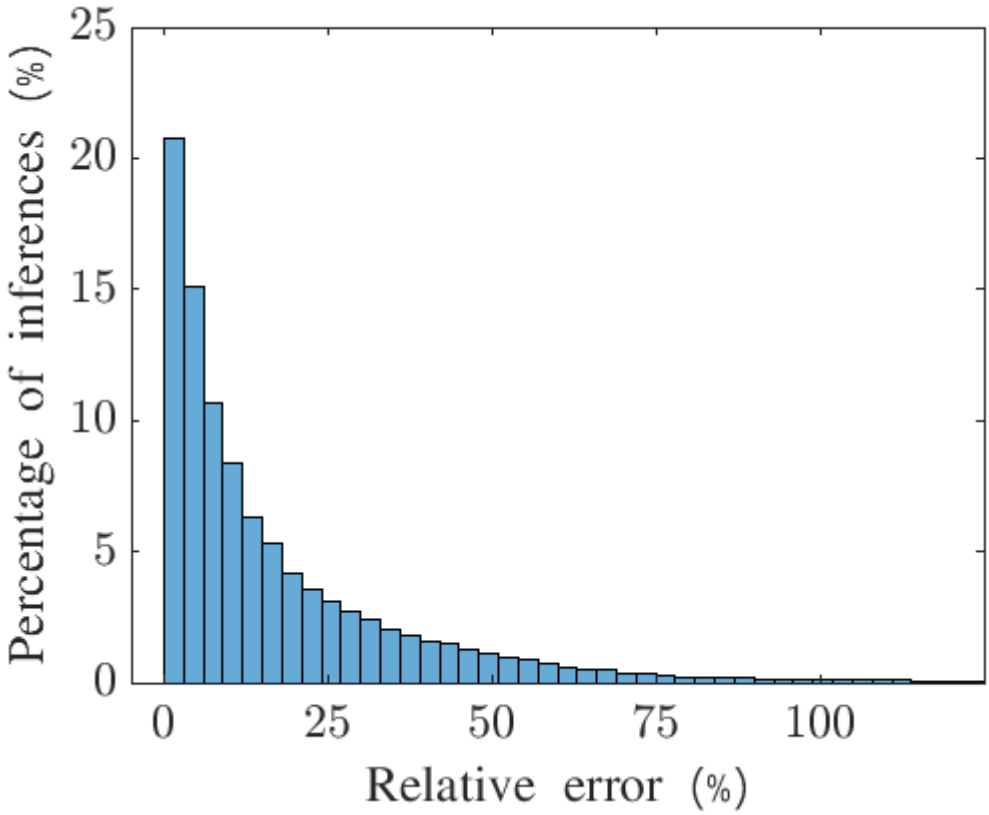
robust.

Fig 6: Robustness check 2. Inference of same epidemic 30 different times for different values of M (and EpiEstim with perfect and imperfect information) Message for Fig 6: As M increases, the upper percentile estimate



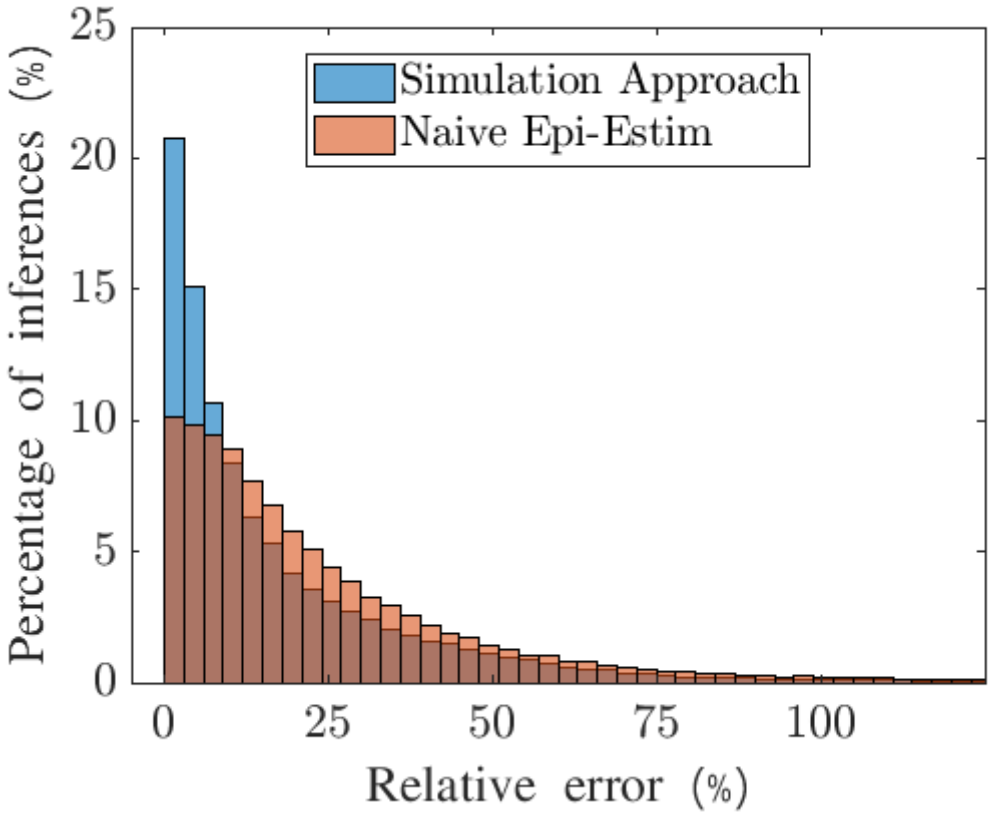
becomes more robust.

Fig 7: Robustness check 3. Inference of same epidemic 30 different times for different values of M (and EpiEstim with perfect and imperfect information) Message for Fig 7: As M increases, the lower percentile estimate



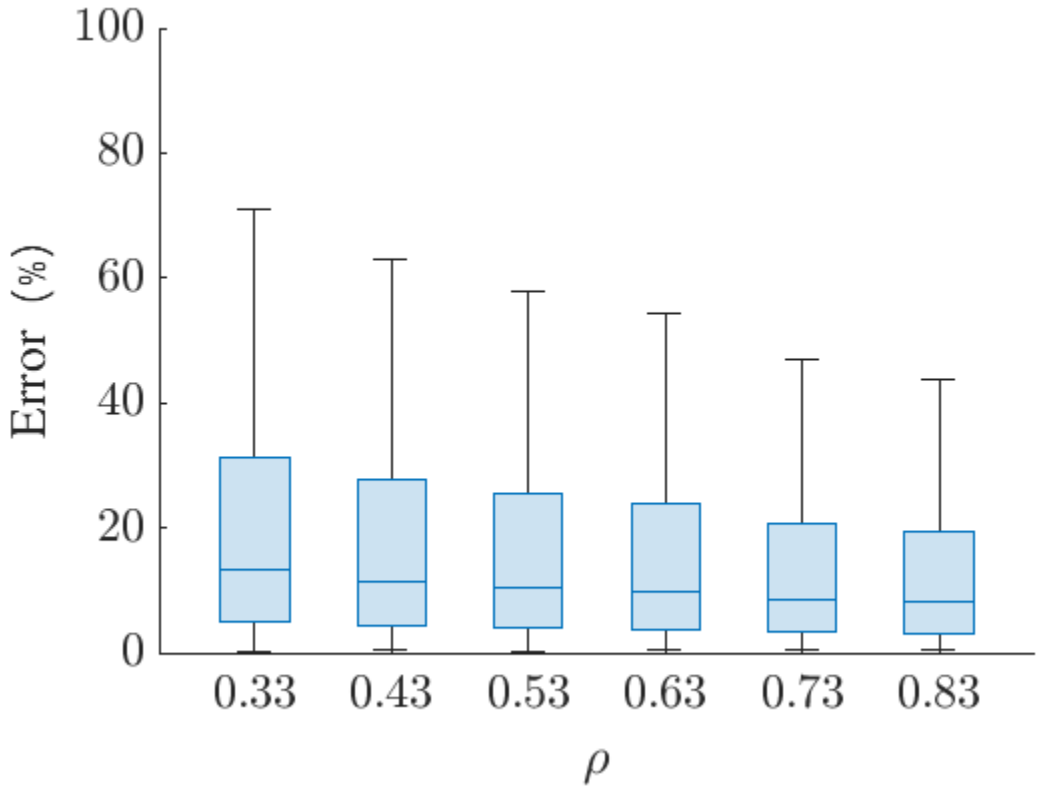
becomes more robust.

Fig 8: Relative error distribution over 6000 simulations Message for Fig 8: Over a wide range of epidemics and a broad spectrum of reporting rates, we find that the relative error distribution to take the following form, with mean value 19.8%. NB: Do we need to qualify this somehow? Is 19.8% good or not? We can also investigate whether there is systematic over-estimation/under-estimation but this will probably be an artifact of what the true R_t and serial intervals are. Perhaps, we simply state what the error is and then look at the coverage to indicate that the method



works.

Fig 9: Comparison of inference for simulation method vs Epi-Estim (with perfect information)



This

could replace Fig 8 if we want to compare our estimate and substantiate our claim more clearly. *Fig 10: Comparison of inference for simulation method when different true values of rho are modelled* Message for Fig 10: For the plausible range of ρ values (0.33-0.83), we see that the error decreases as we reporting rates get higher. This message can be used to motivate higher recording rates in epidemics.

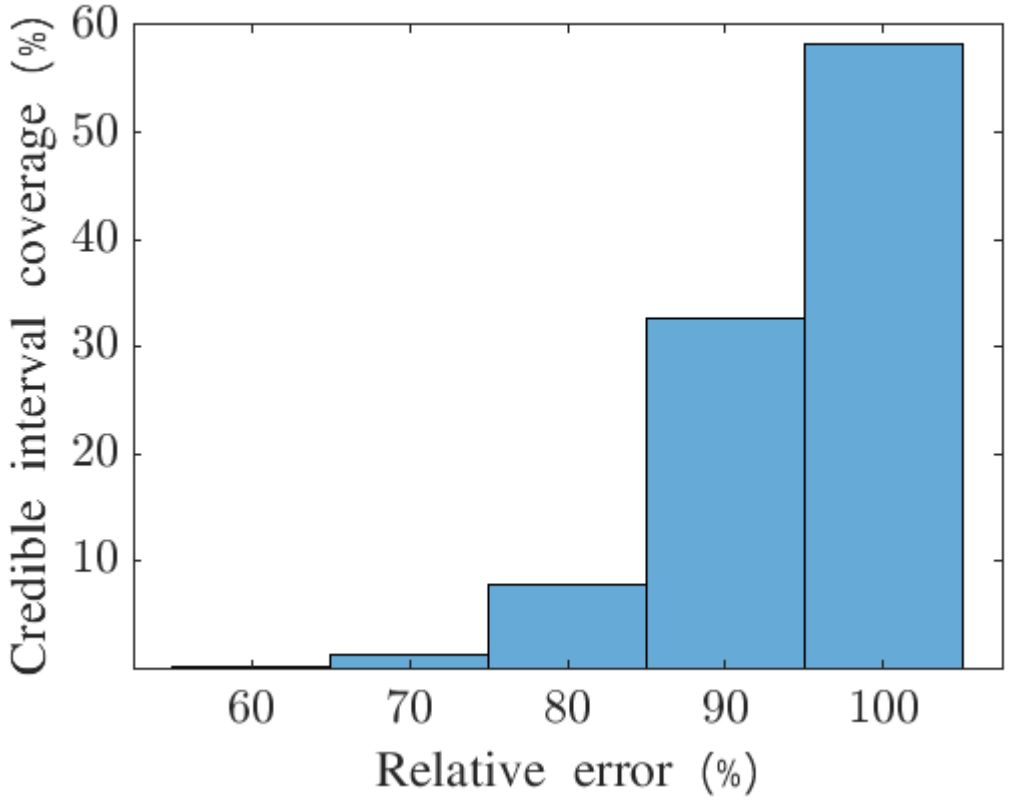


Fig 11: Looking at coverage for over many statistics Message for Fig 11: Along with the statistic that 94.8% of all credible intervals correctly contained the true reproduction number, this figure demonstrates that the coverage is consistent.