# Understand graph-prior ramifications

For a given recurrent state, we assume relationship graphs are uniformly distributed *a priori*. This has various ramifications discussed below.

## Contents



# 1 Posterior bounds that are induced by the graph prior

When the prior probability on relapse is non-zero, posterior probabilities of reinfection and recrudescence never reach one because all genetic data is always compatible with relapse. In this section, we discuss bounds on posterior probabilities of reinfection and recrudescence; they are induced by the uniformity of the prior over relationship graphs. Bounds define a feasible set of posterior probabilities without any information about the genetic data beyond that used to derive the multiplicities of infection (MOIs). For example, in the case of two genotypes in an enrolment episode followed by a monoclonal recurrence, when recurrent states are equally likely *a priori*, the posterior probabilities of reinfection and recrudescence are upper bounded by $^9/_{11}$ and $^9/_{13}$, respectively. The feasible set of posterior probabilities is shown in Figure 1.

## 1.1 How bounds are derived

We start by making some observations about the posterior odds of relapse to reinfection / recrudescence; odds help simplify some derivations.

### 1.1.1 Single recurrence

For clarity of exposition, we start with the case of a single recurrent episode. Let $\boldsymbol{\mathcal{G}}_{\mathrm{C}}, \boldsymbol{\mathcal{G}}_{\mathrm{L}}$, and $\boldsymbol{\mathcal{G}}_{\mathrm{I}}$ denote subsets of the graph space $\boldsymbol{\mathcal{G}}$, containing the relationship graphs compatible with recrudescence, relapse, and reinfection respectively. (To be clear, $\boldsymbol{\mathcal{G}}_{\mathrm{L}}$ is exactly $\boldsymbol{\mathcal{G}}$.) The posterior odds of relapse to recrudescence is given by

$$o_{\mathrm{L:C}} \coloneqq \frac{\mathbb{P}(\boldsymbol{y}|\mathrm{L})\mathbb{P}(\mathrm{L})}{\mathbb{P}(\boldsymbol{y}|\mathrm{C})\mathbb{P}(\mathrm{C})} = \frac{\mathbb{P}(\mathrm{L})}{\mathbb{P}(\mathrm{C})}\frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})\mathbb{P}(\boldsymbol{g}|\mathrm{L})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{C}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})\mathbb{P}(\boldsymbol{g}|\mathrm{C})} = \frac{\mathbb{P}(\mathrm{L})}{\mathbb{P}(\mathrm{C})}\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{C}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|}\frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{C}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})} = \frac{\mathbb{P}(\mathrm{L})}{\mathbb{P}(\mathrm{C})}\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{C}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|}\left(1+\frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{C}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{C}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}\right),$$

where $\boldsymbol{\mathcal{G}}_{\mathrm{L}} \setminus \boldsymbol{\mathcal{G}}_{\mathrm{C}}$ is the subset of graphs compatible with relapse but not recrudescence (graphs where genotypes from the recurrent episode are not all clones from the initial episode). It follows that $o_{\mathrm{L:C}} \geq {}^{\mathbb{P}(\mathrm{L})|\boldsymbol{\mathcal{G}}_{\mathrm{C}}|}/_{\mathbb{P}(\mathrm{C})|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|}$. Note that this inequality is close to equality when $\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{C}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}) \ll \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{C}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})$.

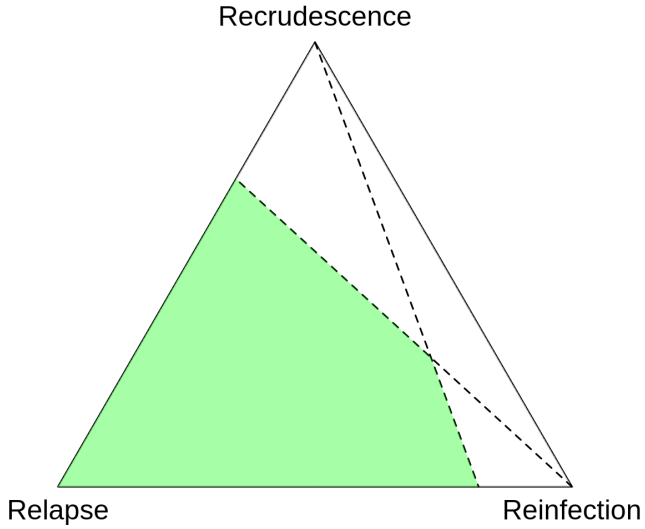



Figure 1: Feasible set of posterior probabilities in the case of two genotypes in the initial episode and one genotype in the recurrent episode. Dashed lines intersect the left and bottom edges of the simplex at $\frac{4}{13}$ and $\frac{2}{11}$, respectively.

Similarly, the posterior odds of relapse to reinfection is given by

$$o_{\mathrm{L:I}} := \frac{\mathbb{P}(\boldsymbol{y}|\mathrm{L})\mathbb{P}(\mathrm{L})}{\mathbb{P}(\boldsymbol{y}|\mathrm{I})\mathbb{P}(\mathrm{I})} = \frac{\mathbb{P}(\mathrm{L})}{\mathbb{P}(\mathrm{I})}\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{I}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|}\left(1 + \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{I}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{I}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}\right).$$

It follows that $o_{\mathrm{L:I}} \geq {}^{\mathbb{P}(\mathrm{L})|\boldsymbol{\mathcal{G}}_{\mathrm{I}}|}/_{\mathbb{P}(\mathrm{I})|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|}$. Note that this inequality is close to equality when $\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{I}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}) \ll \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{I}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})$.

The posterior odds $o_{\mathrm{I:C}}$ can be defined similarly, but analogous lower bounds are not available. This is because $\boldsymbol{\mathcal{G}}_{\mathrm{C}}$ and $\boldsymbol{\mathcal{G}}_{\mathrm{I}}$ do not intersect. Thus, we simply note that $o_{\mathrm{I:C}} \approx 0$ when $\mathbb{P}(\boldsymbol{y}|\mathrm{I}) \ll \mathbb{P}(\boldsymbol{y}\mathrm{C})$.

As for the posterior probabilities, we have

$$\begin{aligned}
\mathbb{P}(\mathrm{C}|\boldsymbol{y}) &= \frac{\mathbb{P}(\boldsymbol{y}|\mathrm{C})\mathbb{P}(\mathrm{C})}{\mathbb{P}(\boldsymbol{y}|\mathrm{C})\mathbb{P}(\mathrm{C}) + \mathbb{P}(\boldsymbol{y}|\mathrm{L})\mathbb{P}(\mathrm{L}) + \mathbb{P}(\boldsymbol{y}|\mathrm{I})\mathbb{P}(\mathrm{I})} \\
&= \frac{1}{1 + o_{\mathrm{L:C}} + o_{\mathrm{I:C}}} \\
&\leq \frac{1}{1 + \frac{\mathbb{P}(\mathrm{L})|\boldsymbol{\mathcal{G}}_{\mathrm{C}}|}{\mathbb{P}(\mathrm{C})|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|} + 0} \\
&= \frac{\mathbb{P}(\mathrm{C})}{\mathbb{P}(\mathrm{C}) + \mathbb{P}(\mathrm{L})\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{C}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|}}.
\end{aligned} \tag{1}$$

This inequality is close to equality when $\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{C}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}) \ll \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{C}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})$ and $\mathbb{P}(\boldsymbol{y}|\mathrm{I}) \ll \mathbb{P}(\boldsymbol{y}|\mathrm{C})$. Note that the former implies the latter as $\boldsymbol{\mathcal{G}}_{\mathrm{I}} \subseteq \boldsymbol{\mathcal{G}}_{\mathrm{L}} \setminus \boldsymbol{\mathcal{G}}_{\mathrm{C}}$.

Similarly, we have

$$\mathbb{P}(\mathrm{I}|\boldsymbol{y}) \leq \frac{\mathbb{P}(\mathrm{I})}{\mathbb{P}(\mathrm{I}) + \mathbb{P}(\mathrm{L})\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{I}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{L}}|}}. \tag{2}$$

Likewise, this inequality is close to equality when $\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{L}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{I}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}) \ll \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{I}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})$.

### 1.1.2 More than one recurrence

When there is more than one recurrence and prior probabilities are non-zero, the maximum probabilities of all sequences are non-certain with the exception of the relapse-only sequence. As before, we can compute odds; for

example

$$o_{\mathrm{LL:CL}} \coloneqq \frac{\mathbb{P}(\boldsymbol{y}|\mathrm{LL})\mathbb{P}(\mathrm{LL})}{\mathbb{P}(\boldsymbol{y}|\mathrm{CL})\mathbb{P}(\mathrm{CL})} = \frac{\mathbb{P}(\mathrm{LL})}{\mathbb{P}(\mathrm{CL})}\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CL}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{LL}}|}\frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{LL}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CL}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})} = \frac{\mathbb{P}(\mathrm{LL})}{\mathbb{P}(\mathrm{CL})}\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CL}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{LL}}|}\left(1+\frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{LL}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{CL}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CL}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}\right).$$

However, unlike before, we cannot always derive maximum posterior probabilities from the odds because the probabilities of the remaining sequences are not necessarily zero (the exception being when the odds of an all-but-one-relapse to all-relapse sequence are maximised). Instead, we must compute maximum probabilities the long way; for example,

$$\begin{aligned}\mathbb{P}(\mathrm{CC}|\boldsymbol{y}) = &\left\{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})^{\mathbb{P}(\mathrm{CC})/|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|}\right\}\left\{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})^{\mathbb{P}(\mathrm{CC})/|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|} + \left(\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}) + \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{LC}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})\right)\frac{\mathbb{P}(\mathrm{LC})}{|\boldsymbol{\mathcal{G}}_{\mathrm{LC}}|}\right.\\ &+\left(\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}) + \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{LL}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})\right)\frac{\mathbb{P}(\mathrm{LL})}{|\boldsymbol{\mathcal{G}}_{\mathrm{LL}}|} + \left(\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g}) + \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CL}}\setminus\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})\right)\frac{\mathbb{P}(\mathrm{CL})}{|\boldsymbol{\mathcal{G}}_{\mathrm{CL}}|} + \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CI}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})^{\mathbb{P}(\mathrm{CI})/|\boldsymbol{\mathcal{G}}_{\mathrm{CI}}|} +\right.\\ &\left.\left.+\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{LI}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})^{\mathbb{P}(\mathrm{LI})/|\boldsymbol{\mathcal{G}}_{\mathrm{LI}}|} + \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{IC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})^{\mathbb{P}(\mathrm{IC})/|\boldsymbol{\mathcal{G}}_{\mathrm{IC}}|} + \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{IL}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})^{\mathbb{P}(\mathrm{IL})/|\boldsymbol{\mathcal{G}}_{\mathrm{IL}}|} + \sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{II}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})^{\mathbb{P}(\mathrm{II})/|\boldsymbol{\mathcal{G}}_{\mathrm{II}}|}\right\}^{-1}\end{aligned} \tag{3}$$

is approximately

$$\frac{\mathbb{P}(\mathrm{CC})}{\mathbb{P}(\mathrm{CC}) + \mathbb{P}(\mathrm{LC})\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{LC}}|} + \mathbb{P}(\mathrm{LL})\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{LL}}|} + \mathbb{P}(\mathrm{CL})\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{CL}}|}}$$

when $\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})$ significantly exceeds the summation over all other subsets of graph space in equation (3).

Bounds on marginal posteriors are harder to compute because computation involves summation over sequences whose probabilities are not necessarily zero. For example, the bound on recrudescence at the first recurrence of two recurrences, may be close to one of three upper bounds each involving non-zero probabilities:

1. bound given maximal CC probability (probabilities of CI & CL are not necessarily both zero)
2. bound given maximal CL probability (probabilities of CC & CI are not necessarily both zero)
3. bound given maximal CI probability (probabilities of CC & CL are not necessarily both zero)

**Digression** In the online article "Understand posterior estimates" we compute maximum marginal probabilities of recrudescence / reinfection at the first recurrence of two / three recurrences. These maxima are not bounds imposed by the prior: they are based on additional knowledge that there are no recurrent data on all but the first recurrence. The computation only holds because all episodes are monoclonal. When all episodes are monoclonal, the likelihood of equivalent sequences (e.g., CC, CI and CL in the case of two recurrences with strong evidence of recrudescence on the first and no data on the second) are equal (otherwise, they are unequal for reasons analogous to those that explain departure from the prior in the section "Data on only one episode") and graph likelihoods are equal for all graphs compatible with equivalent sequences (e.g., $\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})$ is the same $\forall\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}$, $\forall\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CI}}$, and $\forall\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CL}}$ in the case of two recurrences with strong evidence of recrudescence on the first and no data on the second). Given a uniform prior on recurrent states, three monoclonal episodes with strong evidence of recrudescence for the first recurrence and no data on the second recurrence, we thus have

$$\mathbb{P}(s_1 = \mathrm{C}|\boldsymbol{y}) \tag{4}$$

$$\begin{aligned}&\leq \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})/|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|}{\frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|} + \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CI}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{|\boldsymbol{\mathcal{G}}_{\mathrm{CI}}|} + \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CL}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{|\boldsymbol{\mathcal{G}}_{\mathrm{CL}}|} + \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CC}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{|\boldsymbol{\mathcal{G}}_{\mathrm{LC}}|} + \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CI}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{|\boldsymbol{\mathcal{G}}_{\mathrm{LI}}|} + \frac{\sum_{\boldsymbol{g}\in\boldsymbol{\mathcal{G}}_{\mathrm{CL}}}\mathbb{P}(\boldsymbol{y}|\boldsymbol{g})}{|\boldsymbol{\mathcal{G}}_{\mathrm{LL}}|}}\\ &= \frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|/|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|}{\frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|} + \frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CI}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{CI}}|} + \frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CL}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{CL}}|} + \frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CC}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{LC}}|} + \frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CI}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{LI}}|} + \frac{|\boldsymbol{\mathcal{G}}_{\mathrm{CL}}|}{|\boldsymbol{\mathcal{G}}_{\mathrm{LL}}|}}\\ &= \frac{1}{3 + {}^1\!/_3 + {}^1\!/_3 + {}^3\!/_{12}}.\end{aligned} \tag{5}$$

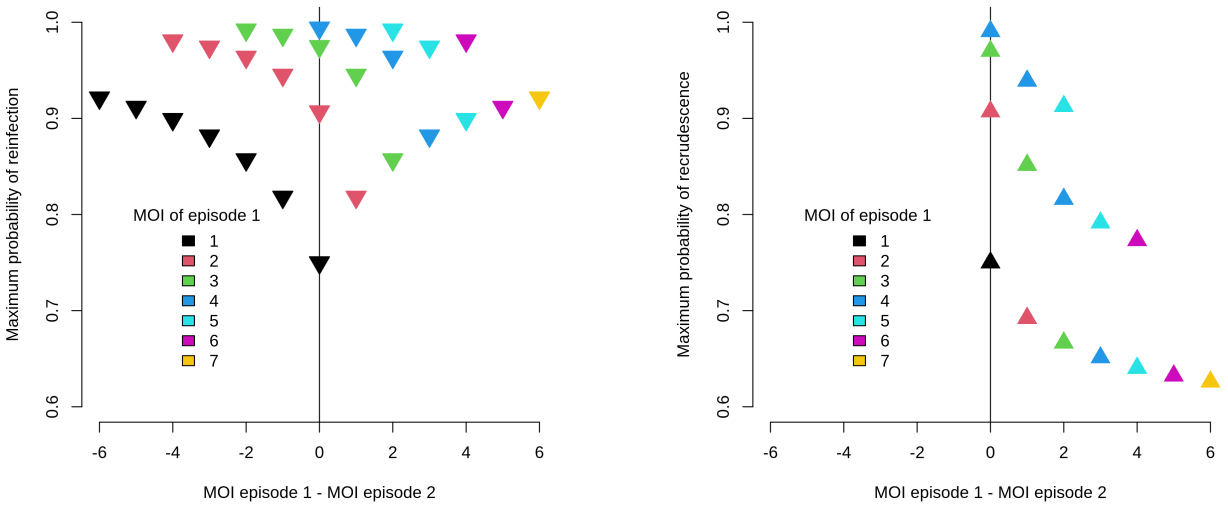
The same reasoning applies given strong evidence of reinfection.

## 2 Bounds as indicators of data informativeness

By comparing probability estimates to their respective bounds, we might like to answer the question: could the posterior probability of my most probable state sequence be higher if we had data on more markers? However, we cannot guarantee convergence onto bounds. In the case of a marginal probability, the bound is the maximum of multiple upper bounds (see example above), and thus convergence is not guaranteed; in the case of a non-marginal probability, we suspect convergence might depend on the relationship graph from which the data are generated.

## 3 How bounds induced by the graph prior change with MOI

In this section we explore how maximum probabilities of recrudescence / reinfection given a single recurrence change with graph size, where maximum probabilities are those induced by the the graph prior. Maximum probabilities of reinfection increase with the size of the graph, both when MOIs of the enrolment episode and the first recurrence are equal (plot 2a, centre) and not (plot 2a, off-centre). Maximum probabilities of recrudescence increase with graph size when the MOIs of the enrolment episode and the first recurrence are equal (plot 2b, centre); they decrease with increasing disparity between the higher MOI of the enrolment episode and the lower MOI of the first recurrence (plot 2b, right of centre). Note that, when the MOI of first recurrence exceeds that of the enrolment episode, recrudescence has zero posterior probability because we assume under the Pv3Rs model that all parasites are detected and that there are no genotyping errors; as such, a recrudescence can be at most as diverse as the preceding episode.




(a) (b)

Figure 2: Maximum probabilities of recrudescence / reinfection given a single recurrence.

## 4 A counter-intuitive property

The *a priori* assumption that all valid relationships graphs $\boldsymbol{g}$ are equally likely given a recurrent state leads to a counter-intuitive property. Specifically, the probability of an edge can vary depending on the graph it is embedded within. Consider the scenario where one monoclonal recurrence follows a monoclonal enrolment episode.

Because we assume that the three relationship graphs are equally likely given relapse, the probability distribution of the edge between the two genotypes given relapse is

$$\mathbb{P}(\text{the two genotypes are clones}|\boldsymbol{s} = \text{L}) = 1/3, \tag{6}$$

$$\mathbb{P}(\text{the two genotypes are siblings}|\boldsymbol{s} = \text{L}) = 1/3, \tag{7}$$

$$\mathbb{P}(\text{the two genotypes are strangers}|\boldsymbol{s} = \text{L}) = 1/3. \tag{8}$$

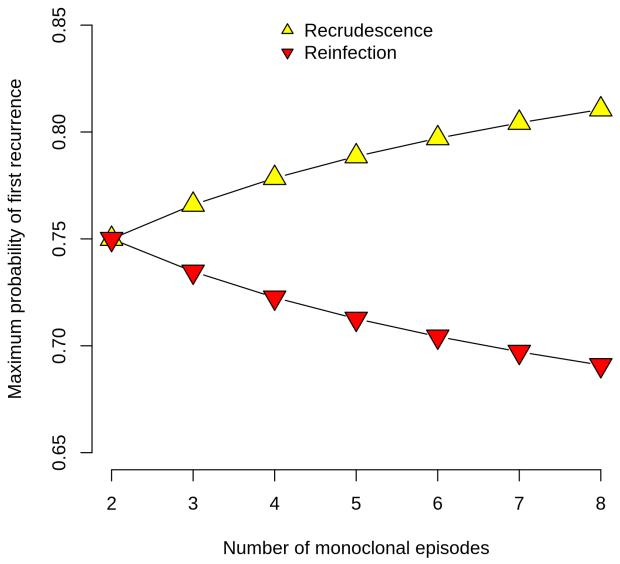



Figure 3: The probability that the first recurrence is a recrudescence / reinfection increases / decreases with additional recurrences devoid of data. The prior on recurrent states is uniform. The maximum probabilities are those based on knowledge that there are no recurrent data for all but the first recurrence and all episodes are monoclonal. For example, from equations (6) and (9) we have $|\mathcal{G}_{\mathrm{C}}|/|\mathcal{G}_{\mathrm{L}}| = 1/3 > |\mathcal{G}_{\mathrm{CL}}|/|\mathcal{G}_{\mathrm{LL}}| = 3/12$. Plugging $|\mathcal{G}_{\mathrm{C}}|/|\mathcal{G}_{\mathrm{L}}| = 1/3$ into equation (1), and $|\mathcal{G}_{\mathrm{CL}}|/|\mathcal{G}_{\mathrm{LL}}| = 3/12$ plus $|\mathcal{G}_{\mathrm{CC}}|/|\mathcal{G}_{\mathrm{LC}}| = 1/3$ and $|\mathcal{G}_{\mathrm{CI}}|/|\mathcal{G}_{\mathrm{LI}}| = 1/3$ into equation (5), we see that the probability that the first recurrence is a recrudescence when recurrent states are equally likely *a priori* increases slightly from $3/4$ to $3/(3 + 11/12)$ with the addition of the second recurrence without data.

Now consider the scenario in which there is an additional monoclonal recurrence.

There are now 12 relationship graphs, which are assumed to be equally likely under relapses. Among them, between the first two genotypes, three have a clonal edge, four have a sibling edge, and five have a stranger edge. The probability distribution of the edge between the first two genotypes given relapses is thus

$$\mathbb{P}(\text{the first two genotypes are clones}|\boldsymbol{s} = \mathrm{LL}) = 3/12, \tag{9}$$

$$\mathbb{P}(\text{the first two genotypes are siblings}|\boldsymbol{s} = \mathrm{LL}) = 4/12, \tag{10}$$

$$\mathbb{P}(\text{the first two genotypes are strangers}|\boldsymbol{s} = \mathrm{LL}) = 5/12. \tag{11}$$

An explanation for the change in the probability distribution of the edge between the first two genotypes upon the addition of the second relapse is that a clonal edge between the first two genotypes imposes a constraint where the two remaining edges must exhibit the same relationship. Similarly, a sibling edge between the first two genotypes is incompatible with a single stranger edge among the two remaining edges. In general, the pattern that stranger edges are more likely *a priori* and clonal edges are less likely *a priori* becomes more prominent when more genotypes are present. For monoclonal episodes, this change in distribution over the edge between the first two genotypes results in an increase / decrease in the maximum probability that the first recurrence is a recrudescence / reinfection with the addition of monoclonal recurrences devoid of data (Figure 3), where the maximum probabilities are those assuming all episodes are monoclonal and there are no recurrent data on all but the first recurrence (e.g., as in equation (5)).