Understanding the effects of date rounding in phylodynamics

Leo A. Featherstone^{*,1}, [Order TBA], Sebastian Duchene^{†,1}

July 11, 2023

[1] Peter Doherty Institute for Infection and Immunity, University of Melbourne, Australia.
email: leo.featherstone@unimelb.edu.au

Abstract

Serving as a guideline for the message of the paper for now. Phylodynamic and genomic epidemiology frequently rely on the sampling times from pathogen isolates to make inference about the development of disease outbreaks over time. By calibrating a rate of substitution against epidemiological timescales, sampling times, in combination with genome sequence, allow for inferences such as the time of onset of an outbreak and intensity of transmission. However, for patient confidentiality, the exact sampling times for many sequences are not given, or rounded to a less precise amount of time such as month or year. Here, we show for the first time how and when such 'date-rounding' induces bias in epidemiological estimates. Broadly, this bias is often substantial, increases with the degree of rounding in provided sampling dates, and affect inference of all parameters including of time of onset and reproductive number. Finally, we close by proposing a solution that prioritises both patient confidentiality and accuracy of inference in genomic epidemiology, by proposing a basic form of encryption of dates in absolute time by translating them by an unknown number.

Introduction

- How and why rounding of dates occurs
- Types of rounding and presumed problems, including distorting clock signal (e.g. 'stress' in the rtt?)
- Consider a figure to show the distortion in the phylogram and in an rtt?
- 'effective number of mutations' do we still need this concept?
- That it might matter for the clock and what downstream parameters

Increased sharing of pathogen genome sequences has been a feature of responses to recent infectious disease threats. This is also the culmination of a broader trend that has build with advances in WGS.

Results

Results Overview

Points to hit

- Bias in all parameters
- R_e / R_0 is biased upwards
- origin ispushed deeper in time, implicating a more severe longstanding outbreak
- Clock rate is increased, suggesting a faster rate of mutation
- This trend is worst for smaller datasets, where the duration of infection is shorter relative to the error induced in rounding
- By contrast TB is affected less, but we need the specificity for emergent outbreaks most
- Show that severity seems to correlate with relationship between error and timing, if I can motivate one in introduction

Simulation Study

Empirical Results

- Here, do I want to group viruses or do each individually? Depends how much I say in the general results overview part.

	organism	resolution	meanR0	R0HPD	meanRe1	Re1HPD
1	h1n1	Day	1.08	[1.05169500944395, 1.11375862585647]		[NA, NA]
2	h1n1	Month	1.14	[1.11518272001637, 1.17471931965058]		[NA, NA]
3	h1n1	Year	115948630.02	[89603766.8312088, 145241108.540186]		[NA, NA]
4	sars-cov-2	Day	1.20	[0.918485254309922, 1.57665249346316]		[NA, NA]
5	sars-cov-2	Month	5.95	[3.83426191124472, 9.17572009709929]		[NA, NA]
6	sars-cov-2	Year	18.54	[10.3581474274774, 29.3223275555288]		[NA, NA]
7	shigella	Day		[NA, NA]	1.08	[1.056765724105
8	shigella	Month		[NA, NA]	1.09	[1.057276633108
9	shigella	Year		[NA, NA]	1.15	[1.116569656645
10	tb	Day		[NA, NA]	2.51	[0.683584249035
11	tb	Month		[NA, NA]	2.80	[0.611546919439
12	tb	Year		[NA, NA]	2.76	[0.49322402519

	organism	resolution	meanR0Err	meanRe1Err	meanRe2Err	meanPErr	meanDeltaErr	mean
1	h1n1	Day	0.03			0.00		
2	h1n1	Month	1.48			0.06		
3	h1n1	Year	327918625.00			0.47		
4	sars-cov-2	Day	0.94				12.56	
5	sars-cov-2	Month	39.78				33.12	
6	sars-cov-2	Year	28.98				74.51	
7	shigella	Day		0.42	0.05	0.25		
8	shigella	Month		0.37	0.62	0.45		
9	shigella	Year		0.23	1579468960284.99	0.28		
10	tb	Day		0.54	0.20	0.02	0.56	
11	tb	Month		0.55	0.20	0.02	0.56	
12	tb	Year		0.57	0.19	0.02	0.60	

Discussion

Brakdown of points to hit

- Why was each parameter biased the way it was?
- Overall, samples get clustered to the same time, but still have differences in sequence
- Suggests a higher mutation rate and transmission rate
- Hence spuriously large results for when we condense to year for H1N1
- Obviously, we would never rely on such results. For example, and R_e of 10^8 for H1N1 suggested the globe's population would be infected in one transmission event. Such is the blindness of our models. More tangibly though, we can always expect bias from what the full dates are suggesting, so the best thing to do is use the full date.

A simple solution

The only information that matters, is the *difference* between sequences and dates, rather than their absolute values. After all, our methods are comparative within a sample. Thus we can prioritise exact information and protect patient identity at the same time. We propose that authorities can provide dates that are all shifted in time by an unknown seed number, and reinterpret results by factoring this in. For example, if the sampling times of a dataset of 3 samples are (2000, 2001, 2002), then public health authorities may randomly draw a seed of 1000 with which to shift and dates and pass onto scientists: (2000, 2001, 2002) → (3000, 3001, 3002). Then results can be reinterpreted with regard to the random seed. If, for example the estimated time of onset was 3 years before the most recent sample, then those receiving the data will not be able to place this in time, while those on the data generation end can interpret this correctly (estimated time of onset = 2002-3 = 1999). In the same vein, transmission parameters such as R_e can be understood to pertain to the true sampling time.

Methods

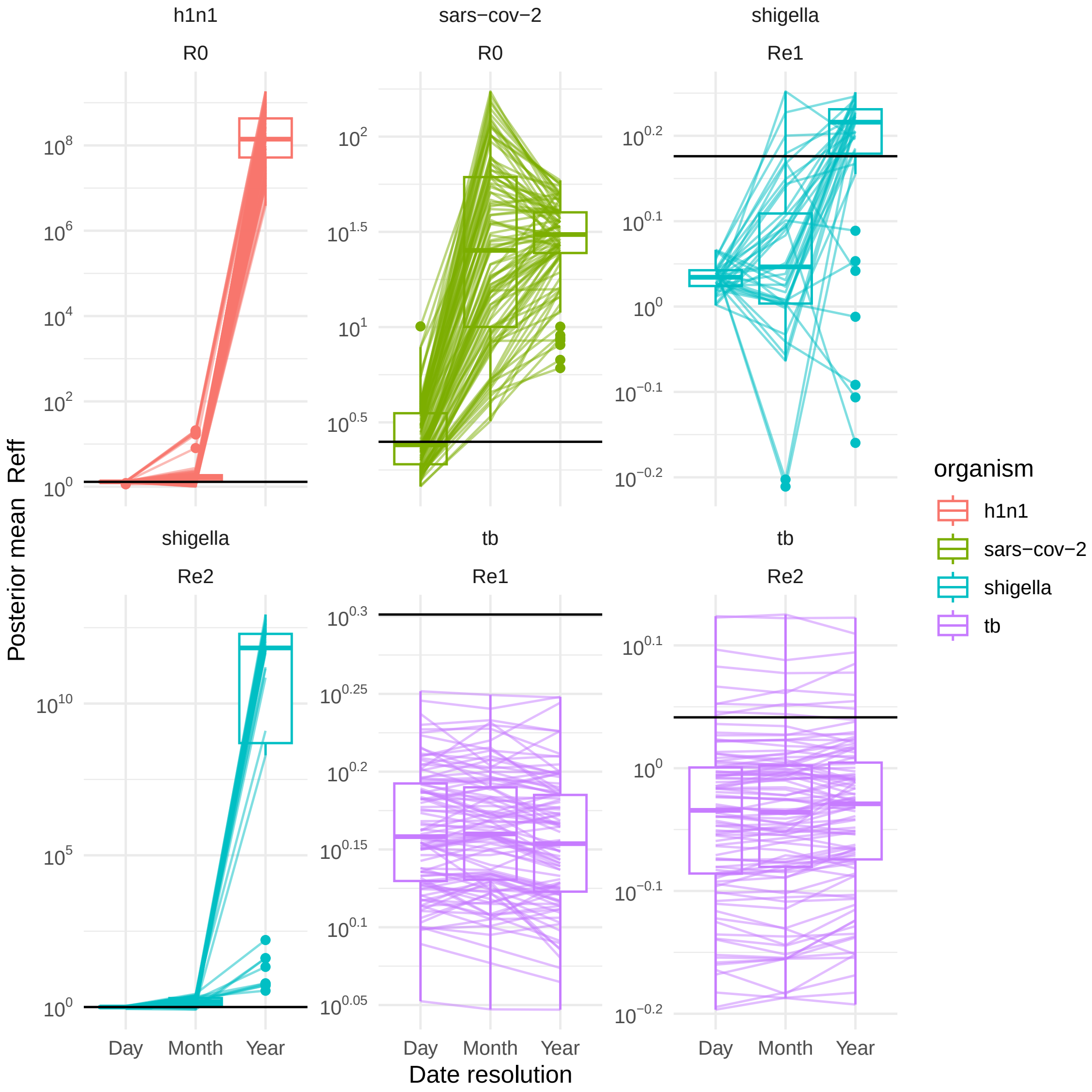

Figure 1: Mean posterior R_0 or R_e for simulated datasets across each level of date resultion, separated over simulation conditions emulating each pathogen (n=100). I.e. One line connects mean posterior estimates for a single simulated datasets analyses under each data resultuiton codition. For H1N1 and SARS-CoV-2 conditions, reducing date resolution (left to right) corresponds to upwards bias mean posterior R_0 and R_e. For year, having idnetical dates corresponds to complete model mispecification and wholly implausible mean posterior $R_\bullet$ values.

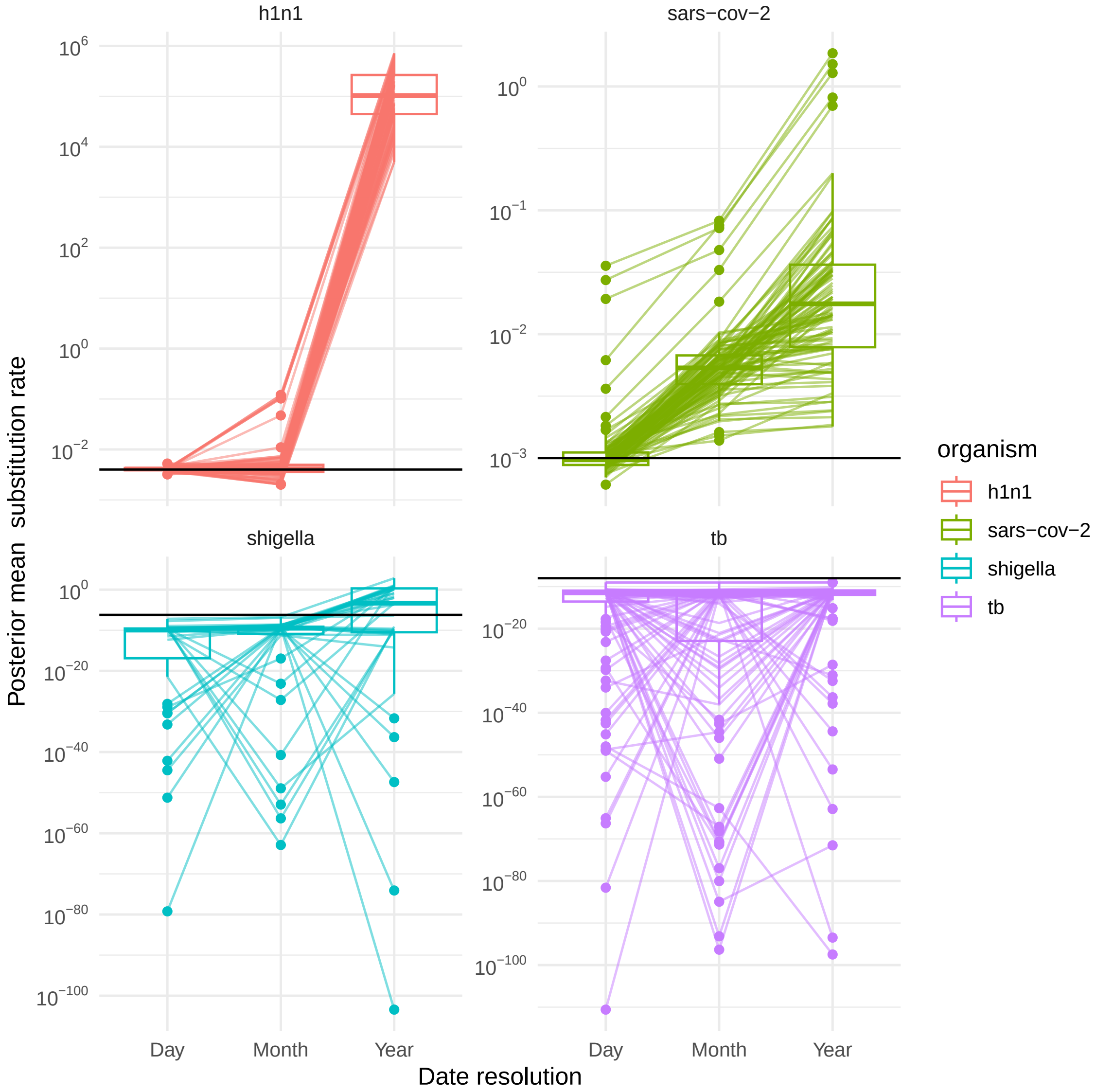

Figure 2: Mean posterior evolutionary rate for simulated datasets across each level of date resultion, separated over simulation conditions emulating each pathogen (n=100). I.e. One line connects mean posterior estimates for a single simulated datasets analyses under each data resultuiton codition. For H1N1 and SARS-CoV-2 conditions, reducing date resolution (left to right) corresponds to upwards bias mean posterior evolutionary rate. For Shigella this effect still present, wlthough diminshed, and estimates for TB appear relatively stable.

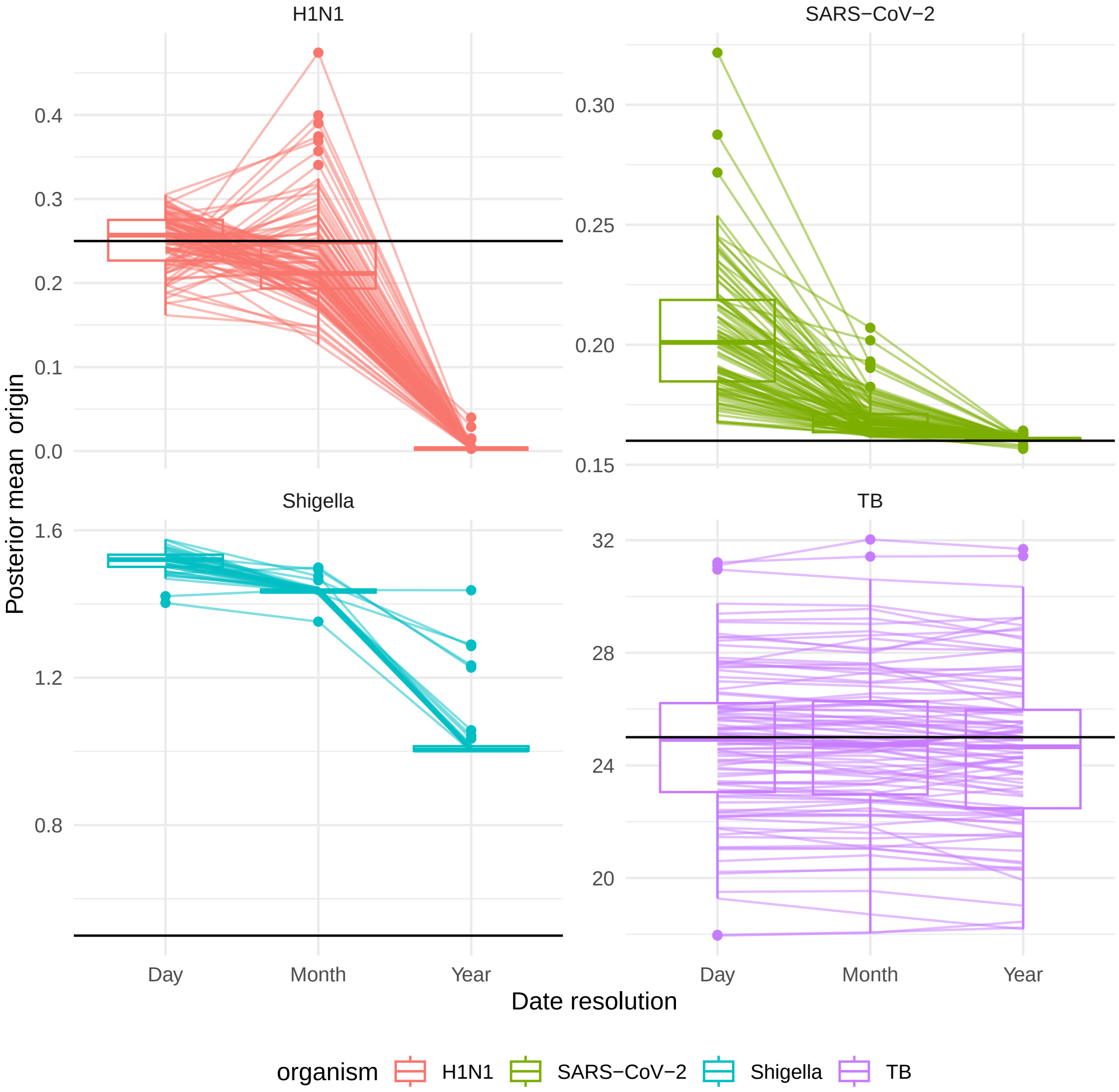

Figure 3: Mean posterior origin for simulated datasets across each level of date resultion, separated over simulation conditions emulating each pathogen (n=100). I.e. One line connects mean posterior estimates for a single simulated datasets analyses under each data resultuiton codition. TO FINISh