Methods for estimating Reproduction Number, R(t)

Laura White, PhD

2025-02-14

Table of contents

Preface

Since the onset of the COVID-19 pandemic in early 2020, there has been a proliferation of software packages that make inference about the current state of an infectious disease outbreak based on daily counts of disease cases. An important and widely used parameter is the instantaneous or effective reproduction number, R(t), defined in Gostic et. al. (2020) 1 as follows:

"The effective reproductive number, denoted as Re or Rt, is the expected number of new infections caused by an infectious individual in a population where some individuals may no longer be susceptible."

Defined as such, R(t) is an unobserved quantity that captures the aggregated combination of disease characteristics (e.g., infectiousness under controlled conditions, mode of transport) and extrinsic factors (e.g., lockdowns that reduce person-to-person contact). An $R(t) = 1$ indicates a "stable" epidemic at time t, where each infected person infects on average one additional person; R(t) values above or below 1 represent a growing or a shrinking epidemic respectively. We focus on R(t) given its relevance for time-sensitive decision-making, and summarize the currently used inputs, data, methods and assumptions in R(t) estimation across the following categories:

- How the relationship between R(t) and infections is defined
- How R(t) is constrained using distributions for key variables
- How R(t) is constrained over time
- Additional data and distributions that are used to constrain R(t)
- Inference frameworks that are used to estimate R(t) ** is constrained the right word?

In this paper we provide a theoretical comparison of the current field of methods for estimating R(t), with the goal of informing user decision-making about which package to choose and in interpreting package outputs. For reference, Table 1 lists packages with an accompanying peer-reviewed journal manuscript, Table 2 lists packages without a peer-reviewed journal manuscript, and Table S1 contains R(t) packages that calculate R(t) but were excluded from this summary. In the text, the package citation is given the first time each package is referenced.

We limit the methods discussed here to those for estimating historical to present-day R(t) values using daily case count data, where a case can be flexibly defined as an individual with a reported positive test (either through healthcare-seeking behavior, routine surveillance, or a hospital admission). Other methods not discussed here include inference of R(t) exclusively

from alternative data sources (e.g., genetic data,2 behavioral data,3 or viral loads in wastewater4), or calculations from compartmental, agent-based models, or network.5–7 We also limit the discussion to packages in the statistical software R,8 which may exclude some packages in other software programs that combine many of the methodological considerations discussed below.9 We do not discuss any packages for now-casting or forecasting, though a number of R(t) estimation packages can be used for this purpose. The methods discussed below and references to specific R packages are current as of December 1, 2024. We attempt to harmonize the mathematical choices between each package using terminology from each.

An evaluative comparison of the performance of these methods would be highly complex, given the following challenges. Some of the most widely-used packages are not accompanied with a peer-reviewed manuscript that describes or evaluates the theory behind modeling choices. Each package contains a subset of the methods below for constraining R(t) in time, but with subtle variations in implementation that are often not well-documented. Most packages have not been recently updated, and even those that have are not maintained on CRAN, instead leaving updates on a development version on GitHub. The combination of differing model frameworks associated with each package make it challenging to easily compare the impacts on estimated R(t), especially when considering additional factors like ease of implementation and computational time

1 Introduction

There are two primary classes methods of estimating R(t) from case count data that are used in most R software packages. The first class of methods assumes there is a formulaic relationship between infections and reproduction number, a relationship known as the renewal equation.10 These infections are then assumed to result in (some fraction of) the observed cases. A second class of methods involves empirically calculating a quantity that approximates the latent quantity represented by a reproduction number by fitting a curve to the case count time-series and finding the time-varying slope in log space (and then performing other transformations). Empirical calculations are discussed in detail below in our examination of ways in which R(t) is constrained over time.

1.1 Renewal equation estimates of R(t)

The renewal equation relates R(t) and infections on day t, I(t), using a third parameter known as the generation interval. The generation interval, , is the time between infection in the infector and infection in the infectee, and assuming independence is the linear combination of incubation time, the time between infection and symptom onset in an individual, and transmission time, the time between symptom onset in the infector and infection of the infectee.11 A similar parameter to the generation interval is the serial interval, which is the time between symptom onset in the infector and symptom onset in the infectee. The serial interval and generation interval are interchangeable if the incubation time is independent from the transmission time, and some formulations of the renewal equation use generation interval. In this paper we use the generation interval described by a probability mass function with non-zero values from day 1 (assuming that disease incubation takes at least 1 day) to a maximum day s, i.e., the longest interval between symptom onset in infector and infectee. Taking care to note that R(t) is undefined on day 0 since there has been no transmission yet (and assuming the initial infections are I(0)), the formulation of the renewal equation is thus:

I(t)=R(t) _(i=max(1, t-s+1))^t (i) I(t-i) (Eq.1)

For brevity, we write the inner sum of (Eq.1) as:

Λ(t)= _(i=max(1, t-s+1))^t (i) I(t-i) (Eq.2)

The assumptions of this formulation, as per Green et. al. 2022,12 are that incident infections can be described deterministically within each window of t [t-s+1,t] and that the generation interval distribution does not change over the modeling time. A common reframing of the renewal equation is to equate R(t) with an exponential growth rate, r. Under specific conditions and within a small time window (t [t-s+1,t]), infections can be assumed to grow exponentially at a constant rate (r).12–14 Using Eq. 1 in the time window t [t- s+1,t] and assuming some initial infections k, R(t) for t [t- s+1,t] can be inferred from only r and : I(t)=ke^rt,t [t-s+1,t] (Eq.3) R(t)=[_(i=max(1, t-s+1))^t (i) e^(-ri)]^(-1) ,t [t- s+1,t] (Eq.4) Again, we will omit the writing the bounds for time in remaining formulae. A single R(t) value, say R_0, can also be put in the form of an infection attack rate, z,15 or in the final size equation,16 to estimate the proportion of all individuals that were affected by a disease with this R_0:

z=1-exp(-R_0 z) (Eq.5)

The attack rate function and others are implemented in the package epigrowthfit.17 The major difference between calculating R(t) from a renewal equation or an exponential growth rate equation is whether I(t) is used. If for a given time window both r and can be estimated independently, then R(t) can be inferred without infection data. Otherwise, infection data are needed to estimate R(t).

Using the renewal equation (Eq. 1) and given that I(t) and are known, R(t) can be solved for algebraically starting with R(t=1) and iterating forwards in time. However, this will produce highly volatile estimates of R(t) that recover the incidence curve directly. This is undesirable for several reasons: real-world infectivity likely does not vary dramatically from day to day, and real-world infection data are rarely complete, especially in an emerging epidemic, meaning that a certain amount of uncertainty must be incorporated into any estimation framework. In addition, infection incidence, I(t), are the data of interest but it is impossible to observe, so many calculations instead may use the observed reported cases, C(t), which requires some additional processing to incorporate into calculations of R(t). Therefore, a variety of constraints on R(t) are added in the inferential process: using distributions on key variables, placing restrictions on how R(t) varies through time, and with additional data sources and delay distributions. These choices dictate which estimation framework is used, which can add additional constraints.

2 Decision Tree

Here's where we will have the decision tree

Part I

Packages

Here's where we will talk about the packages

EpiNow2

EpiEstim

RtEstim

Part II

Methods

Here's where we will talk about the methods