How to measure the controllability of an infectious disease?

Kris V Parag[1,2,*]

[1]MRC Centre for Global Infectious Disease Analysis, Imperial College London, London, UK.

[2]NIHR HPRU in Behavioural Science and Evaluation, University of Bristol, Bristol, UK.

*For correspondence: k.parag@imperial.ac.uk.

Abstract

Quantifying how difficult it is to control an emerging infectious disease is crucial to public health decision-making, providing valuable evidence on if targeted interventions e.g., quarantine and isolation, can contain spread or when population-wide controls e.g., lockdowns, are warranted. The disease reproduction number, R, or growth rate, r, are frequently assumed to measure controllability because R=1 and r=0 define epidemic stability. Outbreaks with larger R or r are therefore interpreted as less controllable and requiring more stringent interventions. We prove this common interpretation is impractical and incomplete. We identify a positive feedback loop among infections intrinsically underlying disease transmission and define controllability from how interventions disrupt this loop. The epidemic gain and delay margin, which describe how much we can scale infections and delay interventions on this loop before losing stability, yield rigorous measures of controllability. Outbreaks with smaller margins necessitate more control effort. Using these margins, we quantify how presymptomatic spread, surveillance limitations, variant dynamics and superspreading shape controllability and show that R or r only measures controllability when interventions do not alter timings between infections and are implemented without delay. Our margins are easily computed, interpreted and reflect complex relationships among interventions, their implementation and epidemiological dynamics.

Keywords: infectious diseases; feedback control; stability margins; reproduction numbers; non-pharmaceutical interventions, growth rates.

Introduction

Understanding and quantifying the effort required to control or contain outbreaks is a principal goal of infectious disease epidemiology [1]. During emergent stages of a potential epidemic, when populations are immunologically naïve, assessments of disease controllability provide critical evidence on whether targeted interventions, for example contact tracing, isolation and quarantines, are sufficient to curb spread [2] or whether non-selective controls, such as population-level lockdowns and closures, are necessary [3]. These assessments typically rely on mathematical models [4] that combine disease surveillance data (e.g., infection times and cases) with intervention mechanisms (e.g., how isolation interrupts transmission chains), to

estimate controllability (in some sense) and have informed the public health responses for influenza, measles, SARS, Ebola virus disease and COVID-19, among others [2,3,5–7].

Despite these applications, a systematic and rigorous definition of controllability is lacking [8–10]. While key factors influencing the difficulty of controlling epidemics such as transmissibility, superspreading levels, the efficiency of contact tracing and the proportion of presymptomatic infections are known [1,8,11], studies generally compute the reproduction number, R, or (less commonly) the growth rate, r, under proposed interventions to measure controllability [12]. For example, the impact of contact tracing efficiency and presymptomatic spread on controllability are assessed by how they effectively change R [1,13,14]. As R=1 or r=0 defines the point at which the epidemic becomes critically stable (i.e., infections neither increase nor decrease) [4], it seems reasonable to base controllability on the distance of R-1 or r-0.

We therefore expect that larger R or r signifies reduced controllability, justifying more stringent interventions, while smaller R or r indicates augmented robustness to transmissibility changes or intervention relaxations. The common interpretation is that we must scale infections by $1/R$ within timeframes proportional to $1/r$ to stabilise the epidemic [12]. Note that R and r are linked by the generation time distribution of the disease [15], w, which captures the times between infections. This interpretation further underlies related measures of intervention efficacy such as the herd immunity threshold (i.e., the proportion of the susceptible population that must be vaccinated or acquire immunity) [4,12] and the proportion of infections that must be targeted by contact tracing [2] (both relate to $1\text{-}1/R$), as well as the speed at which isolation or digital tracing [9,13,16] must be applied to suppress infections (both relate to $\log(2)/r$).

Here we demonstrate that the above interpretations are only valid under impractical and quite restrictive assumptions. We start by recognising that, intrinsically, an epidemic represents a positive feedback loop between past and upcoming infections. Interventions are then control actions that disrupt this loop. This reframing of the disease transmission process allows us to adapt tools from control theory [17] and derive epidemic transfer functions that capture how incident infections are generated under arbitrary generation time distributions, (linear) control actions and imported infection time series. We then propose a rigorous controllability measure defined by the gain and delay margin of the epidemic. These, respectively, quantify how much we must scale infections and delay interventions to achieve critical epidemic stability [18]. A gain margin of 2 and delay margin of 7 days, for example, mean our epidemic is stable unless we more than double infections or introduce delays longer than a week.

This framework yields a number of advantages and results. First, our margins more accurately describe what R and r only attempt to quantify – the scale and speed of required control effort.

We demonstrate that scaling infections by 1/(*kR*) for some *k*>1 only controls epidemics as we might expect (i.e., leads infinite delay margin and a gain margin of 1/*k*) if interventions reduce infections without inducing dynamics and are implemented without delays. This is unrealistic given mounting evidence that interventions change generation time and other key distributions (which is how control actions induce dynamics) and that practical outbreak control constraints always cause lags [7,19–22]. Further, we find that while *r* is the dominant pole of the epidemic transfer function, it only quantifies the asymptotic epidemic growth rate [23] and does not factor in how other poles modulate performance (e.g., these poles can cause unwanted oscillations in infections) and may interact with the induced controller dynamics to set controllability.

Second, our transfer function and margin-based approaches are flexible and both generalise and unify earlier frameworks [1,8]. We characterise how presymptomatic spread, transmission heterogeneities emerging from superspreading, multitype epidemics or co-circulating variants and surveillance limitations (e.g., reporting delays and underreporting) all affect controllability. These complexities can be commonly evaluated using our pair of margins, which always have the same interpretation. This is beneficial because *R* or *r* is not always clearly defined or even meaningful for some of these complexities [24,25]. Importantly, our margins yield thresholds of controllability under these complexities that can be directly compared to decide the relative effectiveness of targeted and population-level interventions. These thresholds reduce to more conventional 1-1/*R* type results under the restrictive conditions mentioned above.

Last, our margins offer a more complete measure of controllability. Because induced dynamics from interventions, implementation delays and surveillance imperfections are pervasive, even if proposed interventions are expected to drive *R*<1 or *r*<0, this does not reliably inform about the required control effort and the robustness of the epidemic once controlled by these actions. We find ample evidence of controlled (*R*<1) epidemics with gain margins under 1/*R*, indicating losses in robustness to increases in infections below what is conventionally expected. We also show that some of these controlled epidemics possess delay margins of 1-2 weeks, signifying that if the combined lag from surveillance and intervention delays rises above this value (e.g., due to reducing sustained surveillance or control) then the epidemic will become destabilised. Neither *r* nor *R* can expose this issue. Our methodology probes the concept of controllability and raises questions about the understudied knock-on effects of interventions.

Results

We start by exploring the conventional assumption that larger R or r signals a less controllable epidemic [1,10,26]. This belief is sensible as increases in R cause infections to multiply more, while rises in r engender faster multiplication. When $R > 1$ and hence $r > 0$ the interpretation

is that we must reduce transmission by a scale factor of R^{-1} i.e., block a fraction $1 - R^{-1}$ of active infections, before $r^{-1}\log R$ time units elapse or we fail to keep up with growth (note that $r^{-1}\log 2$ is known as the epidemic doubling time). Accordingly, when $R < 1$ and hence $r < 0$ we maintain stability for perturbations (e.g., due to pathogen fitness changes) or intervention relaxations that increase transmission by at most R^{-1} within a time frame of $r^{-1}\log R$.

Here we show that these almost ubiquitous interpretations are only true under restrictive and idealistic intervention (i.e., control action) assumptions. We then leverage tools from control theory to construct a rigorous measure of epidemic controllability that accurately reflects the control effort needed to stabilise a growing epidemic and the robustness to perturbations of a controlled epidemic. To maintain analytic tractability and as we focus on deriving fundamental insight into controllability, we only study constant R or r and model spread as a deterministic renewal process (see Methods). In later sections we discuss relaxations of these assumptions and show that our framework can measure the impact of variant dynamics, presymptomatic spread, superspreading, intervention lags and surveillance biases on controllability.

Epidemic models, feedback control and transfer functions

The renewal process is widely used to model acute infectious diseases such as COVID-19, Ebola virus disease and measles and says that new or incident infections at time t, $i(t)$ result from multiplying all active infections by R [27,28]. Past infections are active if they can still be transmitted, the probability of which is determined by the generation time distribution of the disease $w(t)$. The convolution $\int_0^t w(t-\tau)i(\tau)\,\mathrm{d}\tau$ captures this dynamic, weighting the past incidence $i(\tau)$ by the probability $w(t-\tau)$ that a primary infection causes secondary infections $t-\tau$ time units later. Imported infections from other regions $m(t)$ also contribute to onwards transmission and become included within the convolution [29].

We detail the uncontrolled renewal process with importations in **Eq. (M1)** of the Methods but here focus on extending it to include control. We define a generic control strategy as one that reduces infections to $\lambda(\tau) \leq i(\tau)$ so that $\int_0^t w(t-\tau)\lambda(\tau)\,\mathrm{d}\tau$ drives the controlled epidemic. The controller achieves this reduction by weighting past infections by a kernel $k(\tau)$. When this kernel only has mass at the present with $k(0) = k$, we get constant (memoryless) feedback control $\lambda(t) = ki(t)$. Generally, $0 \leq k(\tau) \leq 1$ as we aim to reduce infections. However, if the epidemic is already stable, we let $k(\tau) > 1$ to assess robustness to perturbations in infections i.e., we want to know the largest $k(\tau)$ for which critical or marginal stability is just achieved.

We can define the controlled renewal model with the expressions in **Eq. (1)**. These reduce to the standard renewal model by removing control i.e., by setting constant control at $k = 1$.

$$i(t) = m(t) + R\int_0^t \lambda(\tau)w(t-\tau)\,\mathrm{d}\tau\,, \qquad \lambda(t) = \int_0^t i(\tau)k(t-\tau)\,\mathrm{d}\tau. \quad (1)$$

We analyse **Eq. (1)** in the complex frequency s domain by applying Laplace transforms (see Methods) with $I(s)$, $M(s)$ and $W(s)$ as the transformed infection incidence, importations and generation time distribution. Because convolutions are products in this domain our controller satisfies $\Lambda(s) = K(s)I(s)$. We can represent these operations as a block diagram. We sketch this structure in **Fig 1A**, where we identify that, fundamentally, an epidemic involves a positive feedback loop between past and upcoming infections. Control aims to disrupt this loop.

Using this structure, we can define transmission dynamics by the properties of the closed loop transfer function (TF), $G(s) = I(s)M(s)^{-1}$, which describes how imports drive total incidence. We can write this in terms of the simpler loop TF $L(s)$, obtained by taking the product of blocks along the loop as $G(s) = \big(1 + L(s)\big)^{-1}$ [18]. The poles of $G(s)$ determine the dynamics and stability of the epidemic and are complex number solutions of $L(s) = -1 + j0$ (see Methods for details). We obtain these central TFs from **Eq. (1)** as in **Eq. (2)** below.

$$L(s) = -K(s)RW(s), \qquad G(s) = \frac{1}{1 - K(s)RW(s)}. \quad (2)$$

The uncontrolled epidemic TFs are recovered by setting $K(s) = 1$. We can interpret **Eq. (2)** by recognising that an unstable epidemic (at least one pole of $G(s)$ has positive real part) successively multiplies infections along the loop. This constitutes the positive feedback in **Fig 1A**. Interventions or control actions with magnitude $|K(s)| < 1$ reduce this positive feedback by interfering with the loop to attenuate the multiplication. Modification of the intrinsic epidemic dynamics $RW(s)$ by controller $K(s)$ within the loop achieves this goal. A stable epidemic (all poles of $G(s)$ have non-positive real parts) is also multiplicative, but infections reduce along the loop. We can apply $|K(s)| > 1$ as an amplifier of infections to study the robustness of the epidemic to any destabilising perturbations (e.g., increases in transmissibility).

There are two important corollaries of **Eq. (2)**. First, the poles of the epidemic TF $G(s)$ are the roots of the characteristic polynomial $1 - K(s)RW(s)$. Solving this (see Methods), we find the epidemic growth rate r is the dominant pole i.e., it is the major contributor to the dynamics of the system (see **Eq. (M4)**) and its variations reflect the impact of the controller $K(s)$. Second,

$K(s)$ directly modulates both R and the generation times. For constant control $K(s) = k$, the epidemic has an effective reproduction number of kR. These observations seemingly support the common paradigm of modelling interventions and assessing controllability directly from how R or r (or related parameters such as infectiousness) change [12].

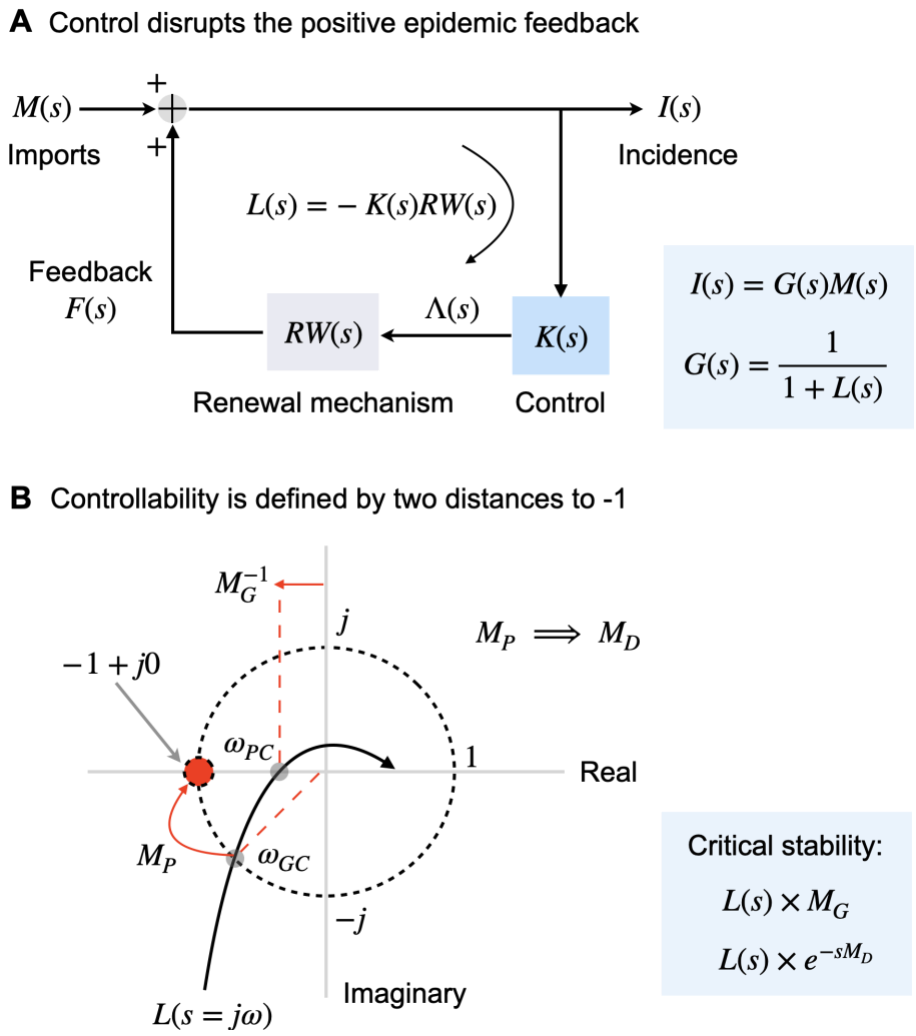

Fig 1: Epidemic architecture and controllability definition. Panel A shows that the renewal epidemic model is a positive feedback system in which a signal $F(s)$ is successively fed back across a loop and added to imported infections $M(s)$. The loop TF $L(s)$ (negative for a positive feedback loop by convention) determines the poles of closed loop TF $G(s)$, which completely expresses how imports combine with the epidemic dynamics to generate new infections $I(s)$. When we intervene or initiate control action, we disrupt the feedback loop via a controller $K(s)$. Panel B sketches a hypothetical polar plot of $L(s)$ for all complex frequencies $s = j\omega$. The closeness of this plot to the critical -1 point describes the controllability of the epidemic. At that point $L(s) + 1 = 0$, $G(s)$ blows up and the epidemic is critically stable. We can measure the

distance from -1 using the gain margin M_G, which defines how much we can scale $L(s)$ to get to -1 and either the phase M_P or delay M_D margin. Both define the angle we can rotate $L(s)$ to get to -1 but M_D expresses this in terms of the maximum system delay we can sustain (this is e^{-sM_D} in the s domain. Computation of these margins depend on phase ω_{PC} and gain ω_{GC} crossover frequencies (see text). Epidemics with larger margins are more controllable.

A framework for investigating epidemic controllability

However, these corollaries actually expose why these parameters are insufficient for defining controllability i.e., the effort required to stabilise an unstable epidemic, or the intensity of the perturbations required to destabilise a stable epidemic. Specifically, the difficulty of controlling the epidemic in real time also depends on its other poles (which may be oscillatory) [18] and only asymptotically are infections completely determined by the dominant pole r. Additionally, the assumption that $K(s)$ is constant and introduces no dynamics is unrealistic (e.g., isolation is known to reduce generation times [20,21]) and only likely true in very limited circumstances. We therefore need to account for transient and intervention-induced dynamics [23,30].

To investigate the implications of these corollaries, we propose a new framework for defining epidemic controllability, which adapts classical control theory as well as generalises and more rigorously quantifies the interpretation frequently ascribed to R or r. **Fig 1B** sketches the polar plot of $L(s)$ in the complex plane. We know from **Eq. (2)** and stability theory [17] that as $L(s)$ approaches $-1 + j0$ the closed loop $G(s)$ becomes critically stable i.e., it is on the verge of instability with $r = 0$. The gain M_{G} and delay M_{D} margins [18] precisely determine the distance of $L(s)$ from $-1 + j0$ (see Methods for how to compute these and related margins) [17].

For stable epidemics ($r < 0$ i.e., all $G(s)$ poles are in the left half of the complex plane), M_{G} and M_{D} respectively measure how much we can scale up or delay infections before the system becomes critical [31]. Accordingly, for unstable epidemics ($r > 0$ i.e., at least one $G(s)$ pole is in the right half plane) they quantify how much we must scale down or limit delay to stabilise an epidemic (assuming certain conditions [17]). Stability is rigorously defined as when $K(s) = 1$ in **Eq. (2)** then r matches $R - 1$ in sign and is the dominant $G(s)$ pole so $L(s) = -1$, $R = 1$ and $r = 0$ all correspond. There is an analogous association with the effective R and r when some control is acting ($K(s) \neq 1$). The crucial distinction we make is that the distance of $L(s)$ from -1 and not that of R from 1 or r from 0 is what actually determines controllability.

The margins we propose to measure this distance precisely and holistically characterise the essence of earlier notions of control effort by quantifying the magnitude and time by which we

must alter infections to attain the brink of stability. Computing these for **Eq. (2)**, we get $M_{\mathrm{G}} = |-K(s = j\omega_{\mathrm{PC}})RW(s = j\omega_{\mathrm{PC}})|^{-1}$ [17] with ω_{PC} as the frequency where the phase of $L(s)$ crosses a critical angle (see **Fig 1B**) and $|.|$ indicating magnitude i.e. $|\sigma + j\omega| \stackrel{\mathrm{def}}{=} \sqrt{\sigma^2 + \omega^2}$. Note that from the properties of distributions, $W(0) = 1$. We confirm this in the Methods for a universal class of phase-type generation time distributions [32], which include realistic models of $w(t)$ for many infectious diseases [15,28]. Accordingly, when $\omega_{\mathrm{PC}} = 0$, $M_{\mathrm{G}} = |-K(0)R|^{-1}$. Critical stability exists when $M_{\mathrm{G}} = 1$ (see **Fig 1B**). The control effort needed to define epidemic controllability, based on the gain margin, is therefore $K^* = |K(0)| = R^{-1}$.

We show in **Fig 2** for constant controllers applied to epidemics with various generation time distribution shapes (**Fig 2A**) that $\omega_{\mathrm{PC}} = 0$ is true and unique. For stable epidemics, we find $M_{\mathrm{D}} \to \infty$ (not shown but code provided in a link at the end of the paper). Consequently, under these conditions, controllability is completely established by the size of R^{-1} (**Fig 2C**), which correlates well with the Euclidean distance in the complex plane between $L(s)$ and -1 (inset). When the epidemic is unstable the gain margin is also set by R^{-1} but there may be ways of removing system lag that also define a dimension of control. However, if we apply a constant controller (so system lag does not change) with $k = \alpha R^{-1}$ with $\alpha < 1$, the controlled epidemic has an effective reproduction number of α and hence a controllability set by α^{-1}. Epidemics with the same controllability can still have diverse responses to imported infections (**Fig 2D**).

The dominant pole and hence the effective growth rate also shifts, from being the solution of $RW(s) = 1$, to that of $(kR)W(s) = 1$. As this equation is only scaled, the growth rate is now related to the effective reproduction number kR. For gamma distribution generation times with parameters (a, b) for example (see Methods), the growth rate changes from $b^{-1}(\sqrt[a]{R} - 1)$ to $b^{-1}(\sqrt[a]{kR} - 1)$ [15]. Consequently, if $\omega_{\mathrm{PC}} = 0$, we can completely describe the controllability of an epidemic using the size of reproduction numbers or growth rates. As growth rates are asymptotic (i.e., other poles decay in impact as $t \to \infty$) we can equally describe controllability from exponential growth models that approximate complex renewal processes (**Fig 2B**).

Our framework therefore supports the conventional definition that larger R or r indicates lower controllability but reveals that this requires $\omega_{\mathrm{PC}} = 0$ and that control is constant. Under these conditions we cannot destabilise the epidemic by perturbations that only add delay (or change phase). This holds across broad classes of fixed generation time distributions. We show next that our more generalised controllability definitions are necessary because these settings are strongly restrictive and unlikely in practice i.e., control often introduces dynamics (for example

by changing incubation periods, generation times and infectiousness durations). Further, we know that presymptomatic spread, superspreading, delays to interventions and surveillance biases all impact controllability and our definitions can rigorously unify these complexities.

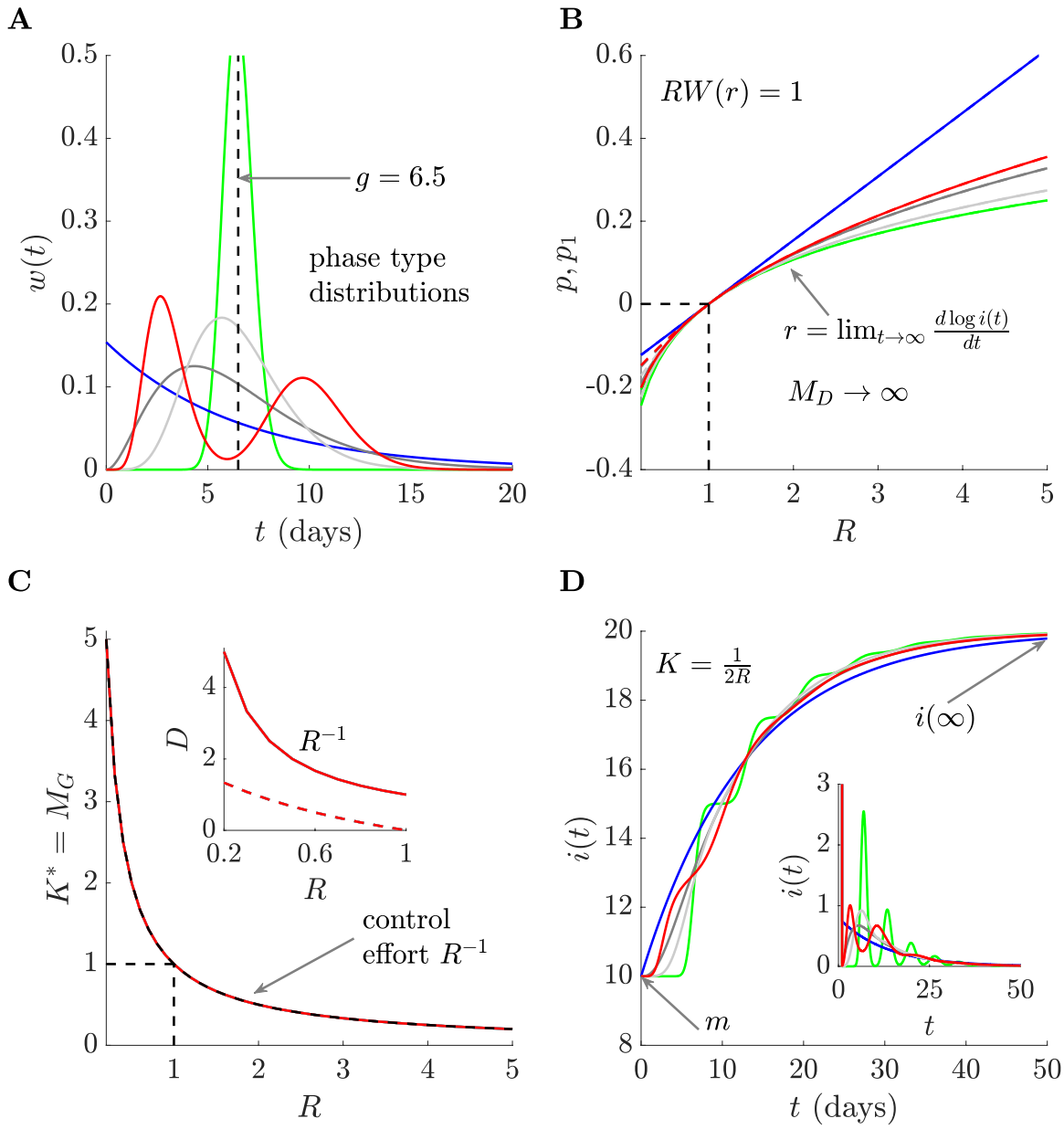

Fig 2: Epidemic controllability under ideal conditions. We assess controllability via gain and delay margins for epidemics subject to constant (non-dynamical) control $K(s) = k$ with phase crossover frequencies of 0 (see text). Panel A shows the generation time distributions $w(t)$ of simulated epidemics that we analyse, which have fixed mean generation time g (taken from COVID-19 [3]) but feature markedly different shapes. Panel B plots the growth rate r of these epidemics (colours match panel A), which is the dominant pole p (solid) of the resulting TFs $G(s)$. These strongly match the dominant pole p_1 (dashed) of an approximating epidemic

described by $i(t) = i(0)e^{p_1 t}$. Panel C plots the gain margin M_G or critical controller K^* that drives the system to the brink of instability (the delay margin M_G here is infinite). The K^* curves from every $w(t)$ exactly equal R^{-1}. These curves correlate well with D (inset), the Euclidean distance between $L(s)$ and -1. Panel D demonstrates that although controllability is the same, transient dynamics of infections may differ (they also depend on non-dominant system poles). We plot incident infections $i(t)$ in response to stable numbers (main) of imported infections ($m(t) = m$) and to a 1-day pulse (inset) of m imports (colours match panel A).

Problems with existing controllability definitions

Previously, we established conditions under which our generalised framework for assessing controllability reduces to the popular but informal definition applied in epidemiology. However, the conditions that allow this interpretation are strongly restrictive for two reasons. First, the only controller guaranteed to satisfy $|K(0)| = R^{-1}$ and have unique $\omega_{\mathrm{PC}} = 0$ is the constant $K(s) = k$. This controller seems unrealistic given that interventions not only scale infections but also change the distribution of generation times and other epidemiological quantities and hence induce additional dynamics (and poles in $G(s)$) [20–22]. Any realistic intervention (e.g., social distancing or contact tracing) likely scales infections and slows them from occurring.

We demonstrate this for the generation time distributions in **Fig 2A** using simple controllers of form $K(s) = \frac{1}{9}\frac{(1+g_1 s)}{(1+g_2 s)}$, which induce minimal dynamics by adding one pole. Here $K(s)$ can model interventions that change the effective reproduction number as well as the generation time properties of the epidemic. For example, if the uncontrolled epidemic has an exponential $w(t)$ with mean g_1 then $W(s) = \frac{1}{(1+g_1 s)}$ and the loop TF changes from $\frac{-R}{(1+g_1 s)}$ to $\frac{-R}{9(1+g_2 s)}$ i.e., the control scaled down infections by a factor of 9 and forced the mean generation time to g_2. This illustrates how controllers can realistically alter dynamics. Intervention-driven changes to generation times have been observed for malaria, COVID-19 and other diseases [19,20].

When $K(s)$ is applied to epidemics with $R = 4$, if $\omega_{\mathrm{PC}} = 0$, then $M_{\mathrm{G}} = |-K(0)R|^{-1} = \frac{9}{4}$. This controller is strongly stabilising (we can multiply infections by $\frac{9}{4}$ before facing critical stability), attenuating infections so that the effective reproduction number of the controlled epidemic is $\frac{4}{9}$. However, this standard interpretation is misleading and incomplete. Some controllers of this form cause $\omega_{\mathrm{PC}} \neq 0$. In **Fig 3A** we analyse one $K(s)$ that maintains $\omega_{\mathrm{PC}} = 0$ and another that violates this setting. For the first ($g_1 = 1, g_2 = 8$) the gain and delay margins as well as response to a stable input of $m(t) = 10$ infections over time is in accordance with **Fig 2**.

Strikingly, for the $\omega_{\mathrm{PC}} \neq 0$ case ($g_1 = 8, g_2 = 1$), the response is markedly different, showing oscillations and large infection peaks. The gain margin for these cases falls from $\frac{9}{4}$ to about 2 but, importantly, the delay margin for one of the $w(t)$ in **Fig 2A** becomes finite and small (not shown but $M_{\mathrm{D}} \approx 3.8$ days). This effect is more pronounced and smaller and finite M_{D} values occur for more types of $w(t)$ if we apply the slightly more complex control $K(s) = \frac{1}{9}\frac{(1+g_1 s)}{(1+g_2 s+s^2)}$ (not shown). Our generalised controllability formulation is necessarily more accurate, even in categorising infection scaling, and exposes important destabilising factors.

The finite delay margin is especially valuable as in reality we rarely apply interventions without some latency [13]. If control is applied after a 3.5-day delay, we obtain infection curves as in **Fig 3B**. There we observe that the red curve approaches instability and realise that there is a hard limit from M_{D} on how late we can respond to an epidemic if we want control to work. The importance of delays in epidemic control is a known issue [9,16] but it is rarely factored into epidemic controllability directly. Our $(M_{\mathrm{D}}, M_{\mathrm{G}})$ framework is comprehensive and exposes the pitfalls of measuring controllability only in terms of R or r (while not shown, the dominant poles and hence r in **Fig 3** are similar for both the finite and infinite M_{D} cases as well).

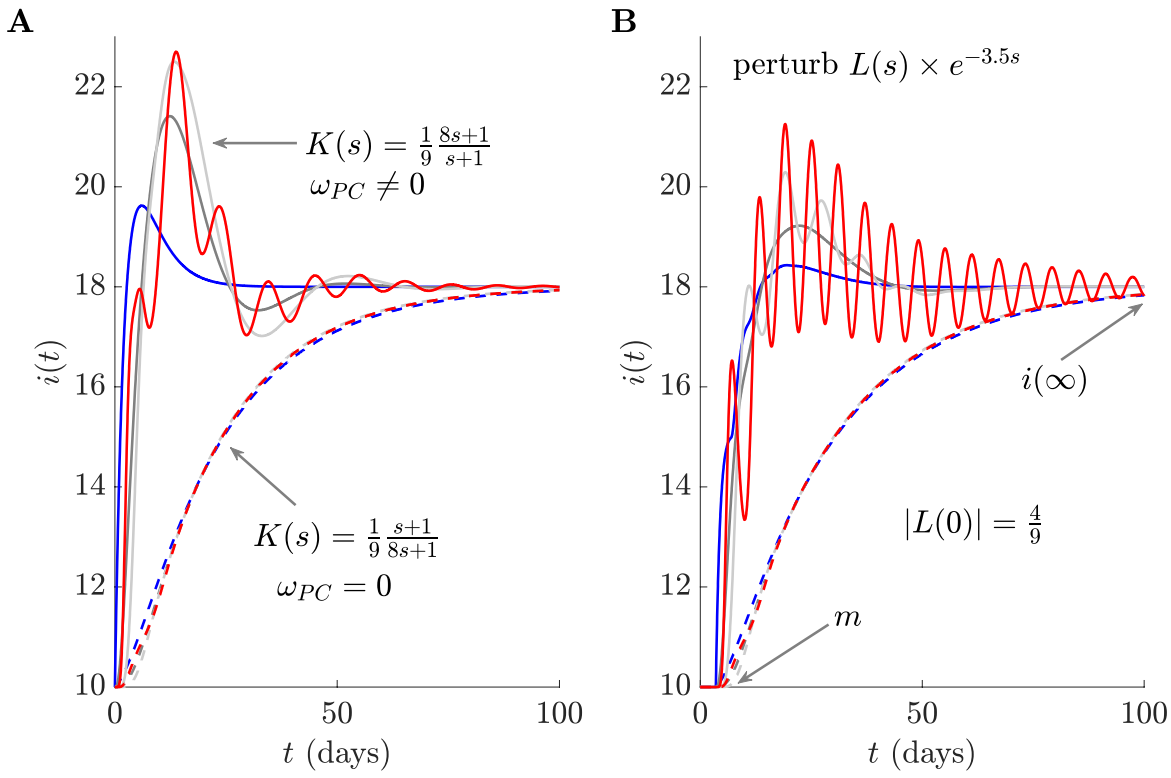

Fig 3: Controllers introducing additional dynamics. We simulate epidemics that are forced by a constant supply of imported infections ($m(t) = 10$). Panel A shows the resulting curves

of incidence for epidemics with generation time distributions from **Fig 2** (excluding the green one because this becomes unstable, curves match in colour) when non-constant control $K(s)$ is applied. There are major discrepancies among responses to interventions (controllers) that change the phase crossover frequency ω_{PC} (solid) and those maintaining $\omega_{PC} = 0$ (dashed). The former show salient transients that disrupt controllability and feature smaller gain M_G and finite delay M_D margins (conventional interpretations expect $M_G = \frac{9}{4}$, $M_D \to \infty$). The steady state incidence $i(\infty)$ remains, however, unchanged for all our controllers. In panel B we apply a perturbation of a 3.5-day delay ($e^{-3.5s}$ in the s domain). This pushes the curve from panel A with finite $M_D \approx 3.8$ days towards instability. The value of a two-margin description is clear.

The second major problem with conventional definitions of controllability is that they are not easily computed, interpreted or compared when practicalities such as presymptomatic spread, superspreading, variant dynamics and surveillance imperfections (e.g., reporting delays and incomplete case ascertainment) occur [1,24]. In the next two sections, we expand our models and demonstrate that the $(M_{\mathrm{D}}, M_{\mathrm{G}})$ framework presents a unified and interpretable approach to measuring and monitoring epidemic controllability under all of these complexities. No matter the specific model structure, the boundaries of controllability specified by our $(M_{\mathrm{D}}, M_{\mathrm{G}})$ pair are directly comparable and possess exactly the same interpretation as in **Fig *1***.

Surveillance limitations and presymptomatic spread

Until now we have assumed that we can observe and apply control to all new infections. This is unrealistic as commonly we can only count cases or deaths, which are delayed and scaled versions of infections [33,34]. Here we generalise **Eq. (1)** and **Eq. (2)** to include these effects. We denote the proportion of infections that we observe as cases by probability $0 \leq \rho \leq 1$ and model the latency in observing these cases with a distribution $h(t)$. Our controller acts on the incidence of cases $c(t)$, and $i(t) - c(t)$ infections remain unobserved. This yields **Eq. (3)**.

$$\lambda(t) = \int_0^t c(\tau)k(t-\tau)\,\mathrm{d}\tau, \quad c(t) = \rho\int_0^t i(\tau)h(t-\tau)\,\mathrm{d}\tau. \quad (3)$$

The unobserved infections continue to propagate the epidemic as they remain uncontrolled. We therefore construct the combined renewal model of **Eq. (4)** below.

$$i(t) = m(t) + R\int_0^t (i(\tau) - c(\tau))w(t-\tau)\,\mathrm{d}\tau + R\int_0^t \lambda(\tau)w(t-\tau)\,\mathrm{d}\tau\,. \quad (4)$$

This collapses into **Eq. (1)** when reporting is perfect i.e., $\rho = 1$ and $h(t)$ has all its probability mass at the present ($h(0) = 1$) so that $c(t) = i(t)$.

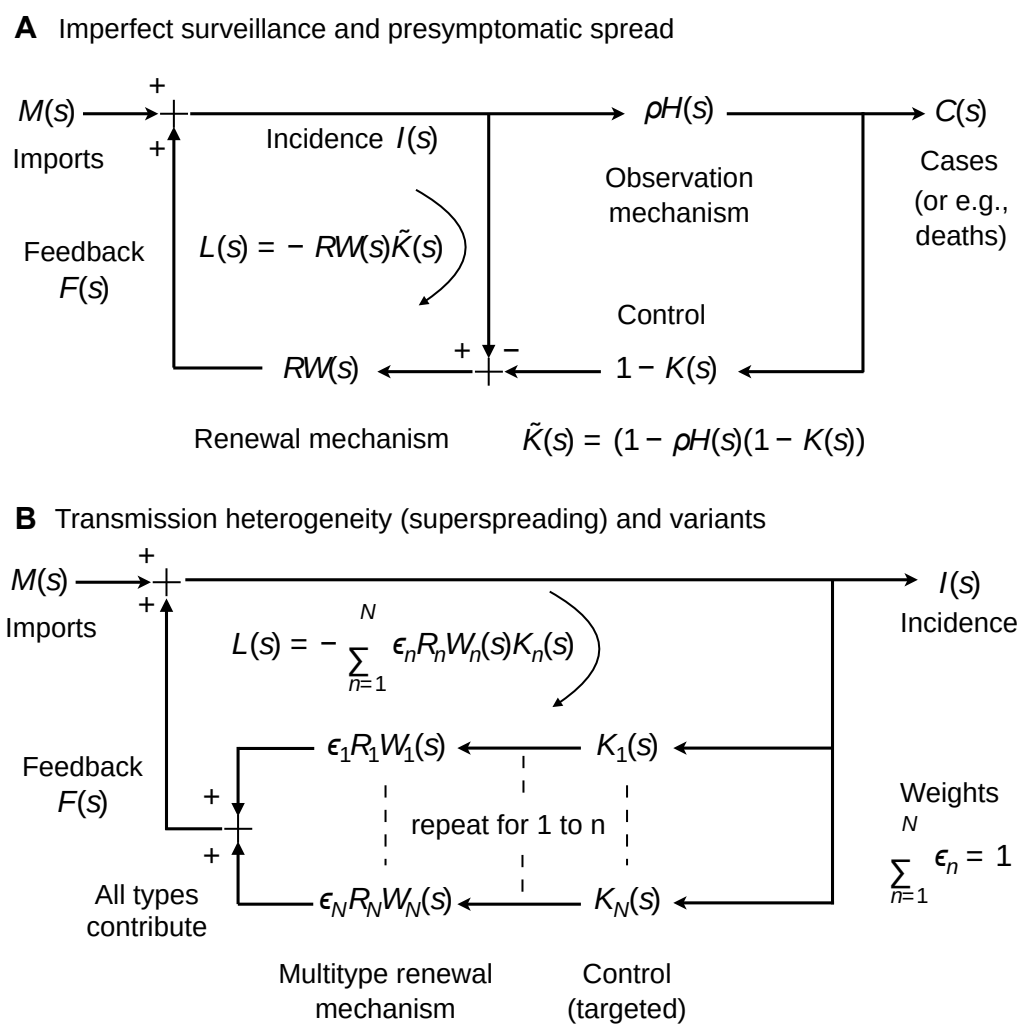

Fig 4: Generalised controlled renewal model architectures. Panel A illustrates the block diagram of a renewal model for which only a portion of the new infections $I(s)$ are observable and hence can be controlled by $K(s)$. This portion $C(s)$ may model cases, deaths or any other time series that is mediated by a scale factor ρ and a lag distribution $H(s)$. This architecture represents imperfect surveillance mechanisms or presymptomatic spread. Panel B shows the structure of a multitype, controlled renewal model describing N infectious types or stages with diverse reproduction numbers R_n and generation time distributions $W_n(s)$. The weight ϵ_n is the fraction of new infections of type n. This architecture models transmission heterogeneity including superspreading, co-circulating variants and diseases with multiple routes for spread. Both panels have closed loop TFs $G(s) = I(s)M(s)^{-1} = \big(1 + L(s)\big)^{-1}$, with loop TF $L(s)$ as described. See main text for details on how $K(s)$ and the $K_n(s)$ define controllability.

We again take Laplace transforms of **Eq. (3)** and **Eq. (4)** to obtain our key TFs for evaluating epidemic controllability in **Eq. (5)**. We illustrate this architecture in **Fig 4A** and observe that we also obtain TFs for the observed cases easily since $C(s)M(s)^{-1} = \rho H(s)G(s)$.

$$L(s) = -RW(s)(1 - \rho H(s)(1 - K(s)), \qquad G(s) = \frac{1}{1 + L(s)}. \quad (5)$$

When $K(s) = 1$ in **Eq. (5)** we recover the uncontrolled epidemic TFs (see **Eq. (M1)**). Perfect surveillance means $\rho H(s) = 1$ and reverts **Eq. (5)** to **Eq. (2)**. If we instead perform control on another proxy of infections, for example deaths or hospitalisations, then ρ is the proportion of infections that lead to mortality or hospitalisation (e.g., for the incidence of deaths this includes the infection fatality ratio and the proportion of deaths that are observed). The distribution $h(t)$ models the lag from becoming infected to mortality or being admitted to hospital [34,35].

This formulation equally models presymptomatic and asymptomatic spread, with $h(t)$ defining the delay between infection and presenting symptoms and ρ as the proportion of infections that never become symptomatic. We compute our $(M_{\mathrm{D}}, M_{\mathrm{G}})$ pair to assess how these differing transmission and surveillance characteristics impact controllability. **Eq. (5)** includes all the key controllability factors outlined in [1] and describes targeted interventions such as quarantine, contact tracing or isolation but not widescale lockdowns (we only control observed infections). Lockdowns and other non-selective interventions conform more closely to **Eq. (2)** as they act indiscriminately on all infections, including those we never observe.

We know from earlier that critical stability is achieved when $L(s) = -1$. We substitute this into **Eq. (5)** and find that our control needs to satisfy the left side of **Eq. (6)**. As a constant $K(s) = 0$ represents the maximum possible control effort (i.e., all observed infections are suppressed completely), we insert this condition and rearrange to derive the threshold on the right side of **Eq. (6)**, outlining the requirements on the surveillance noise or level of presymptomatic spread for the epidemic to just be controllable. A smaller $|\rho H(s)|$ causes loss of controllability and provides evidence that widescale interventions or surveillance improvements are needed. The relations of **Eq. (6)** are only required to hold at the $s = j\omega$ satisfying $L(s) = -1$.

$$K(s) = 1 - \frac{1 - \left(RW(s)\right)^{-1}}{\rho H(s)}, \qquad |\rho H(s)| \geq \left|1 - \frac{1}{RW(s)}\right|. \quad (6)$$

If $\omega_{\mathrm{PC}} = 0$ then this requirement is met at $\rho \geq 1 - R^{-1}$, as $W(0) = H(0) = 1$. This matches the critical contact tracing efficiency derived in [2] and the presymptomatic condition of [1] and

confirms how our methodology generalises more conventional notions of controllability (it also generalises the herd immunity threshold). **Eq. (6)** verifies that we need both margins because $\omega_{\mathrm{PC}} = 0$ is not guaranteed here, even if control is constant. The temporal impact of imperfect surveillance or presymptomatic spread via $H(s)$ means that the dynamics leading to situations as in **Fig 3** always exist. Transient dynamics are crucial and unavoidable.

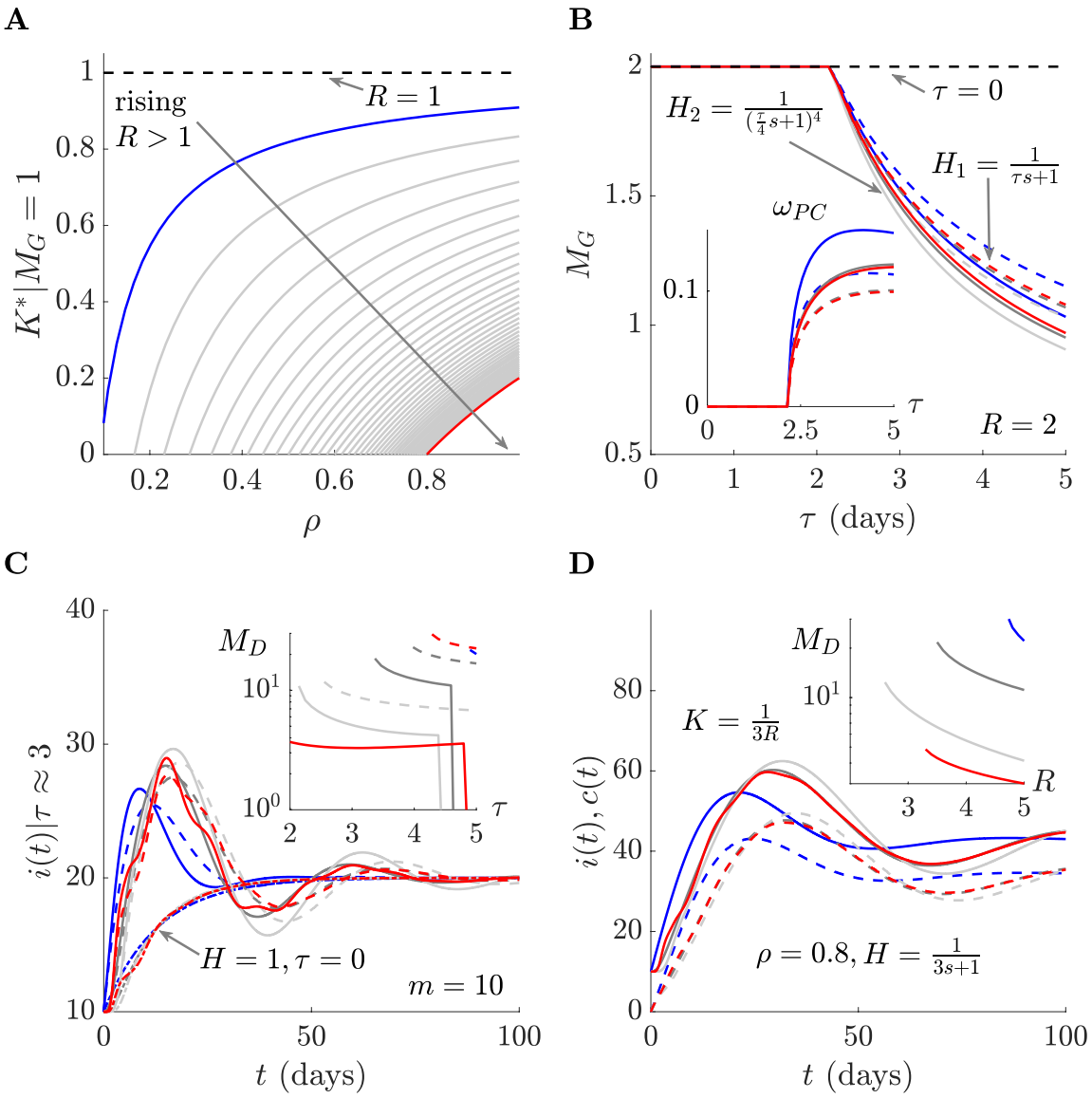

Fig 5: Surveillance noise and presymptomatic spread. We investigate how imperfect case reporting, or equivalently presymptomatic spread, limits the controllability of epidemics using our (M_D, M_G) framework. Panel A shows for curves of constant $R \geq 1$ (rising from blue to red, which is at $R = 5$) how the reporting rate or proportion of symptomatic infections, ρ, reduces controllability. Smaller ρ requires more control effort to attain critical stability i.e., a smaller K^* is needed for a gain margin $M_G = 1$. There is no reporting delay or presymptomatic distribution

in this analysis so $H(s) = 1$. Panel B sets $\rho = 1$ and investigates the influence of two $H(s)$ distributions, H_1 (dashed) and H_2 (solid) modelling exponential and gamma distributions. Both have mean lag τ, $R = 2$ and a controller applied that achieves $M_G = 2$, if the phase crossover frequency $\omega_{PC} = 0$. We find that as τ increases $M_G < 2$ indicating a decline in controllability. This results from ω_{PC} increasing above 0 (inset). Colours in this and panels C-D match the generation times modelled from **Fig 2A** (excluding the green). Panel C confirms $H(s)$ causes the delay margin M_D to become finite (inset, dashed or solid corresponding to panel B). This reduced controllability is visible from the peaked, oscillatory response in new infections $i(t)$ for a constant number of imports $m(t)$ (main). This effect is similar to that in **Fig 3**. Here dot-dashed lines plot the response in the absence of $H(s)$. Panel D shows the combined influence of lags and under-reporting given the constant controller of $K = \frac{1}{3R}$. The inset demonstrates how M_D falls with R and the main shows the infection (solid) and case $c(t)$ (dashed) epidemic curves in response to constant imports (colours match generation time distributions).

We verify this point in **Fig 5**, showing how controllability depends on ρ and $H(s)$. We first set $H(s) = 1$ and explore the controller gain needed to get $M_{\mathrm{G}} = 1$, which sets critical stability. In the absence of under-reporting, we have $\rho = 1$ and $K^* = R^{-1}$ for any R. **Fig 5A** shows that our required K^* substantially deteriorates, highlighting that we need additional control effort to stabilise the epidemic as ρ decreases. When $K^* = 0$, the epidemic is no longer controllable by these targeted interventions. If we cannot improve surveillance quality or, equally diminish asymptomatic spread (so ρ rises), then population-level controls are warranted. Strikingly, at $R = 5$ (red) , we cannot control the epidemic unless more than 80% of all new infections are observed (sampled) or symptomatic. **Eq. (6)** defines fundamental limits on controllability.

In **Fig 5B** and **Fig 5C** we assume perfect reporting and test the influence of delays in reporting or equivalently lags in infections becoming symptomatic. We investigate two $h(t)$ distributions, $H_1(s)$ and $H_2(s)$ in the frequency domain, with results respectively as dashed or solid. These, model exponential and gamma distributed delays with means τ. We apply controls that force $M_{\mathrm{G}} = 2$ when $\omega_{\mathrm{PC}} = 0$ but find in **Fig 5B** that our gain margin declines with τ. This occurs as $\omega_{\mathrm{PC}} > 0$ (inset). **Fig 5C** further shows that the delay margin M_{D} becomes finite, decaying with τ (inset). Hence, $H(s)$ reduces both the scaling and delays that the controlled epidemic can robustly support. Incident infections $i(t)$ display oscillatory dynamics with substantial peaks (main). This contrasts the plots featuring no delay i.e., $\tau = 0$ (dot-dashed).

Colours indicate the $w(t)$ from **Fig 2A** underlying results in **Fig 5B**, **Fig 5C** and **Fig 5D**. On its own, $H(s)$ substantially reduces our controllability. At $\tau \geq 4$ we find that $M_{\mathrm{D}} \to 0$ (with also $M_{\mathrm{G}} < 1$) signifying that the epidemic is now unstable. Epidemics with larger τ are necessarily uncontrollable. We combine both ρ and $H(s)$ in **Fig 5D** but vary R and apply a strong controller that scales down cases by $\frac{1}{3R}$. Even for this constant control we find a finite M_{D} that declines with R (inset) and large amplitude oscillations in $i(t)$ (solid, main). We also plot the observed cases $c(t)$ (dashed), which are the fraction of infections we can control. Both of the $(M_{\mathrm{D}}, M_{\mathrm{G}})$ pair are therefore critical to accurately quantifying epidemic controllability.

Superspreading, variants and multiple infector types

Our $(M_{\mathrm{D}}, M_{\mathrm{G}})$ framework can also evaluate the controllability of epidemics that are composed of multiple infectious types or transmission routes. This models superspreading, co-circulating variants and pathogens with multiple pathways of spread. We unify these multitype epidemics using the renewal process of **Eq. (7)**, which features N distinct types or pathways.

$$i(t) = m(t) + \sum_{n=1}^{N} R_n \int_0^t \lambda_n(\tau) w_n(t-\tau)\,\mathrm{d}\tau, \quad \lambda_n(t) = \int_0^t \epsilon_n i(\tau) k_n(t-\tau)\,\mathrm{d}\tau. \qquad (7)$$

We denote the reproduction number, generation time distribution and controller of the n^{th} type with subscript n. The parameters ϵ_n define the proportion of incidence associated with the n^{th} type and $\sum_{n=1}^{N} \epsilon_n = 1$. By dividing control into N functions, we allow for type-specific control. This includes non-targeted control (all $k_n(\tau)$ are the same) and situations where some types are uncontrolled (those $k_n(\tau) = 1$), perhaps due to being unobservable.

Specialisations of **Eq. (7)** can model superspreading or transmission heterogeneity (e.g., we set $N = 2$, $R_1 \gg R_2$, $\epsilon_1 = \frac{1}{5}$ and $\epsilon_2 = \frac{4}{5}$ to describe cases where 20% of new infections have substantially larger transmissibility [36]), pathogenic variants with differing transmissibility and generation times (e.g., with N as the number of co-circulating variants, although we assume early growth so that the ϵ_n are fixed [37,38]) and diseases with diverse transmission pathways (e.g., Ebola virus disease has sexual and non-sexual pathways with distinct $w_n(t)$ [39]). These models do not include explicit interaction among types (though all types compose $i(t)$) as this requires additional cross-type reproduction numbers and auxiliary data (e.g., contact matrices) [40]. Such extensions are possible by altering the integral within the sum in **Eq. (7)**.

We take Laplace transforms of **Eq. (7)** to construct **Eq. (8)**, which is amenable to our gain and delay margin controllability analyses. We sketch the architecture of this model in **Fig 4B**.

$$L(s) = -\sum_{n=1}^{N} \epsilon_n R_n W_n(s) K_n(s), \qquad G(s) = \frac{1}{1 + L(s)}. \quad (8)$$

Using the fact that $W_n(0) = 1$, we find that if $\omega_{\mathrm{PC}} = 0$ then $M_{\mathrm{G}} = |-\sum_{n=1}^{N} K_n(0)\epsilon_n R_n|^{-1}$. We can therefore scale the epidemic by a quantity that is a weighted sum of control, reproduction numbers and proportions of the contributing infectious types. As we showed in above sections, this condition is only likely to be met if every controller is constant (at which also $M_{\mathrm{D}} \to \infty$). If controllers introduce dynamics, which is realistic, then we expect effects similar to **Fig 3**.

Eq. (8) provides the flexibility to investigate several controllability problems. We focus on two questions about the limitations of targeted control for heterogeneous populations. We let $N = 2$ and assume that $R_1 \geq R_2$ so that type 1 represents individuals with the more transmissible variant or superspreading nodes. We consider non-selective control where $K_1(s) = K_2(s) = K(s)$ and targeted control, in which only one type is controlled. We only target type 1, which is more transmissible, so type 2 is uncontrolled and $K_2(s) = 1$. Our first question asks under what conditions the targeted approach, which is often proposed as an efficient control scheme [11,36], fails to suppress the overall epidemic, making non-selective control unavoidable.

For this two-type epidemic $L(s) = -(\epsilon_1 R_1 K_1(s) W_1(s) + \epsilon_2 R_2 W_2(s))$ for targeted control and $-K(s)(\epsilon_1 R_1 W_1(s) + \epsilon_2 R_2 W_2(s))$ for non-selective control, with $\epsilon_2 = 1 - \epsilon_1$. In both cases, $\omega_{\mathrm{PC}} = 0$ and $W_1(0) = W_2(0) = 1$ (see Methods). If we only apply constant controllers, then $M_{\mathrm{D}} \to \infty$ and controllability is exclusively defined by the values of M_{G}, which are computed as $|\epsilon_1 R_1 K_1(0) + \epsilon_2 R_2|^{-1}$ and $|K(0)|^{-1}|\epsilon_1 R_1 + \epsilon_2 R_2|^{-1}$. To attain some specific M_{G}, we require $K_1(0) = (M_G^{-1} - \epsilon_2 R_2)(\epsilon_1 R_1)^{-1}$ and $K(0) = M_G^{-1}(\epsilon_1 R_1 + \epsilon_2 R_2)^{-1}$. We can combine these relations to get the left side of **Eq. (9)**, which shows how much smaller $K_1(0)$ needs to be than $K(0)$ i.e., how much more targeted control effort is required to attain our desired M_{G}.

$$K_1(0) = K(0) - \left(\frac{\epsilon_2 R_2}{\epsilon_1 R_1}\right)\big(1 - K(0)\big), \qquad \epsilon_1 \geq 1 - \frac{1}{M_G R_2}. \quad (9)$$

We plot the control efforts $K^* = K(0)$ and $K_1^* = K_1(0)$ from both strategies that are necessary to achieve critical stability ($M_{\mathrm{G}} = 1$) in **Fig 6A**. There we observe the limits of targeted control as a critical ϵ_1 cut-off (dashed vertical). This follows from the positivity constraint $0 \leq K_1(0) < 1$, where 1 is no control and 0 defines perfect control, in which type 2 infections are neutralised.

We derive this for any desired gain margin on the right side of **Eq. (9)**. Interestingly, this cut-off does not depend on R_1 and, if $M_G = 1$, it indicates that targeted control only works when the proportion of superspreading nodes or type 1 variants is above $1 - R_2^{-1}$. This procedure is easily generalised to N-type epidemics where we can control a subset χ of the types. The controllability cut-off then requires the uncontrolled proportion $\left|\sum_{n\notin\chi} \epsilon_n R_n W_n(s)\right| \leq M_G^{-1}$.

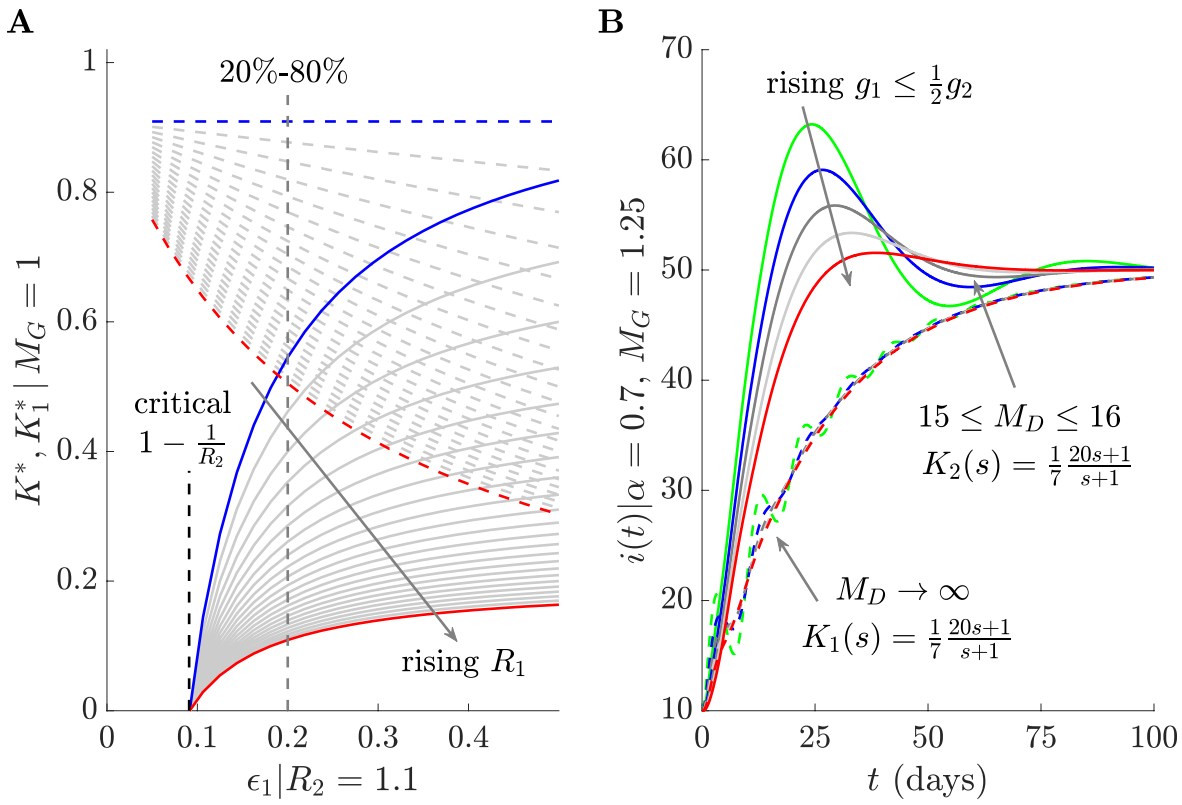

Fig 6: Targeted control in multitype epidemics. We explore controllability and performance limits for epidemics that involve two distinct types, modelling superspreading or co-circulating variants. Panel A plots the constant control effort necessary for critical stability ($M_G = 1$) under a non-selective strategy with controller K^* that reduces infections of both types (dashed) and a targeted strategy with controller K_1^* that only reduces infections of type 1 (solid), which has larger transmissibility $R_1 \geq R_2 = 1.1$. For both strategies, we vary the proportion of type 1, ϵ_1, and curves are for increasing R_1 from blue (1.1) to red (5.5) with intermediate values in grey. We use a vertical grey line to show the ϵ_1 for the commonly used 20-80 superspreading rule. Targeted control requires substantially more effort (as it must also account for the uncontrolled type 2), and the epidemic is uncontrollable if ϵ_1 is smaller than the critical black line (see **Eq. (9)**). Panel B considers targeted controllers that introduce dynamics and only apply $K_1(s)$ or $K_2(s)$ to reduce either type 1 or 2 infections. We fix $\epsilon_1 R_1 = \epsilon_2 R_2$ so there is no difference in

how the types contribute to overall transmissibility and both controllers lead to the same $M_G > 1$. We show how infections $i(t)$ change due to both schemes (dashed and solid respectively), where type 1 is the faster variant with mean generation time $g_1 \leq g_2 = 8$ days. Targeting the slower type 2 leads to worse performance and is sensitive to g_1 (curves are not grouped).

Our second question relates to the interaction between differing generation times of the types and induced controller dynamics. We consider targeted control of either type or variant with type 1 having smaller mean generation time and hence being faster than type 2 i.e., $g_1 < g_2$. We set $\epsilon_1 R_1 = \epsilon_2 R_2 = \alpha$ to remove any relative transmissibility advantage between the types. Consequently, variations in the infections caused by the types emerge from their generation time distribution differences. Targeted control applies non-constant control $K_1(s)$ exclusively to type 1 or $K_2(s)$ exclusively to type 2, yielding loop TFs $L(s) = -\alpha(K_1(s)W_1(s) + W_2(s))$ and $-\alpha(W_1(s) + K_2(s)W_2(s))$. Because the controller induces additional dynamics, we are neither guaranteed $\omega_{\mathrm{PC}} = 0$ nor $M_{\mathrm{D}} \to \infty$ and must evaluate the complete $(M_{\mathrm{D}}, M_{\mathrm{G}})$ pair.

We compute these margins and dynamical responses to constant importations in **Fig 6B** for a range of fast type 1 generation time distributions $w_1(t)$ and a fixed (slow) type 2 distribution $w_2(t)$. Although M_{G} is the same for both schemes, controlling type 2, which may occur when transmission chains of slower variants are easier to interrupt, yields worse performance. The overshoots and oscillations are also accompanied by a finite M_{D}, highlighting that neglecting the faster variant can potentially reduce robustness of the controlled epidemic to perturbations or equally reduce controllability below what we may expect from conventional measures based on reproduction numbers or asymptotic growth rates. For certain controllers (not shown) we also find that $\omega_{\mathrm{PC}} > 0$ can occur and reduce M_{G} for either targeted scheme. This underscores the importance of our two-margin solution to understanding controllability.

Discussion

Measuring the controllability of an infectious disease subject to various intervention options is a fundamental contribution of mathematical modelling to epidemiology [4,12]. However, there exists no rigorous and precise definition of what controllability means [8,10] and studies have highlighted a need for robust analytical frameworks to better appraise the impacts of targeted and reactive interventions [1]. Currently, the distance from the epidemic threshold of *R*=1 or *r*=0 is frequently used to measure controllability. Here we have demonstrated that this notion of controllability, although reasonable, is idealistic and likely misleading because neither *R* nor *r* completely and unambiguously measures distance from stability. We proposed an alternative

and analytic definition of this distance by reformulating the disease transmission process as a positive feedback loop and leveraging results from control engineering.

We derived epidemic transfer functions to describe the dynamics of this loop and model how stabilising interventions interrupt and attenuate this positive feedback (or for stable epidemics we test robustness to perturbations that amplify infections along this loop). This allowed us to develop stability margins that accurately measure the distance from stability (**Fig *1***) in units of the scale and speed of required control efforts. The gain and delay margins are key metrics from control engineering [17,18], a field that studies stability and feedback problems across dynamical systems. Although there is increasing interest in using tools from this field to better understand infectious disease spread [9,41–44], our study appears to be among the earliest to construct margins for epidemics and appraise existing notions of disease controllability.

Our central contribution is a flexible method for quantifying epidemic controllability that is both computable and easily interpreted across many salient characteristics of infectious diseases. This is important for three main reasons. First, R and r can lose their meaning or comparability as threshold parameters when characteristics such as superspreading and multitype spread are included [24,45]. Second, for a given transmission model there can be numerous ways of constructing and defining valid epidemic thresholds and these are not always consistent when assessing interventions [25,45,46]. For example, when interventions change generation times then we can find situations where r increases yet R decreases [47]. Third, earlier frameworks were unable to directly include reactive or feedback effects within their measures and did not account for how the implementation of interventions might modify effectiveness.

In contrast, our gain and delay margins maintain their interpretation, validity, uniqueness and comparability across complex disease models and explicitly reflect feedback loops intrinsic to transmission and intervention. These properties allowed controllability to be measured across realistic generation time distributions (**Fig *2***), constraints on interventions (**Fig *3***), surveillance imperfections (**Fig *5***) and transmission heterogeneities (**Fig 6**). Principal insights emerging from this unified approach were that (i) R and r only track controllability in restrictive settings where interventions do not alter temporal disease characteristics and are applied instantly, (ii) sharp thresholds of controllability exist due to presymptomatic spread, superspreading, delays and under-reporting and co-circulating variants that generalise $1-1/R$ type results and (iii) the delay margin is crucial because lags along feedback loops (from both intervention delays and surveillance biases) can destabilise epidemics that are conventionally deemed controlled.

While our approach rigorously incorporates many realistic epidemic complexities and extends earlier frameworks [1,8,10], it depends on several simplifying assumptions, which we made to

ensure tractability and to extract general insights. Specifically, our analysis uses deterministic renewal models and assumes constant *R* or *r*. Although some or all of these assumptions are common to seminal studies and recent works on controllability thresholds [1,38], the influence of stochasticity in disease transmission can be substantial [7,11]. We recommend computing our margins as an initial step to quantifying the impacts of interventions, which can then guide the running of more complex stochastic models. Our margins are also only well-defined for linear systems, which include any epidemics represented by renewal models with constant *R*. If *R* varies on the timescale of interventions or involves non-linear effects such as saturation, this assumption may be invalidated. However, we can use piecewise-constant transmissibility approximations and fit renewal models to each piece, to partially circumvent this issue.

Moreover, we examine linear and reactive control actions only (i.e., convolutions of kernels with past infections). This improves upon many studies, where controllers simply multiply and reduce *R* or *r* but may not model other notable types of interventions, such as those reducing infections due to non-linear switching triggers or those that completely ignore feedback signals in favour of predetermined action [30,48]. Understanding the relative benefits of these different strategies is an ongoing area of research. Last, we comment that controllability here focussed on intrinsic epidemic dynamics and neglected the costs of actions. Including how these costs further constrain the realisable limits of controllability, as well as incorporating key behavioural effects within our feedback loops are the future directions of this research.

In summary, we demonstrate that controllability is only completely and accurately measured by the distance of the loop transfer function $L(s)$ from -1. This generalises and improves upon the conventionally used distances of *R* from 1 or *r* from 0, but still admits interpretable margins or safety factors that quantify how much we can scale infections or delay interventions to attain critical stability. This allows us to better evaluate when targeted interventions are insufficient and hence when non-selective controls such as lockdowns are justified from the viewpoint of curbing transmission. We find that targeted controls fail when the dynamics of the unobserved or untargeted infectious population, together with constraints on surveillance and intervention implementation cross margin thresholds that are analytically derived from our framework.

Methods

Renewal models and transfer functions

The renewal branching process [27] is a fundamental and popular infectious disease model that has been applied to describe epidemics of COVID-19, pandemic influenza, Ebola virus disease, measles, SARS and many others [12,28]. This model defines how incident infections at time t, $i(t)$ depend on the disease reproduction number R and incidence at earlier times

$i(\tau)$ ($\tau \leq t$, with a limit infinitesimally before t) via the autoregressive relationship in the left of **Eq. (M1)**. We assume that R is constant over the period that we investigate.

$$i(t) = m(t) + R\int_0^t i(\tau)w(t-\tau)\,\mathrm{d}\tau\,, \qquad I(s) = \frac{1}{1-RW(s)}M(s). \quad \text{(M1)}$$

Eq. (M1) also includes infections $m(t)$ that have been imported or introduced into our region of interest. The kernel of the autoregression is determined by $w(t-\tau)$, which is the probability of an infection being transmitted after a lag of duration $t-\tau$. The set of coefficients $\{w(\kappa), \kappa \geq 0\}$ composes the generation time distribution of the disease [15].

Since **Eq. (M1)** defines a linear model, we can analyse it in the frequency or s domain using Laplace transforms e.g., $I(s) \stackrel{\text{def}}{=} \int_0^\infty i(t)e^{-st}\,\mathrm{d}t$ is the transform of $i(t)$. We get the right side of **Eq. (M1)** after some algebra with capitalised forms as the transformed version of variables from the time domain. We can visualise this formulation from the block diagram of **Fig 1**. The ratio $G(s) = I(s)M(s)^{-1}$ defines an epidemic transfer function. The roots of the characteristic polynomial $1-RW(s)$ are the poles of the system and completely define the stability of the epidemic [17]. Characterising these poles for generalised versions of $G(s)$ that model different control schemes and intervention practicalities forms the subject of the main text.

The epidemic dynamics therefore depend explicitly on both the reproduction number and the generation time distribution, which are two of three key quantities often used to describe the transmissibility of infectious diseases. The third, which is the asymptotic growth rate of new infections $r = \lim_{t\to\infty} \frac{\mathrm{d}\log i(t)}{\mathrm{d}t}$, also emerges from our formulation. Since $W(-s)$ is equivalent to the moment generating function of the generation time distribution evaluated at s, we know from [15] that $W(r) = R^{-1}$. Interestingly, this is also the dominant pole of $G(s)$. We convert the growth rate into $t_R = \log \sqrt[r]{R}$, the time it takes for infections to (asymptotically) grow (or decline) by a factor of R. If we replace R by 2 we obtain the epidemic doubling time.

Generation time distributions and Laplace transforms

The dynamics of infectious diseases are largely determined by the generation time distribution since $W(s)$ is the only non-constant component of the transfer function in **Eq. (M1)**. We model $W(s)$ as a phase-type distribution, which is an expansive class built from combinations and convolutions of exponential distributions that can approximate any distribution [32]. Erlang, exponential, deterministic (degenerate) and bimodal distributions that we consider in the main

text are all special phase-type distributions conforming to the relations in **Eq. (M2)**. Erlang (or related gamma), deterministic and exponential distributions are often used to model influenza, measles and COVID-19, among others [3,15,27,28]. Multimodal and mixture distributions are commonly applied to diseases featuring multiple stages (which may even involve vectors) or pathways of transmission, for example malaria and Ebola virus disease [19,39].

$$W(s) = \boldsymbol{\alpha}(s\mathbf{I} - \mathbf{T})^{-1}(-\mathbf{T}\mathbf{1}'), \qquad W(0) = \boldsymbol{\alpha}\mathbf{1}' = 1. \quad \text{(M2)}$$

We use bold to denote vectors or matrices. In **Eq. (M2)** $\mathbf{T}$ is a n^2 matrix of the transition rates among the n distribution states, $\mathbf{I}$ the n^2 identity matrix, $\boldsymbol{\alpha}$ is a row vector of length n summing to 1 (providing weights to the states) and $\mathbf{1}'$ is the transpose of a row vector of n ones. Mixtures of phase-type distributions remain phase-type and we see that their Laplace transforms are always 1 at $s = 0$ (equivalent to the fact that probability distributions integrate to 1). We find this basic property important for computing controllability in the main text.

For a mean generation time of g we can obtain an Erlang distribution with shape a and scale b such that $g = ab$ by setting $n = a$, $\boldsymbol{\alpha} = [1\ 0\ \ldots\ 0]$, and $\mathbf{T}$ as a matrix with non-zero elements of $\mathbf{T}_{\kappa\kappa} = -b^{-1}$ and $\mathbf{T}_{\kappa\kappa+1} = b^{-1}$ for $1 \leq \kappa \leq n$. As a result, we obtain $W(s)$ as in **Eq. (M3)**.

$$W(s) = \frac{1}{(bs+1)^a}, \qquad 1 - RW(s) = \frac{(bs+1)^a - R}{(bs+1)^a}. \quad \text{(M3)}$$

We also find the characteristic polynomial or denominator from **Eq. (M1)**. This has roots when $s = b^{-1}\left(\sqrt[a]{R} - 1\right)$, which is the formula for the growth rate as expected from [15]. Exponential and deterministic distributions have $a = 1$ and $a \to \infty$, respectively. We get the roots of the characteristic polynomial of the exponential distribution by simply substituting in **Eq. (M3)**. The deterministic distribution yields $W(s) = e^{-sg}$ at the limit, is equivalent to applying a delay of g time units and has solution to its characteristic polynomial of $s = \log \sqrt[g]{R}$.

The bimodal distribution we consider is a mixture of two Erlang distributions with state sizes $n_1 = a_1$ and $n_2 = a_2$ and $\boldsymbol{\alpha} = [\alpha_1\ 0\ \ldots\ 0, 1 - \alpha_1\ 0\ \ldots\ 0]$, which has $n_1 - 1$ and then $n_2 - 1$ zeros respectively. The choice of α_1 defines the mixture weighting. The state matrix has size $(n_1 + n_2)^2$ with $\mathbf{T}_{\kappa\kappa} = -b_1^{-1}$ and $\mathbf{T}_{\kappa\kappa+1} = b_1^{-1}$ for $1 \leq \kappa \leq n_1$ and $\mathbf{T}_{\kappa\kappa} = -b_2^{-1}$ and $\mathbf{T}_{\kappa\kappa+1} = b_2^{-1}$ for $n_1 + 1 \leq \kappa \leq n_2$. The b_1 and b_2 are chosen to get mean generation time g. We obtain $W(s) = \sum_{\kappa=1}^{2} \alpha_\kappa (b_k s + 1)^{-a_\kappa}$ and numerically compute roots of its characteristic polynomial. We can easily extend this formulation to higher order mixtures. The phase-type structure allows us to describe complex distributions without losing analytical tractability.

Margins of stability and notions of controllability

In the above subsections we described the elements of the renewal epidemic model and its characteristic polynomial $1 - RW(s)$. Here we review the concepts of gain, phase and delay margin from classical control theory, which underpin our results in the main text and provide measures of how distant linear systems are from stability [17]. The loop transfer function is $L(s) = -RW(s)$ and describes the dynamics around the loop as shown in the block diagram of **Fig 1A**. In the main text we expand this $L(s)$ formulation to include some controller $K(s)$ and other system architectures but the general principles that we detail here remain valid.

When $L(s) = -1 + j0$ our closed loop TF $G(s) = \big(1 + L(s)\big)^{-1}$ just becomes unstable (i.e., it is infinite) with $j = \sqrt{-1}$. We can describe the distance of $L(s)$ from this critical stability point in terms of the multiplicative factor i.e., the gain, and the angular change i.e., the phase, that respectively scales and rotates $L(s)$ onto $-1 + j0$. These distances are termed the gain and phase margin [17] and relate to the polar description of a complex number. The gain margin $M_{\mathrm{G}} \stackrel{\text{def}}{=} |L(j\omega_{\mathrm{PC}})|^{-1}$ is the inverse of the magnitude of $L(s)$ evaluated at ω_{PC} the first frequency at which the phase crosses $-\pi$ radians. We can therefore multiply our feedback loop signals by the factor M_{G} to drive the epidemic to the critical stability point [31].

The phase margin $M_{\mathrm{P}} \stackrel{\text{def}}{=} \pi + \Phi(L(j\omega_{\mathrm{GC}})$ is the phase $\Phi(.)$ evaluated at the frequency ω_{GC} at which the gain of $L(s)$ crosses 1 and added to π. We can rotate the phase of the loop by M_{P} to drive the system to the critical stability point [31]. The phase margin is not intuitive for our epidemic analyses but can be transformed into the more interpretable delay margin M_{D}. This measures how much lag or pure delay forces $L(s)$ to the critical point (lag tends to decrease phase). If we multiply $L(s)$ by $e^{-sM_{\mathrm{D}}}$ we attain critical stability. We compute all margins (using in-built functions from the MATLAB control system toolbox. Note that for systems with multiple gain and phase crossover frequencies we use the minimum margin to ensure stability.

Importantly, the distance of our system from stability requires specifying both the gain margin and one of either the phase or delay margin [17]. We propose this pair representation as our measure of epidemic controllability, which quantifies the effort required to control an unstable epidemic or the perturbation required to destabilise an epidemic that is already under control. This is different from the formal definition of controllability in control theory, which says that a system is controllable if inputs exist to drive it from any initial state to any desired state in finite time. Our definition only considers what inputs can force epidemics to critical stability and the

required intensity of those inputs. This relates to margins (also termed relative stability) and corresponds directly to the informal definition commonly applied in infectious diseases [1,10].

Although controllability here defines the control effort needed to stabilise infections, it does not measure performance. Performance depends on our control objectives i.e., what we want our interventions to achieve [17]. These may include desired gain and delay margins but generally we may want to stipulate characteristics of the system response, $i(t)$ in our setting, to some desired dynamics $u(t)$. One key performance measure is the ability to accurately track $u(t)$, set by the steady state error $\lim_{t\to\infty} i(t) - u(t) = \lim_{s\to 0} s(I(s) - U(s))$. The second limit follows from the final value theorem [17]. If $U(s) = us^{-1}$ is our desired equilibrium of infections, then $\lim_{s\to 0} sG(s)M(s) - u$ defines accuracy. Other important measures of performance are the peak overshoot $\max_t i(t) - u(t)$ and the level of oscillation of $i(t)$ about $u(t)$ or the equilibrium.

Bibliography

1. Fraser C, Riley S, Anderson RM, Ferguson NM. Factors that make an infectious disease outbreak controllable. Proc Natl Acad Sci USA. 2004;101: 6146–6151. doi:10.1073/pnas.0307506101
2. Eames KTD, Keeling MJ. Contact tracing and disease control. Proc Biol Sci. 2003;270: 2565–2571. doi:10.1098/rspb.2003.2554
3. Ferguson N, Laydon D, Nedjati-Gilani G, Others. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID- 19 mortality and healthcare demand. Imperial College London; 2020.
4. Hethcote HW. The Mathematics of Infectious Diseases. SIAM Rev. 2000;42: 599–653. doi:10.1137/S0036144500371907
5. Team WER. Ebola Virus Disease in West Africa – The First 9 Months of the Epidemic and Forward Projections. N Engl J Med. 2014;371: 1481–95.
6. Walport MJ, Professor Sir Mark Walport on behalf of the Expert Working Group for the Royal Society's programme on non-pharmaceutical interventions. Executive Summary to the Royal Society report "COVID-19: examining the effectiveness of non-pharmaceutical interventions". Philos Trans A, Math Phys Eng Sci. 2023;381: 20230211. doi:10.1098/rsta.2023.0211
7. Bauch CT, Lloyd-Smith JO, Coffee MP, Galvani AP. Dynamically modeling SARS and other newly emerging respiratory illnesses: past, present, and future. Epidemiology. 2005;16: 791–801. doi:10.1097/01.ede.0000181633.80269.4c
8. Peak CM, Childs LM, Grad YH, Buckee CO. Comparing nonpharmaceutical interventions for containing emerging epidemics. Proc Natl Acad Sci USA. 2017;114:

4023–4028. doi:10.1073/pnas.1616438114

9. Casella F. Can the COVID-19 Epidemic Be Controlled on the Basis of Daily Test Reports? IEEE Control Syst Lett. 2021;5: 1079–1084. doi:10.1109/LCSYS.2020.3009912
10. McCaw JM, Glass K, Mercer GN, McVernon J. Pandemic controllability: a concept to guide a proportionate and flexible operational response to future influenza pandemics. J Public Health. 2014;36: 5–12. doi:10.1093/pubmed/fdt058
11. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438: 355–359. doi:10.1038/nature04153
12. Anderson R, Donnelly C, Hollingsworth D, Keeling M, Vegvari C, Baggaley R. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: methods of. The Royal Society. 2020;
13. Ferretti L, Wymant C, Kendall M, Zhao L, Nurtay A, Abeler-Dörner L, et al. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science. 2020;368. doi:10.1126/science.abb6936
14. Grantz KH, Lee EC, D'Agostino McGowan L, Lee KH, Metcalf CJE, Gurley ES, et al. Maximizing and evaluating the impact of test-trace-isolate programs: A modeling study. PLoS Med. 2021;18: e1003585. doi:10.1371/journal.pmed.1003585
15. Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B. 2007;274: 599–604.
16. Pates R, Ferragut A, Pivo E, You P, Paganini F, Mallada E. Respect the unstable: delays and saturation in contact tracing for disease control. SIAM J Control Optim. 2022;60: S196–S220. doi:10.1137/20M1377825
17. Ogata K. Modern Control Engineering, Volume 1. 3, illustrated ed. Prentice Hall; 1997.
18. Åström KJ, Murray RM. Feedback systems: an introduction for scientists and engineers. Princeton University Press; 2010. doi:10.2307/j.ctvcm4gdk
19. Churcher T, Cohen J, Ntshalintshali N, Others. Measuring the path toward malaria elimination. Science. 2014;344: 1230–32.
20. Sun K, Wang W, Gao L, Wang Y, Luo K, Ren L, et al. Transmission heterogeneities, kinetics, and controllability of SARS-CoV-2. Science. 2021;371. doi:10.1126/science.abe2424
21. Ali ST, Wang L, Lau EHY, Xu X-K, Du Z, Wu Y, et al. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science. 2020;369: 1106–1109. doi:10.1126/science.abc9004
22. Favero M, Scalia Tomba G, Britton T. Modelling preventive measures and their effect on generation times in emerging epidemics. J R Soc Interface. 2022;19: 20220128.

doi:10.1098/rsif.2022.0128

23. Stott I, Hodgson DJ, Townley S. Beyond sensitivity: nonlinear perturbation analysis of transient dynamics. Methods Ecol Evol. 2012;3: 673–684. doi:10.1111/j.2041-210X.2012.00199.x
24. Pellis L, Ball F, Trapman P. Reproduction numbers for epidemic models with households and other social structures. I. Definition and calculation of R0. Math Biosci. 2012;235: 85–97. doi:10.1016/j.mbs.2011.10.009
25. Parag KV, Obolski U. Risk averse reproduction numbers improve resurgence detection. PLoS Comput Biol. 2023;19: e1011332. doi:10.1371/journal.pcbi.1011332
26. Hu J, Qi G, Yu X, Xu L. Modeling and staged assessments of the controllability of spread for repeated outbreaks of COVID-19. Nonlinear Dyn. 2021;106: 1411–1424. doi:10.1007/s11071-021-06568-z
27. Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS One. 2007;2: e758. doi:10.1371/journal.pone.0000758
28. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178: 1505–1512. doi:10.1093/aje/kwt133
29. Roberts MG, Nishiura H. Early estimation of the reproduction number in the presence of imported cases: pandemic influenza H1N1-2009 in New Zealand. PLoS One. 2011;6: e17835. doi:10.1371/journal.pone.0017835
30. Morris DH, Rossine FW, Plotkin JB, Levin SA. Optimal, near-optimal, and robust epidemic control. Commun Phys. 2021;4: 78. doi:10.1038/s42005-021-00570-y
31. Seiler P, Packard A, Gahinet P. An introduction to disk margins [lecture notes]. IEEE Control Syst. 2020;40: 78–95. doi:10.1109/MCS.2020.3005277
32. Cox DR. A use of complex probabilities in the theory of stochastic processes. Math Proc Camb Phil Soc. 1955;51: 313–319. doi:10.1017/S0305004100030231
33. Yan P, Chowell G. Quantitative Methods for Investigating Infectious Disease Outbreaks. Cham, Switzerland: Springer; 2019.
34. Parag KV, Donnelly CA, Zarebski AE. Quantifying the information in noisy epidemic curves. Nat Comput Sci. 2022;2: 584–594. doi:10.1038/s43588-022-00313-1
35. Goldstein E, Dushoff J, Ma J, Plotkin JB, Earn DJD, Lipsitch M. Reconstructing influenza incidence by deconvolution of daily mortality time series. Proc Natl Acad Sci USA. 2009;106: 21825–21829. doi:10.1073/pnas.0902958106
36. Woolhouse M, Dye C, Etard J, Others. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. PNAS. 1997;94: 338–42.
37. Blanquart F, Hozé N, Cowling BJ, Débarre F, Cauchemez S. Selection for infectivity

profiles in slow and fast epidemics, and the rise of SARS-CoV-2 variants. Elife. 2022;11. doi:10.7554/eLife.75791

38. Britton T, Ball F, Trapman P. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. Science. 2020;369: 846–849. doi:10.1126/science.abc6810
39. Lee H, Nishiura H. Sexual transmission and the probability of an end of the Ebola virus disease epidemic. J Theor Biol. 2019;471: 1–12. doi:10.1016/j.jtbi.2019.03.022
40. Glass K, Mercer GN, Nishiura H, McBryde ES, Becker NG. Estimating reproduction numbers for adults and children from case data. J R Soc Interface. 2011;8: 1248–1259. doi:10.1098/rsif.2010.0679
41. Nowzari C, Preciado VM, Pappas GJ. Analysis and control of epidemics: A survey of spreading processes on complex networks. IEEE Control Syst. 2016;36: 26–46. doi:10.1109/MCS.2015.2495000
42. How Control Theory Can Help Us Control COVID-19 - IEEE Spectrum [Internet]. [cited 12 Sep 2023]. Available: https://spectrum.ieee.org/how-control-theory-can-help-control-covid19
43. Parag KV, Cowling BJ, Donnelly CA. Deciphering early-warning signals of SARS-CoV-2 elimination and resurgence from limited data at multiple scales. J R Soc Interface. 2021;18: 20210569. doi:10.1098/rsif.2021.0569
44. Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. Nat Med. 2020;26: 855–860. doi:10.1038/s41591-020-0883-7
45. Li J, Blakeley D, Smith RJ. The failure of R0. Comput Math Methods Med. 2011;2011: 527610. doi:10.1155/2011/527610
46. Heffernan JM, Smith RJ, Wahl LM. Perspectives on the basic reproductive ratio. J R Soc Interface. 2005;2: 281–293. doi:10.1098/rsif.2005.0042
47. Parag KV, Cowling BJ, Lambert BC. Angular reproduction numbers improve estimates of transmissibility when disease generation times are misspecified or time-varying. Proc Roy Soc B. 2023; 290: 20231664. doi:10.1098/rspb.2023.1664
48. Bin M, Cheung PYK, Crisostomi E, Ferraro P, Lhachemi H, Murray-Smith R, et al. Post-lockdown abatement of COVID-19 by fast periodic switching. PLoS Comput Biol. 2021;17: e1008604. doi:10.1371/journal.pcbi.1008604

Funding

KVP acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO

Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. For the purpose of open access, the author has applied a 'Creative Commons Attribution' (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Data availability statement

All data and code underlying the analyses and figures within this study are freely available in MATLAB at: https://github.com/kpzoo/EpidemicControllability

Author Contributions

KVP – Conceptualisation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Validation, Visualisation, Writing – original draft, Writing – review and editing.